Temporal decomposition of speech and its relation to phonetic information

Temporal decomposition of speech and its relation to phonetic information

Proefschrift

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de Rector Magnificus, prof. ir. M. Tels, voor een commissie aangewezen door het College van Dekanen in het openbaar te verdedigen op dinsdag 30 mei 1989 te 14.00 uur

door

Astrid Maria Louise Van Dijk-Kappers

geboren te Haarlem

Dit proefschrift is goedgekeurd door de promotoren:

Prof. Dr. S.G. Nooteboom

en

Prof. Dr. H. Bouma

Dit onderzoek werd uitgevoerd aan het Instituut voor Perceptie Onderzoek (IPO) te Eindhoven, en werd financieel gesteund door de Stichting Taalwetenschap van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO).

Contents

List of definitions

The following list gives the definitions of some terms used throughout this thesis, in alphabetic order. Italicized words in the definitions are defined separately.

Acoustic event	Perceptibly distinct acoustic segment of speech.
Articulators	The parts of the oral tract that can be used to form sounds (e.g. lips and tongue).
Articulatory gesture	Movement of the whole configuration of the vocal tract aimed at realizing a *speech event*. This term will not be applied to movements of single *articulators*.
Articulatory position	Configuration of the vocal tract.
Articulatory target	Target configuration of the vocal tract.
Coarticulation	Temporal overlap of two adjacent *articulatory gestures* which influences the acoustic realizations of adjacent *phonemes*.
Phoneme	The smallest distinctive segment of sound which brings about a change of meaning between words.
Phoneme boundary	Perceptually determined location in the acoustic speech signal which best approximates a boundary between the acoustic realizations of two neighbouring, phonologically transcribed *phonemes*.
Speech event	Segment of speech corresponding to the acoustic realization of a *phoneme*, or, if *subphonemic events* can be distinguished (for instance in the case of plosives or diphthongs), corresponding to a *subphonemic event*.
Subphonemic event	*Acoustic event* which can be considered as a distinct part of the acoustic realization of a *phoneme* (for instance burst and occlusion of a plosive).

Target function	Function describing the temporal evolution of the *target vector* which is associated with it. Determined by means of temporal decomposition.
Target vector	Vector of speech parameters which possibly models an ideal *articulatory target*. Determined by means of temporal decomposition.

List of symbols

The following list summarizes the specific symbols used throughout this thesis. Vectors and matrices are denoted by lower and upper case characters respectively, both printed bold faced.

N	Total number of frames.
$n, \quad 1 \leq n \leq N$	Frame number.
I	Total number of speech parameters per frame.
$\mathbf{y}$	Frame of speech parameters.
$\tilde{\mathbf{y}}$	Approximation of $\mathbf{y}$.
$y_i(n)$	i^{th} speech parameter of n^{th} frame.
K	Total number of target functions and target vectors.
$\phi_k(n)$	k^{th} target function.
$\mathbf{a}(k)$	k^{th} target vector, consisting of I elements.
$\theta(n_c)$	Measure of spread around frame n_c.
$\alpha(n)$	Weighting factor.
E	Euclidian distance or error measure.

List of abbreviations

The following list gives an overview of abbreviations used in this thesis.

A	Area
CVC	Consonant-vowel-consonant
BF	Filter bank output
F	Formant
LA	Log area
LAR	Log-area ratio
LPC	Linear predictive coding
RC	Reflection coefficient
S	Spectral coefficient
SED	Smallest Eucledian distance
TD	Temporal decomposition

IPA symbols

IPA symbols used in this thesis, together with Dutch example words.

IPA symbol	example word	IPA symbol	example word
/ɑ/	m<u>a</u>t	/p/	<u>p</u>aal
/ɛ/	l<u>e</u>s	/b/	<u>b</u>as
/I/	p<u>i</u>t	/t/	<u>t</u>ak
/ɔ/	r<u>o</u>t	/d/	<u>d</u>as
/œ/	h<u>u</u>t	/k/	<u>k</u>at
/a/	l<u>aa</u>t	/f/	<u>f</u>ier
/e/	h<u>ee</u>t	/v/	<u>v</u>os
/i/	b<u>ie</u>t	/s/	<u>s</u>oep
/o/	b<u>oo</u>t	/z/	<u>z</u>on
/y/	m<u>uu</u>r	/x/	<u>g</u>ok
/u/	b<u>oe</u>r	/m/	<u>m</u>an
/ø/	k<u>eu</u>s	/n/	<u>n</u>at
/ɛɪ/	d<u>ij</u>k	/ŋ/	la<u>ng</u>
/ʌy/	t<u>ui</u>n	/l/	<u>l</u>aag
/au/	k<u>ou</u>d	/r/	<u>r</u>ood
/ə/	d<u>e</u>	/j/	<u>j</u>as
		/w/	<u>w</u>al
		/h/	<u>h</u>ok

Chapter 1

General introduction

1.1 Introduction

During the speech process, the tongue, lips and other articulators move continuously from one articulatory position to the next. As a consequence, the acoustic speech signal also changes in time continuously. In spite of this continuous variation, the signal conveys information that can be represented by a sequence of discrete units such as words, letters or phonetic symbols. Examples of both kinds of representations can be seen in Fig. 1.1. From the top downwards an orthographic text, a time-aligned phonetic transcription, an amplitude-time waveform and a spectrogram of the same utterance are shown.

It will be clear immediately that the relations between the various representations are far from trivial. Not even A and B have a one-to-one relationship, let alone A or B to C or D. One can notice that spaces between words cannot be found as silences in the spectrogram or waveform and, conversely, silences that do occur nearly always indicate plosives (e.g. /b/ or /g/) rather than word boundaries. Furthermore, acoustic realizations of phonemes do not appear as discrete units in waveform or spectrogram.

Listening carefully to small portions of speech of the length of one or at most two phonemes by means of a gating technique (e.g. 't Hart and Cohen, 1964) confirms the above-mentioned visual observations: sharp boundaries between the acoustic realizations of phonemes do not exist. In the transition region both phonemes can often be perceived simultaneously.

The fact that the acoustic realizations of phonemes do overlap in time contributes to the efficiency of speech as messages. 15 phonemes per second, which is not unusual in normal conversation, would be more than the ear could cope with if phonemes were a string of discrete acoustic events (Liberman,

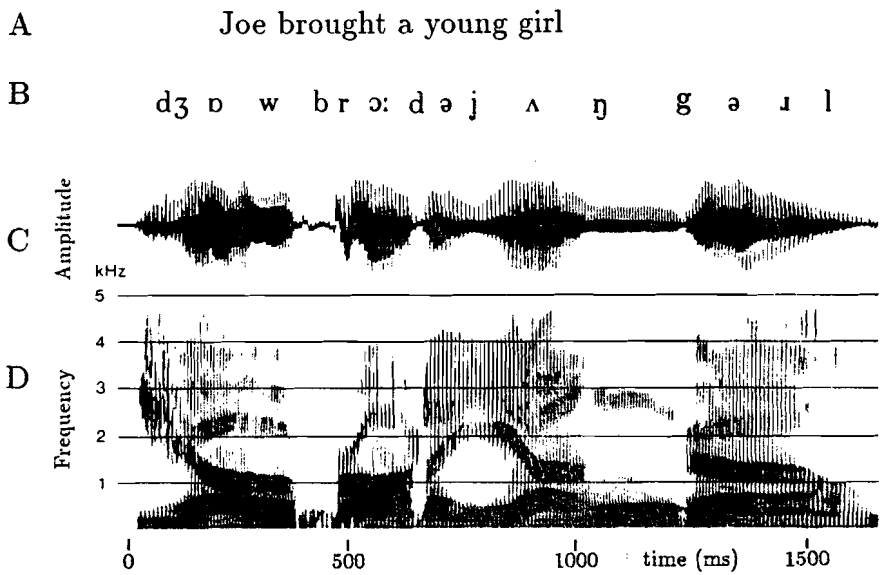

Figure 1.1: This figure shows four different representations of the utterance "Joe brought a young girl". A. orthographic text, B. time-aligned phonetic transcription, C. amplitude-time waveform and D. spectrogram. (Figure from Marcus and Van Lieshout (1984)).

Cooper, Shankweiler and Studdert-Kennedy, 1967). Producing such a high speech rate also implies that phonemes cannot be realized separately. Due to the inertness of the vocal tract, anticipation for the next phoneme is necessary. According to Liberman et al. (1967) perception mirrors articulation more closely than it does sounds. This is in agreement with the outcome of the perception experiments of Fowler (1984). She states that listeners recover the overlapping segments that talkers produce.

The mappings between discrete and continuous representations of the same speech utterance are quite complex. Both listeners and speakers perform this mapping almost automatically, hardly hindered by variations of speaking style or speech rate. Speech researchers, however, encounter serious problems. For many applications, such as automatic speech recognition or speech synthesis, phonemes are the most important entities, but the overlapping, non-stable, non-invariant acoustic realizations of phonemes are very difficult to cope with.

As a consequence, most approaches ignore these phoneme features, and treat the acoustic signal as if it consisted of a sequence of discrete invariant non-overlapping speech units. This is a severe simplification of reality which, of course, has its impact on the ultimate results which can be obtained.

This thesis centres on the so-called temporal decomposition technique, which tackles this mapping problem from a relatively new angle.

1.2 Short history of temporal decomposition

In 1983 a new technique for economical speech coding was introduced by Atal (1983). Based on the considerations that speech events do not occur at uniformly spaced time intervals and that articulatory movements are sometimes fast, sometimes slow, he concluded that uniform time sampling of speech parameters is not efficient. Thus, he proposed the temporal decomposition method in order to break up the continuous variation of speech parameters into discrete overlapping units of variable lengths located at non-uniformly spaced time intervals. Although this method was based on articulatory considerations, Atal did not attempt to interpret the possibly phonetic meaning of the so-determined units. Economical speech coding was his primary objective.

In 1984 Marcus and Van Lieshout suggested that temporal decomposition was promising as a means of segmenting speech into a sequence of overlapping events closely related to the phonetic structure of the speech signal. They also pointed out that some of the first few steps of Atal's original method caused some problems, which, however, were not fundamental to the philosophy of the temporal decomposition method. Unfortunately, due to these problems, they could not give any quantitative data about the phonetic relevance of the overlapping units. In their paper they gave some suggestions for improving the method. It was along these lines that the present research developed.

At the time this research started, in 1985, this was the only literature about temporal decomposition. In the meantime, however, a few other articles on the subject have been published. Most of these very concise articles focussed attention on a description of the method and gave some suggestions for specific applications. Speech synthesis was advocated by Chollet, Grenier and Marcus (1986), Ahlbom, Bimbot and Chollet (1987) and Bimbot, Ahlbom and Chollet (1987). Niranjan and Fallside (1987) gave a geometric interpretation of the temporal decomposition results and proposed some improvements. Finally, Bimbot, Chollet, Deleglise and Montacie (1988) and Marteau, Bailly

and Janot-Giorgetti (1988) did some pilot experiments on speech recognition, obtaining promising results.

Further details of these articles can be found in subsequent chapters in relation to the research presented in this thesis.

1.3 Aims of this thesis

The current research is a continuation of the work initiated by Marcus and Van Lieshout (1984). The principal objective is to investigate systematically the possibilities of temporal decomposition as a tool with which to segment the speech parameters into phonetically relevant speech units. Since Marcus and Van Lieshout encountered some shortcomings of the original temporal decomposition method, the method itself will also be under investigation. Especially those parameters which have a strong influence on the decompositions will be evaluated. The phonetic relevance of the decomposition will be used as a criterion for good performance.

Many segmentation methods make use of phonetic information. For instance, a common method is to use the phonetic transcription of an utterance in order to optimize the number and locations of segment boundaries (e.g. Bridle and Chamberlain, 1983; Lennig, 1983; Van Hemert, 1987). We, however, are interested in how much information can be derived directly from the acoustic speech signal. Thus, the use of phonetic knowledge will be excluded explicitly. Though this will probably limit the achievements of the method right now, a good view can be expected of the possibilities and shortcomings of temporal decomposition as a tool for extracting phonetically relevant units from the speech signal.

This research is mainly exploratory in nature; no specific application is aimed at. Nevertheless, if the overlapping units turn out to be phonetically interpretable, the results may provide information which could be of interest to workers in various fields of speech research, such as speech coding, synthesis, segmentation and recognition.

1.4 Contents of this thesis

In Chapter 2 a detailed description of Atal's original temporal decomposition method is given, together with its shortcomings. For each problem an alternative solution is proposed, which results in a modified and extended temporal decomposition method.

The choice of input parameters for temporal decomposition is not restricted to a specific set of speech parameters. The parameters proposed by Atal were the log-area parameters which, probably due to their close relationship to the positions of the articulators, gave reasonably satisfactory results. However, since it is not inconceivable that better candidates exist, temporal decomposition results are compared using nine different sets of input parameters. This is the subject of Chapter 3.

In Chapter 4 some other explorations are described. After optimization of the method and the choice of the speech parameters which give the best results, the question remains how well the decomposed speech signal models the original signal. For the derivation of phonetic information the quality of the resynthesized speech signal is only of secondary importance; intelligibility experiments, however, may reveal important information about the usefulness and the limitations of the model on which temporal decomposition is based and the ultimate results which may be obtained with it. Perception experiments are described, evaluating the intelligibility of temporally decomposed and resynthesized CVC utterances. This chapter also examines the achievements of temporal decomposition using a database consisting of 100 phonologically balanced German sentences. The phonetic relevance of target functions is judged both from a phonological and a phonetic point of view.

Finally, Chapter 5 gives an evaluation of the state of the art of temporal decomposition. Future directions and applications as well as possible improvements and remaining shortcomings are discussed.

Since chapters 2 and 3 are also meant for publication as articles, they contain some overlapping of information.
Chapter 2: A.M.L. Van Dijk-Kappers and S.M. Marcus (1989), Temporal decomposition of speech, *Speech Communication 8*, (in press).
Chapter 3 is a modified version of an article accepted for publication by *Speech Communication*: A.M.L. Van Dijk-Kappers (1989), Comparison of parameter sets for temporal decomposition.

Chapter 2

Temporal decomposition of speech

Abstract

In articulatory phonetics, speech is described as a sequence of distinct articulatory gestures towards and away from articulatory targets, resulting in a sequence of speech events. Due to overlap of the gestures, these articulatory 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 target 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 some 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, segmentation and synthesis may profit.

2.1 Introduction

Articulatory phonetics is based on a description of speech as a sequence of articulatory gestures. Each gesture can be considered as a movement towards and away from an articulatory target. Different articulatory targets will result in different acoustic consequences. A limited number of adjacent gestures can overlap one another, resulting in the characteristic transitions between the acoustic realizations of 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 taking 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). Applications in the field of speech synthesis were also 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 recognized 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 chapter is to propose some improvements and extensions to the original method to overcome these deficiencies. Although some of our choices stem from 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 to other possible applications of this technique. As this chapter 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 needed for describing the modifications.

2.2 Temporal decomposition

Atal (1983) assumed that, given some suitable parametric representation of the input speech signal, 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 been produced by a sequence of K movements aimed at realizing K articulatory targets. Let us denote the speech parameters corresponding to the k^{th} target by a *target vector*, $\mathbf{a}(k)$, 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, $\mathbf{y}(n)$, by the following linear

combination of target vectors and functions:

$$\tilde{\mathbf{y}}(n) = \sum_{k=1}^{K} \mathbf{a}(k)\phi_k(n), \quad 1 \leq n \leq N \tag{2.1}$$

or, in matrix notation,

$$\tilde{\mathbf{Y}} = \mathbf{A}\mathbf{\Phi}. \tag{2.2}$$

$\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{y}}(n)$ are approximations of the observed speech parameters. The acoustic parameters chosen by Atal to describe the speech signal $\mathbf{y}(n)$ are the log-area parameters. These parameters model the cross-sectional areas of an acoustic tube, vary slowly in time and show a high mutual linear dependence, which makes them eminently suited for temporal decomposition (Van Dijk-Kappers, 1989). These parameters are derived from the filter parameters of an LPC analysis; the source parameters (filter gain, pitch and voiced/unvoiced parameter) are not used for the purpose of temporal decomposition.

In Eqs. (2.1) and (2.2), both the target vectors and target 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 Eq. (2.1) for the $\phi_k(n)$; after that, the optimal target vectors can be computed.

Eq. (2.1) can be inverted to give the k^{th} target function $\phi_k(n)$ as a linear combination of the speech parameters $y_i(n)$

$$\phi_k(n) = \sum_{i=1}^{I} w_{ki} y_i(n) \tag{2.3}$$

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 time window over the matrix of speech parameters $\mathbf{Y}$. 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 $\mathbf{Y}_w$. In matrix notation this can be expressed as

$$\mathbf{Y}_w^T = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{2.4}$$

where both $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices, their columns containing the singular vectors. $\mathbf{D}$ is a diagonal matrix of singular values, the square roots of the eigenvalues of $\mathbf{Y}_w^T\mathbf{Y}_w$. 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 explain more than 99 % 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. 2.1, 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 $\mathbf{u}_i$ from the matrix $\mathbf{U}$. It can be seen that only the first few singular vectors are important and account for most of the variance.

It follows from Eq. (2.4) that the speech parameters $y_i(n)$ can be expressed as a linear combination of the parameters $u_i(n)$. Substituting this in Eq. (2.3) and taking only the s most significant singular vectors results in an important data reduction in solving Eq. (2.3). Thus, the target function $\phi_k(n)$ can be represented as

$$\phi_k(n) = \sum_{i=1}^{s} b_{ki} u_i(n) \tag{2.5}$$

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}.

2.2.1 Determination of the target functions

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

$$\theta(n_c) = [\sum_n \alpha(n)\phi_k^2(n) / \sum_n \phi_k^2(n)]^{\frac{1}{2}} \tag{2.6}$$

where $\alpha(n)$ is a weighting factor. The sum over n extends over the N_w frames of the analysis window of which 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, $\phi_k(n)$ of Eq. (2.6) is replaced by the expression of Eq. (2.5), and the derivatives of

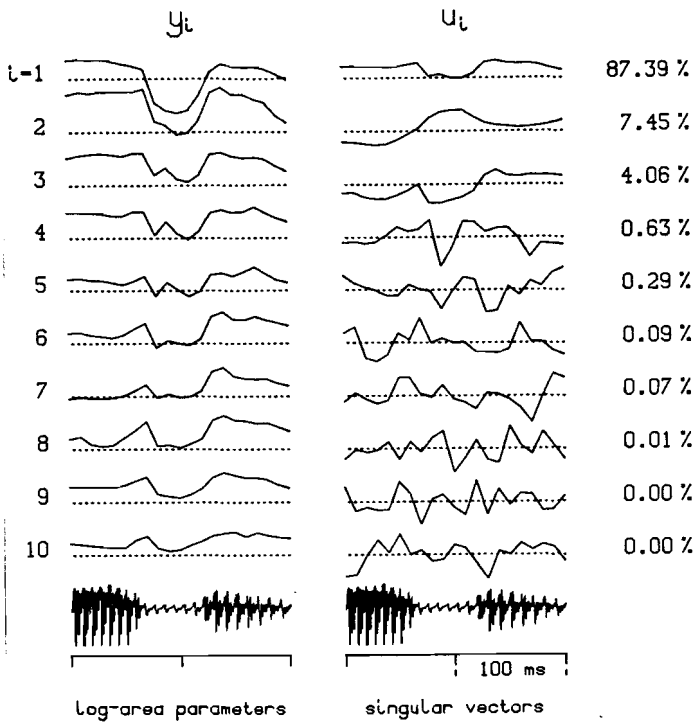

Figure 2.1: Plot of the 10 log-area parameters, $y_i(n)$, of a 210 ms window and the singular vectors, $u_i(n)$, of the same speech segment.

$\theta(n_c)$ (or $\ln\theta(n_c)$, which gives the same results but with less computational effort) with respect to the coefficients b_{ki} are set equal to 0. This results in the eigenvalue equation

$$\mathbf{R}\mathbf{b} = \lambda\mathbf{b} \tag{2.7}$$

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

$$r_{ij} = \sum_n \alpha(n) u_i(n) u_j(n). \tag{2.8}$$

The smallest (or largest) eigenvalue λ provides the optimal choice of the coefficients b_{ki}, and with Eq. (2.5) the target function $\phi_k(n)$ is determined (Lawley & Maxwell, 1971; Atal, 1983).

2.2.2 Weighting factor of Atal

Atal proposed a quadratic weighting factor:

$$\alpha(n) = (n - n_c)^2 \tag{2.9}$$

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.

2.2.3 An alternative weighting factor

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

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

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

This iteration procedure, consisting of the successive models and the resulting target functions, can be seen in Fig. 2.2 for three different segments of speech. With this procedure a single target function is found within each analysis window. The target functions are 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.

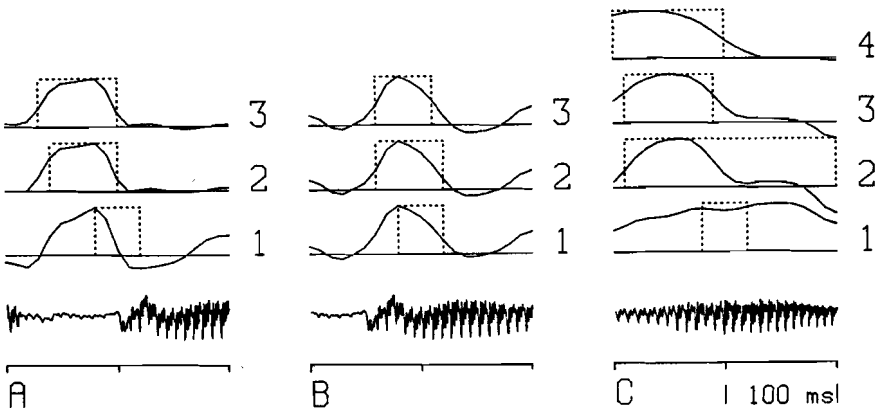

Figure 2.2: Target functions resulting from successive iterative models for three segments of speech. The initial choice of the model is fixed in the centre of the window, and the model converges in successive iterations.

The choice of the initial model m_1 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.

2.2.4 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 only be non-zero during a limited number of consecutive frames, but this requirement is not always fulfilled. Sometimes, as in Fig 2.2B, 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 Eq. (2.5). At other times the resulting target function is not complete, as in Fig. 2.2C, 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 for adapting the window size.

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. 2.3. If there is no frame which satisfies the conditions for n_{left}, the first frame number, n_0, will be assigned to n_{left}; likewise the last frame number, n_n, will be assigned to n_{right} if no frame to the right of $n_{\max}$ has a value smaller than h_w.

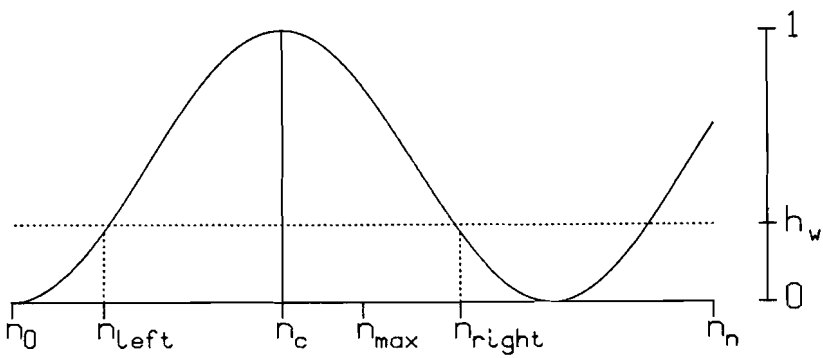

Figure 2.3: Schematic overview of the variables used for modifying the window.

As a measure of the amount of (left) edge effects we use $\sum_{n=n_0}^{n_{\text{left}}} \phi^2(n)$. 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 ensure that the resulting target function is not located completely outside the initial 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 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 is achieved in two or three iterations.

The results of this window adaptation procedure are shown in Fig. 2.4, where the same speech segments are used as in Fig. 2.2. The target functions numbered 1 indicate the resulting target functions of Fig. 2.2; 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.

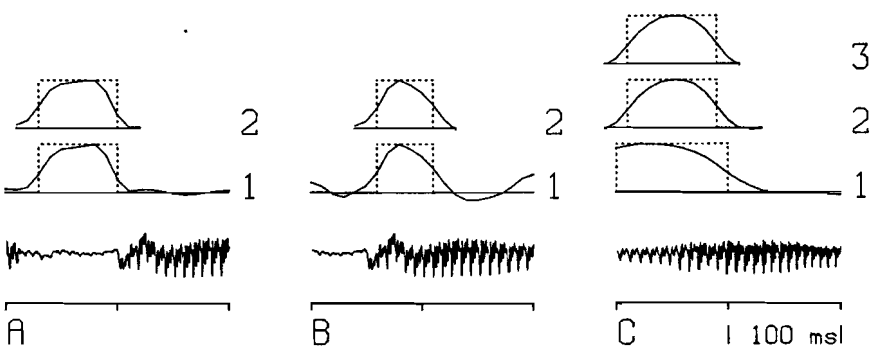

Figure 2.4: Iterative modification of the analysis window size and position, for the same three segments of speech as in Fig. 2.2.

2.3 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 such an analysis of a number of successive windows is shown in Fig. 2.5. 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.

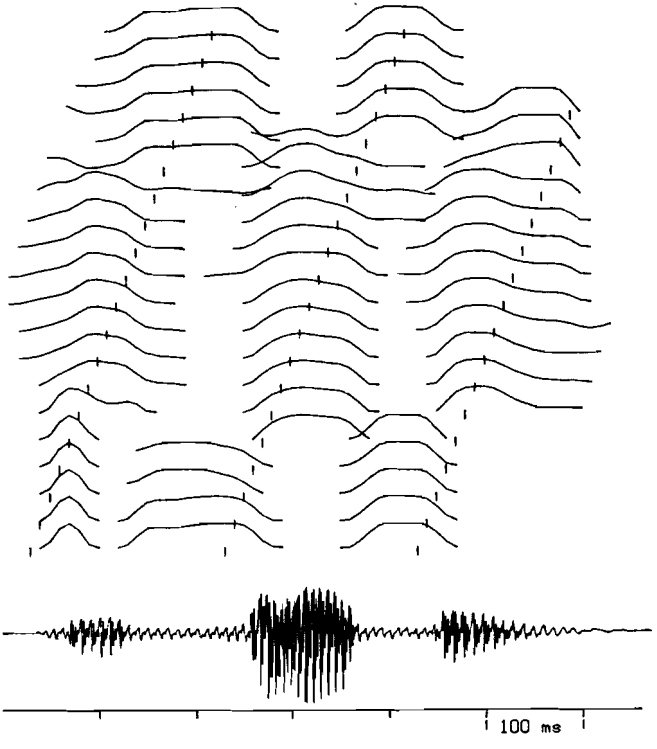

Figure 2.5: Target functions determined within the successive analysis windows of the utterance /dəbabə/ using the original method of Atal. The vertical bars indicate the centres of the analysis windows.

For comparison, we also show the analysis of the same utterance derived with the modified temporal decomposition method described in the previous sections (Fig. 2.6). 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.

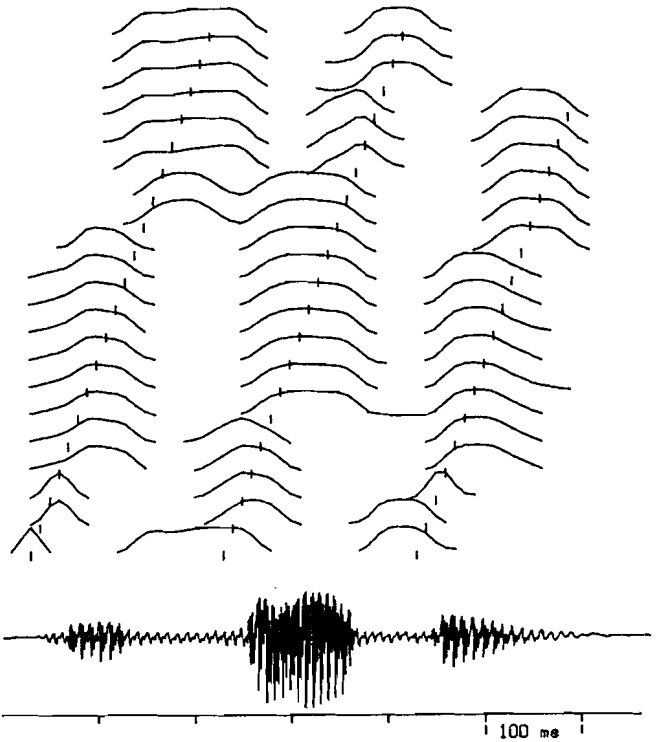

Figure 2.6: Target functions determined within the successive analysis windows of the utterance /dəbabə/ using our modified method.

Shifting the analysis window in steps of 10 ms means that the total number of target functions equals 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.

2.3.1 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 $\nu(n_c)$:

$$\nu(n_c) = \sum_{n=n_0}^{n_n} (n - n_c)\phi_k^2(n) / \sum_{n=n_0}^{n_n} \phi_k^2(n). \tag{2.10}$$

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.

2.3.2 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 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 is a slight chance of finding the same function once again, 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:

$$\cos(\alpha) = \sum_n \phi_{k-1}(n)\phi_k(n) / [\sum_n \phi_{k-1}^2(n) \sum_n \phi_k^2(n)]^{1/2}. \tag{2.11}$$

The summation extends over the overlapping frames n. If $\cos(\alpha)$ 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.

2.3.3 Determination of the target vectors

For the determination of the target vectors we use the same procedure as proposed by Atal. The target vectors $\mathbf{a}(k)$ associated with the target functions $\phi_k(n)$ can be determined by minimizing the mean-squared error E, defined by:

$$E = \sum_n [\mathbf{y}(n) - \tilde{\mathbf{y}}(n)]^2, \tag{2.12}$$

or, by substituting Eq.(2.1)

$$E = \sum_n [\mathbf{y}(n) - \sum_{k=1}^{K} \mathbf{a}(k)\phi_k(n)]^2. \tag{2.13}$$

This equation can be solved for the $\mathbf{a}(k)$ by setting the partial derivatives of E with respect to $\mathbf{a}(k)$ equal to zero (Atal, 1983). This results in a set of target vectors $\mathbf{a}(k)$, each consisting of a frame of 10 log-area parameters.

2.3.4 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 target 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. 2.7. 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. 2.7. In Fig. 2.7A there is a clear correspondence between the target functions and speech events, although a function associated with the burst of the second /b/ is missing. In Fig. 2.7B there is one speech event described by two target functions. In Chapter 4 the phonetic relevance of the decomposition will be further investigated and discussed.

2.4 Evaluation and discussion

In several respects, our modified temporal decomposition method gives better results than the original method of Atal. An important improvement is that target functions are now found in all situations, whereas in the original method sometimes a gap occurred between two functions. Of course, a gap is

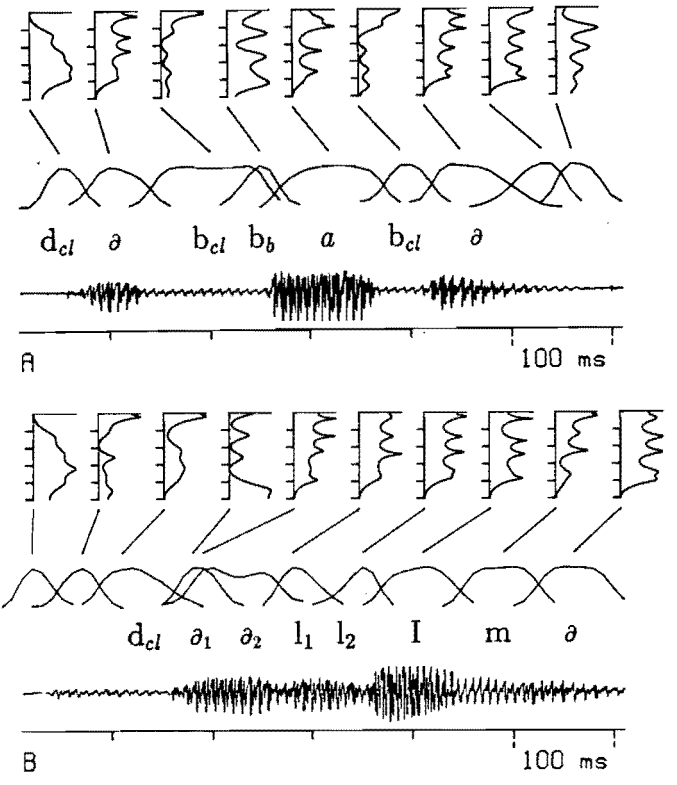

Figure 2.7: Temporal decomposition of some CVC utterances: A. /dəbabə/, B. /dəlImə/. The subscripts *cl* and *b* stand for closure and burst respectively.

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 2.5 and 2.6 will illustrate this. In spite of these modifications, the computation time has remained more or less the same, since the singular value decomposition is the most time-consuming part of the procedure.

2.4.1 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 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 differences 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 combinations embedded in the context /d$\partial$$C_1VC_2\partial$/. The consonants C_1 and C_2 were one of the phonemes /l/, /m/, /b/ or /p/ and the vowel V was one of the phonemes /a/, /I/ or /$\supset$/. 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. 2.7, 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 1A shows for each weighting factor the percentage of speech events described by zero, one, two or more target functions respectively. Although a reasonable percentage of the speech events is described by only one target function, 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.

Table 2.1: Percentages of the speech events associated with zero, one, two or more than two target functions. The results given are averaged over all speech events, including and excluding the bursts respectively.

	A. including bursts				B. excluding bursts			
	0	1	2	>2	0	1	2	>2
Rectangular	18	63	18	1	1	73	25	1
Atal	15	55	27	3	0	60	36	4

To get a better idea of the achievements of temporal decomposition, we show in Table 1B the results of the analysis of the same words, leaving out

the bursts. As can be seen, the improvement is considerable; only a very small percentage of the speech events is 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 1A and 1B 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 Eq. (2.5). 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, 1988).

2.4.2 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 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 (e.g. Bimbot et al., 1987).

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 with our target functions. In any case, the clustering technique causes a substantial increase in computation time.

2.5 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 powerful 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 desirable.

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

Chapter 3

Comparison of parameter sets for temporal decomposition

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 target vectors may correspond to ideal articulatory targets of which the target functions describe the temporal evolution. Although developed for economical speech coding, this method also provides an interesting tool for deriving phonetic information from the acoustic speech signal.

The speech parameters used by Atal (1983) when he proposed this method are the log-area parameters. Our modified temporal decomposition method (Van Dijk-Kappers and Marcus, 1987; 1989) also works with the log-area parameters as input. The method is not, however, restricted to log-area parameters; in principle, most commonly used parameter sets can be used. In this chapter we compare the results obtained with nine different sets of speech parameters, among are which log-area parameters, formants, reflection coefficients and filter bank output parameters.

The main performance criterion was the phonetic relevance of the target functions. The phonetic interpretation of the target vectors was also considered, but that turned out to be an ineffective criterion. Finally, 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 the filter bank output parameters form the most suitable parameter set available for temporal decomposition if only the phonetic relevance is considered. However, with respect to resynthesis, the log-area parameters must be classified as a better set.

3.1 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 target. 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). However, the so-called temporal decomposition method, proposed by Atal (1983) for economical speech coding, decomposes the speech signal into overlapping units, each 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 phonemes or subphonemic events. Indeed, we have shown (Van Dijk-Kappers and Marcus, 1987; 1989) that with some modifications and extensions, promising results were obtained. Using a restricted database consisting of 48 CVC combinations embedded in a neutral context, and excluding the bursts of the plosives, 74 % of the phonemes could be associated with precisely one target function and one target vector. Furthermore, only 1 % of the phonemes was missed. The remaining 25 % of the phonemes were associated with two or more target functions, which is possibly due to the fact that the acoustic realizations of these phonemes can be considered to consist of more than one acoustic event.

These results were obtained with our modified temporal decomposition method, which is more robust than Atal's original method. Important parameters (for instance the length of the analysis window) the optimal values of which depend on the speech segment under evaluation, are adjusted iteratively. Other parameter choices, such as iteration thresholds and initial values of parameters, have been optimized. The only parameters which have not been varied are the input parameters; up till now log-area parameters have always been used. These speech parameters have yielded the reasonably satisfactory results mentioned above, possibly owing to their close relationship to the positions of the articulators. It is, however, not inconceivable that better candidates exist. At this moment, fundamental insight into the properties that make a parameter set suitable for temporal decomposition is still lacking. As a consequence, the search for speech parameters which might give better results than log-area parameters must be explorative.

In this chapter, two objectives are pursued. In the first place, the aim is to find a set of speech parameters which improves the achievements of the method as compared to the results obtained with the log-area parameters. The second objective is to gain more insight into the temporal decomposition method with respect to the influence of the choice of input parameters. To reach both objectives, we compare the temporal decomposition results using nine different sets of speech parameters. As, in practice, the results cannot be predicted theoretically, we use sets which are often used for other purposes, such as speech coding or synthesis. Parameter sets proposed for temporal

decomposition in recent papers (Chollet et al., 1986; Ahlbom et al., 1987; Bimbot et al., 1987) are also included in the comparison. Moreover, the number of parameters, the amount of detail in the parameters and the relation to log-area parameters are varied over the sets. Except for the filter bank output parameters, all of these sets are LPC-derived. The main performance criterion will be the phonetic relevance of the target functions, since this gives a good indication of the phonetic relevance of the decomposition. In addition, the phonetic meaning of the associated target vectors will be considered. For those parameter sets which can be transformed into one another, a reconstruction error will be defined and evaluated.

The work reported here is part of a project studying the relationship between the target functions and target vectors determined by means of temporal decomposition and a phonetic transcription of the same utterance. The results may provide deeper insight into the structure of the speech signal. Such knowledge can be applied to economical speech coding or speech synthesis. Applications of the method as a preprocessor for automatic speech recognition or transcription may also be feasible.

In the following sections we will first give a brief description of the temporal decomposition method. Next we will devote a section to the various speech parameter sets 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 parameter spaces in which the target functions and target vectors should be determined.

3.2 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 k^{th} target by a target vector $\mathbf{a}(k)$ consisting of I speech parameters, and the temporal evolution of this target by a target function $\phi_k(n)$, the observed speech parameters $\mathbf{y}(n)$ can be approximated by the following linear combination of target vectors and functions:

$$\tilde{\mathbf{y}}(n) = \sum_{k=1}^{K} \mathbf{a}(k)\phi_k(n), \quad 1 \leq n \leq N \tag{3.1}$$

where $\tilde{\mathbf{y}}(n)$ is the approximation of $\mathbf{y}(n)$. 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 $\mathbf{y}(n)$ any kind of parameter set can be chosen, and a number of these will be compared in the following sections. In this equation, the target vectors and target functions, as well as their number and locations are unknown.

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

$$E = \sum_n [\mathbf{y}(n) - \tilde{\mathbf{y}}(n)]^2, \tag{3.2}$$

or, by substituting Eq. (3.1):

$$E = \sum_n [\mathbf{y}(n) - \sum_{k=1}^{K} \mathbf{a}(k)\phi_k(n)]^2. \tag{3.3}$$

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

According to Eq. (3.1), the target functions and target 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. 3.1. The plot shows the amplitude-time waveform of the utterance, the phonetic transcription and the automatically extracted target functions. The speech parameters, in this case log-area parameters, of the associated target vectors are transformed into the spectral domain and the corresponding log amplitude spectra are also shown.

Although temporal decomposition results in a description of speech in terms of linear combinations of the input parameters, the method itself is nonlinear, and complex to such a degree that its behaviour cannot be made explicit. As a consequence, the results obtainable with different parameter sets cannot be predicted theoretically from the log-area results, not even if the relationship between the two sets is linear.

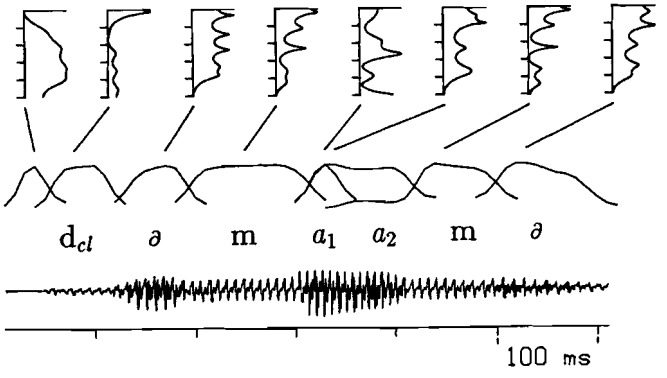

Figure 3.1: Temporal decomposition of the CVC utterance /dəmamə/.

3.3 Speech parameters to be compared

The parameter sets used in this chapter for the comparison of suitability for temporal decomposition form together a representative subset of the available sets for such an experiment. They are all characterized by the fact that they are often used for other purposes, such as speech coding or synthesis. Some of these sets are also used by others for temporal decomposition. Three sets are closely related, and differ only in the amount of detail or the number of parameters. It is hoped that the comparison of these three sets will add to the insight into temporal decomposition. Eight parameter sets were derived from the prediction coefficients 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 voice/unvoiced parameter) were left out of consideration. In the following 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 latter set, amplitude information is integrated in the parameters.

3.3.1 Parameter sets

Four of the LPC-derived parameter sets are directly related to the physical parameters of a model in which the vocal tract consists of an acoustic tube of I sections, each of the same length but of different cross-sectional area. They are

all convertible into one another through linear or non-linear transformations. Here they are presented in terms of the LPC prediction coefficients a_i (not to be confused with the elements of the target vector $\mathbf{a}(k)$).

(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:

$$\begin{aligned} k_i &= a_i^{(i)} \\ a_j^{(i-1)} &= \frac{a_j^{(i)} - a_i^{(i)} a_{i-j}^{(i)}}{1 - k_i^2}, \qquad 1 \le j \le i-1, \end{aligned} \tag{3.4}$$

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

(2) Area coefficients (A) Area coefficients are the cross-sections of the I successive sections of the vocal tube. Eq. (3.5) relates these parameters, A, to the reflection coefficients k:

$$\begin{aligned} A_{I+1} &= 1 \\ A_i &= A_{i+1} \frac{1+k_i}{1-k_i}, \qquad 1 \le i \le I. \end{aligned} \tag{3.5}$$

If a frame of I area 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 possible 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 coefficients 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:

$$\log A_i, \qquad 1 \le i \le I. \tag{3.6}$$

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

$$g_i = \log \frac{1 + k_i}{1 - k_i}, \qquad 1 \leq i \leq I, \tag{3.7}$$

or, by substituting Eq. (3.5) into Eq. (3.7), in terms of the area coefficients A:

$$g_i = \log \frac{A_i}{A_{i+1}}, \qquad 1 \leq i \leq I, \tag{3.8}$$

thereby immediately explaining their name. Along with the LA parameters, 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). Bimbot et al. (1987) reported that the LAR were the most suitable parameters for temporal decomposition they had found so far. It is, however, unclear whether they based this on experimental or theoretical grounds.

Although Eq. (3.7) 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. 3.2. Viswanathan and Makhoul have shown that the LAR provide an approximately optimal set for quantization.

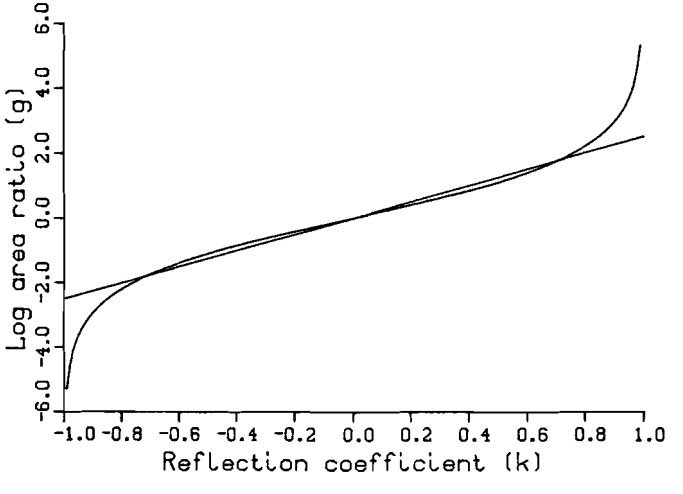

Figure 3.2: Log-area ratio (LAR) plotted as a function of the reflection coefficient (RC). For comparison the linear characteristic $g_i = 4 * k_i$ is also shown.

The following five parameter sets are derived from a spectral analysis of the speech signal.

(5) Formant frequencies (F) The F frequencies are defined as the acoustical resonances of the vocal tract. Besides their frequent usage by phoneticians, they are often used in speech synthesizers (e.g. Flanagan, 1972). To determine the F frequencies and the associated bandwidths from the LPC coefficients, we have used the robust formant analysis method of Willems (1986, 1987). This method, based on the Split Levinson Algorithm, always yields $I/2$ ordered formant tracks. After that, the optimal bandwidth values can be found from a table. Only the F frequencies will be used as input parameters, since at several stages of the temporal decomposition procedure (e.g. singular value decomposition and the computation of the target vectors) it is undesired to have a mixed set of input parameters. Bimbot et al. (1987) suggested that F frequencies are less suitable for temporal decomposition because their number is not constant and they are not always ordered. This objection, however, does not apply to the F frequencies obtained with the above-mentioned algorithm.

(6), (7), (8) Three sets of spectral coefficients (S_*) These spectral coefficients are calculated by means of a discrete Fourier transform (DFT) of the prediction coefficients a_i. The order of the Fourier transform determines the number of resulting log-amplitude coefficients (I') which describe the spectral transfer characteristic of the prediction filter. Both the prediction order and the order of the Fourier transform are varied, 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_*) are used to vary the amount of detail in the input parameters.

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

3.3.2 Time variations of the speech parameters

As temporal decomposition is based on the assumption that the speech

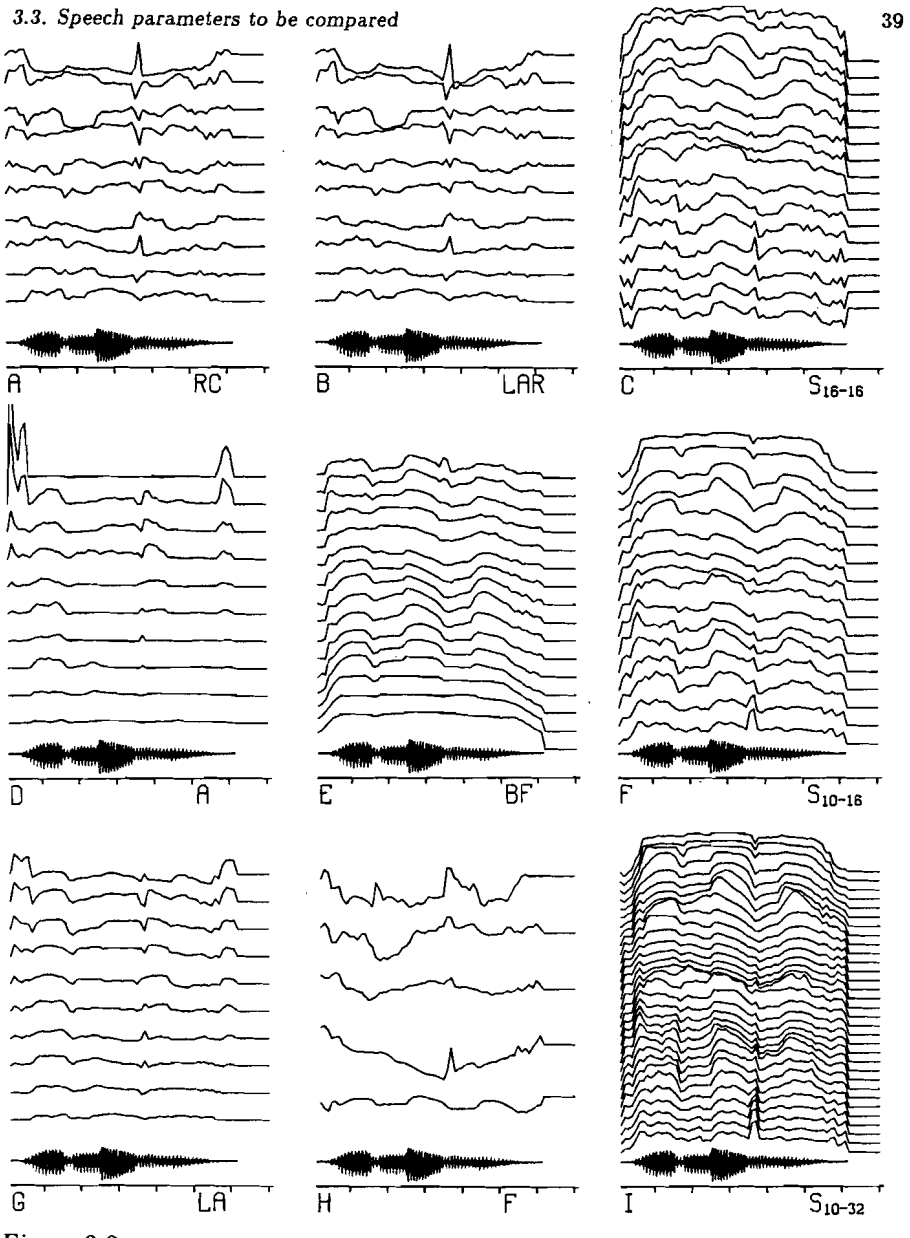

Figure 3.3: Time variations of the parameters of several sets and the waveform of the speech signal. In all nine cases the speech utterance is /dəlɔmə/. The time marks are 100 ms apart. A. reflection coefficients (RC), B. Log-area ratios (LAR), C. spectral coefficients (S_{16-16}), D. area coefficients (A), E. filter bank output parameters (BF), F. spectral coefficients (S_{10-16}), G. log-area parameters (LA), H. formant frequencies (F) and I. spectral coefficients (S_{10-32}).

parameters can be described by linear combinations of articulatory target positions, the speech parameters should somehow reflect this linearity. Therefore, there should be a high linear dependency between the time variations of the different parameters of a single set.

The time variations of the parameters of the sets used in our experiment can be seen in Fig. 3.3. Both the RC and the LAR (Fig. 3.3A and B) exhibit a capricious behaviour in time. It can be seen that their time variations are almost identical, which is explained in Fig. 3.2. Clearly, the three spectral sets S_* (Fig. 3.3C, F and I) are closely related, S_{16-16} showing more details than S_{10-16} and S_{10-32}. The BF parameters (Fig. 3.3E), and especially the low-frequency ones (upper lines), also show a resemblance to the coefficients of the S_* sets. The BF parameters, however, vary more smoothly. The time variations of the A coefficients show distinct peaks between rather long periods of almost constant value (Fig. 3.3D). The LA parameters, on the other hand, vary constantly though rather smoothly (Fig. 3.3G). Finally, in Fig. 3.3H the F frequency tracks are shown.

3.3.3 Temporal decomposition of a speech utterance

Fig. 3.4 gives an example of the decompositions of the speech utterance /dəbɑbə/ using five different sets of input parameters. From the top downwards, RC, LAR, A, BF and LA have been used. It can clearly be seen 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 the other. Consequently, the corresponding target vectors too can 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 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. 3.2). 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.

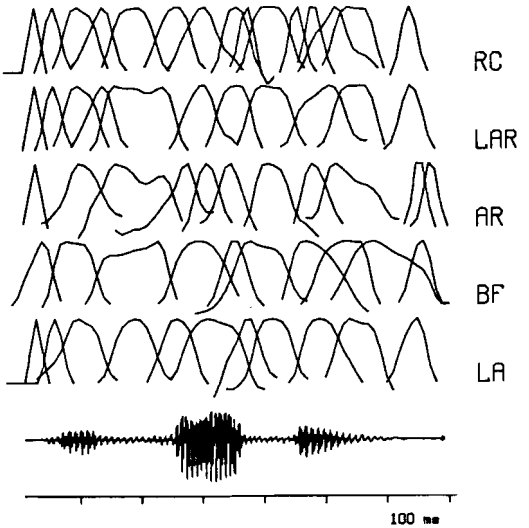

Figure 3.4: Temporal decomposition of the speech utterance /dəbəbə/ using five different sets of input parameters: reflection coefficients (RC), log-area ratios (LAR), areas (A), filter bank output parameters (BF) and log-area parameters (LA).

3.4 Phonetic relevance of the target functions

Most phonemes are realized in the acoustic speech signal as a single speech event. Clear exceptions are plosives and diphthongs, both consisting of two speech events. Though this gives a simplified view of reality, it yields a possibility to judge the phonetic relevance of the decomposition. An important criterion for good performance is a one-to-one relation between target functions and speech events. In the following, each target function will be associated with a particular speech event, and for each speech event the number of target functions associated with it will be counted.

3.4.1 Experimental procedure

In order to perform this experiment, a small database was constructed consisting of CVC combinations embedded in a neutral context: /d∂C$_1$VC$_2\partial$/. 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 /ɔ/. Each

of the 48 possible combinations was produced by a single male speaker. Of course, this material does not exhaust all possibilities of the language, but for practical reasons the size of the database had to be limited. Moreover, since much is yet unknown of the possible achievements of temporal decomposition, it is better to restrict oneself to a small set of carefully articulated words than to use fluent speech produced by various speakers. The results presented in the next chapter will legitimate this approach in which only CVC words spoken by one speaker are used.

Phonetic labelling of the CVC combinations was carried out by hand. Closure and burst of the plosives 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 speech event. As may be gathered from Fig. 3.4, this assignment was not always a straightforward matter. Sometimes a target function is located at the transition of two consecutive speech events and thus difficult to classify. As we have opted not to label transitions, in each case a decision had to be made as objectively as possible. However, since we are mainly interested in the number of target functions describing a speech event, a wrong decision will not substantially influence the results as they are averaged over all speech events. Furthermore, these transition-describing target functions appear to occur much more frequently when the overall number of target functions is also relatively high. Thus, uncertain classifications will occur specifically for parameter sets which are obviously not very suitable for temporal decomposition.

3.4.2 Results

As described above, for each speech event the number of associated target functions was counted. Next, the percentage of speech events associated with zero, one, two or more than two target functions was determined for each set of input parameters. The results are shown in Table 3.1A. The order in which the data of the several sets are presented may give an indication of the quality of the performance.

As we aim at a one-to-one correspondence of target functions to speech events, the second column gives the best indication of the performance. One can see that both BF parameters and LA parameters give the fair result that about 64 % of the speech events are associated with only one target function. At a distance of some 6 % they are followed by the sets of spectral coefficients S_* and the F frequencies. The remaining three sets, consisting

Table 3.1: Percentages of speech events associated with zero, one, two or more than two target functions. The results given are averaged over all speech events, including and excluding the bursts respectively.

	A. including bursts				B. excluding bursts			
parameters	0	1	2	>2	0	1	2	>2
BF	20	65	15	1	0	79	20	1
LA	18	63	18	1	1	73	25	1
S_{10-16}	20	58	21	2	0	70	28	2
S_{10-32}	22	56	21	1	1	70	28	1
S_{16-16}	20	52	27	2	1	61	36	2
F	14	59	23	4	0	64	31	5
A	21	47	30	2	4	54	40	3
LAR	16	43	34	7	0	46	46	9
RC	14	42	37	7	0	41	50	9

of A coefficients, LAR and RC, only got as far as some 44 %. The other columns, however, also reveal important information. A high percentage of speech events associated with precisely one target function is useless if the remaining speech events are not associated with any target function at all and thus are not detected. As can be seen in the first column, all sets show an unacceptably high percentage of missed speech events. However, since the bursts are only of very short duration and already spread out by the LPC analysis, the question arises whether these speech events can be correctly modelled by a target function which of necessity has a longer duration. It might be expected that a fair amount of them will not be detected. Thus, it is important to examine whether the high percentages in the first column of Table 3.1A can be attributed to missed bursts.

In Table 3.1B the results are shown for the same speech events but excluding the bursts of the plosives. Clearly, the overall results are much improved, all percentages in the first column showing a dramatic decrease, while all percentages in the second column are increased, as compared with the results in Table 3.1A. For most parameter sets, the percentage of missed speech events even gets as low as 0 %. Especially the relatively good parameter sets of Table 3.1A, such as BF and LA, profit from this alternative way of presenting the results, as their percentages of phonemes associated with precisely one target function increase to 79 and 73, respectively. Again, the S_* sets and the F frequencies form a middle group, while the same three sets as before lag

behind. Furthermore, a high percentage in the second column correlates with a low percentage in the fourth column, indicating that only a small number of speech events is associated with more than two target functions.

It should be noted that in Table 3.1 as well as in the next table the remaining percentages of missed speech events do not always indicate a gap in the sequence of overlapping target functions. Rather, this can be attributed to a strong coarticulation, because of which two consecutive speech events 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 in these particular cases hardly any speech signal exists.

Table 3.2: Percentages of the consonants (excluding the bursts) and the vowels associated with zero, one, two or more than two target functions.

	A. consonants				B. vowels			
parameters	0	1	2	>2	0	1	2	>2
BF	0	86	13	1	0	65	35	0
LA	2	77	20	1	0	65	35	0
S_{10-16}	0	73	24	3	0	65	35	0
S_{10-32}	2	70	27	1	0	70	30	0
S_{16-16}	1	62	35	2	2	59	37	2
F	0	70	24	7	0	52	46	2
A	4	59	35	2	2	44	50	4
LAR	0	47	45	9	0	44	48	9
RC	0	40	49	11	0	41	52	7

Given the results of Table 3.1 it will be interesting to investigate whether these results apply for all categories of phonemes. The most obvious division is that into consonants and vowels. The results for both the consonants and the vowels are shown in Table 3.2. The consonants show an increase in the percentages of the second column for almost all parameter sets. The BF parameters even get as high as 86 % of the consonants described by only one target function. Here, the LA parameters cannot match the BF parameters, although the achievement of 77 % is also relatively high. The other parameter sets follow in almost the same order as in the previous table.

From these results it can already be derived that, for the vowels, the percentages in the second column of Table 3.2B are lower than the corresponding ones in Table 3.1B. It is interesting to notice that, for the vowels, the results

of the BF parameters and the LA parameters are identical. Also the sets of S_{10-16} and S_{10-32} show similar results.

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

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. 3.2 and 3.3), the former set always yields a slightly better performance.

3.5 Phonetic relevance of target vectors

The target vectors are assumed to model articulatory target positions. It will be clear that this is only possible if for each speech event only one target function, and thus also one target vector, is found. In the previous section we have found that this is not always the case; part of the speech events of our database is associated with two or more target functions. For this reason, we will restrict ourselves in the following to evaluating only target vectors belonging to speech events associated with precisely one target function. Since this is a first exploration of the target vectors, 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 LA parameters, we will first investigate the interpretation and phonetic relevance of the target vectors determined with LA parameters. Next, we will extend or adapt our findings to the other sets of parameters.

3.5.1 Target vectors of log-area parameters

After the determination of the target functions, a target vector is computed for each function by solving Eq. (3.3). In the case of the LA parameters, the target vectors describe in fact 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 identical speech events, target vectors belonging to the same speech events should show a close resemblance. A convenient way of judging this resemblance is in terms of the first two formants.

In order to obtain more vowels associated with precisely one target function, the database was extended with two more productions of the same utterances by the same speaker. For all vowels associated with only one target function, the target vectors were transformed from the LA space to the formants and bandwidths 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. 3.5, where the target vectors belonging to an /*a*/ are represented by filled circles (●), to an /I/ by filled squares (■) and to an /ɔ/ 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. In order to better appreciate the location of these clusters, we also show, for comparison, the points belonging to the middle frames of the same vowels. These frames, which we use as a reference, were extracted by hand from the original matrix of speech parameters and thus, in contrast with the target vectors, were actually realized in the speech signal. In Fig. 3.5 the original vowels /*a*/, /I/ and /ɔ/ 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 vectors form three separate clusters. Also, the middle frames of the vowels form separate clusters which are slightly more compact. What is most important, however, is that the two clusters belonging to the same vowel do not occur at exactly the same location, although there is a fair amount of overlap. This can be seen most clearly for the vowel /ɔ/; there is not much overlap between the groups of ▲ and of △. One might argue that this is due to the fact that the target vectors represent idealized targets and are thus not necessarily realized in the acoustic speech signal. This argument

would be supported by the fact that the shift of the target-vector clusters with respect to the vowel clusters is in a direction away from a neutral vocal tube; that is, the target-vector points are more pronounced than the actually realized vowels. However, in that case one would have to expect more compact clusters, as all different realizations of the same vowel are supposed to belong to the same target.

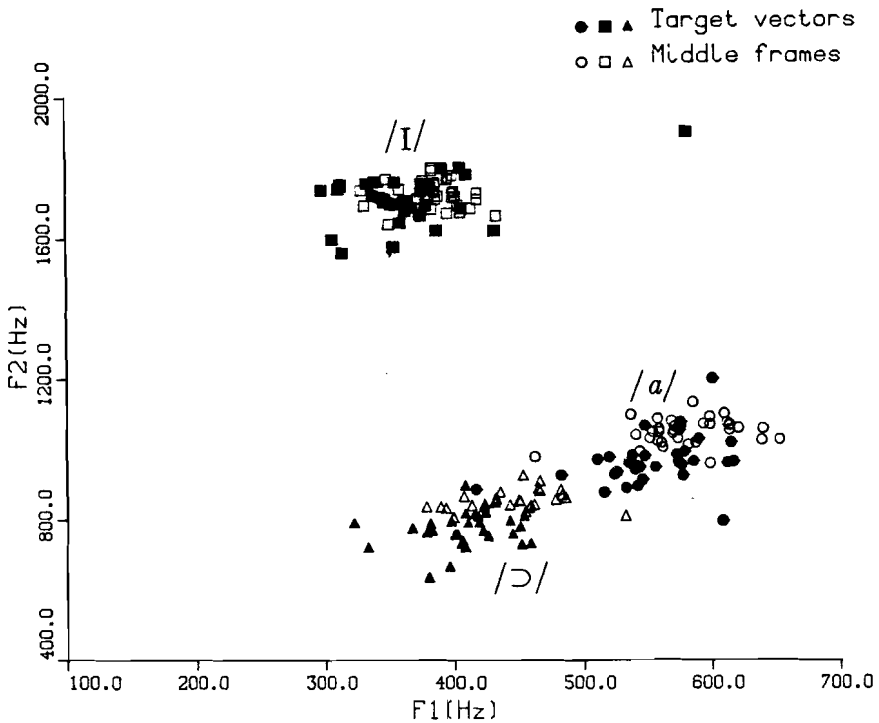

Figure 3.5: The first two formants F_1 and F_2 plotted against each other for some target vectors and some middle frames of vowels. The target vectors are associated with the short vowels /a/ (●), /I/ (■) and /ɔ/ (▲). The middle frames are taken from the same vowels: /a/ (○), /I/ (□) and /ɔ/ (△).

There is another, more plausible explanation. The $\mathbf{a}(k)$ are chosen subject to the condition that the product of $\mathbf{a}(k)$ and $\phi_k(n)$ approximates as closely as possible the original speech parameters $\mathbf{y}(n)$. Thus, the length of the target vectors is determined by both $\mathbf{y}(n)$ 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 all 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 in the following extension of Eq. (3.1):

$$\tilde{\mathbf{y}}(n) = \sum_{k=1}^{K} \mathbf{a}(k)\phi_k(n) = \sum_{k=1}^{K} (\frac{1}{x} \cdot \mathbf{a}(k))(x \cdot \phi_k(n)), \tag{3.9}$$

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. 3.6. The F_1-F_2 points corresponding to the vectors of original length are represented by filled circles (●). Increasing the length up to a factor of 2 results in the tracks from ● to □, while decreasing the length down 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 of the value of x. The value of x actually used (namely $x = 1$) is based upon the following grounds: if a target function does not have much overlap with neighbouring functions, the target can be reached. Since the input vectors at that particular place are approximated by the product of only one target function and target vector, the target vector should resemble the input vectors, so the target function has to be normalized to 1. However, in practice consecutive target functions often show a considerable overlap. In those cases it is less clear what the normalization factor should be. The results given in Figs. 3.5 and 3.6 suggest that a normalization factor of slightly more than 1 would be a better choice. The lengths of the target vectors will then be a little bit shorter, causing a shift of the target vector clusters in the direction of the original vowel clusters. It will be clear that in order to give optimal results all target vectors require different normalization factors. However, up till now it has been impossible to find boundary conditions for the temporal decomposition method which solve this problem satisfactorily. A more detailed study of a wider range of target vectors will be necessary.

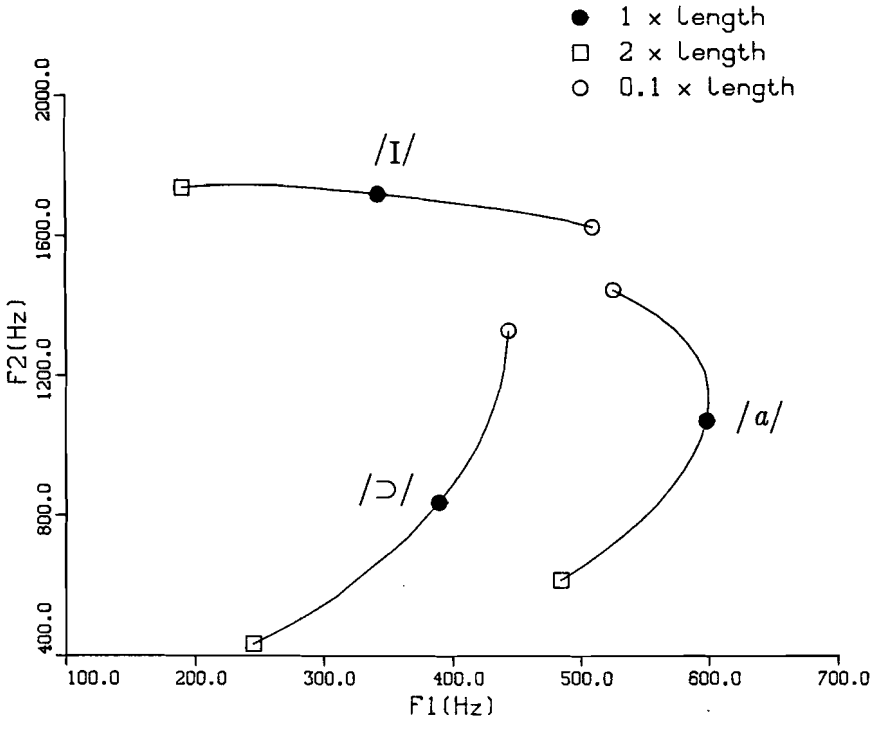

Figure 3.6: Tracks in the F_1-F_2 plane of three vectors which are changed in length in the LA space. The three vectors correspond to the three vowels /a/, /I/ and /ɔ/. The F_1-F_2 belonging to the vectors of original length are represented by the filled circles (●). If the length of the vector is doubled, F_1-F_2 take the place of the squares (□). Multiplying the length by 0.1 results in the F_1-F_2 at the places of the open circles (○).

3.5.2 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. Clearly, 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.

3.6 Resynthesis

The temporal decomposition method of Atal was originally proposed for economical speech coding. Thus, after the decomposition of the speech signal in terms of target functions and target vectors, the original speech signal has to be reconstructed and resynthesized again. Reconstructed speech parameters, approximating the original ones, can be obtained by substituting the target functions and target vectors in Eq. (3.1). 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 to test 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 speech signal starting from BF parameters, a special synthesizer, which was not available for our experiment, is needed. Furthermore, such synthesizers are known to yield, in general, unsatisfactory results (e.g. Pols, 1977). The S_* coefficients can only be used for resynthesis if phase information is also available, but this is lost during the several stages of the analysis. Finally, among the reconstructed speech parameters of RC, A coefficients and F frequencies unphysical values sometimes occur. Such unphysical values would have to be corrected before they could be passed on to a speech synthesizer. Thus, only the LA parameters 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 errors signals are computed in the same parameter space. Since not all sets

are transformable into one another, only LA parameters, A coefficients, LAR and RC can be compared in this way. The following section will deal with this.

3.6.1 Reconstruction errors in the resynthesis

The LA space is used as a reference; all the reconstructed speech parameters are transformed into 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 distance measure would be the most appropriate, but although many attempts have been made (e.g. Gray and Markel, 1976; Nocerino, Soong, Rabiner and Klatt, 1985; Applebaum, Hanson and Wakita, 1987), the definition of such a distance measure does not exist yet. Therefore, we confine ourselves to a simple Euclidian distance measure, the same which is minimized in Eq. (3.3) for the determination of the target vectors. The reconstruction error $E(n)$ for one particular frame is defined as:

$$E(n) = \left[\sum_{i=1}^{I}(y_i(n) - \tilde{y}_i(n))^2\right]^{\frac{1}{2}}. \tag{3.10}$$

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 target 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:

$$E_m = \sum_{n=10}^{50} E(n). \tag{3.11}$$

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 get a perceptually more relevant error measure, it is possible to weight the reconstruction 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:

$$E_A = \sum_{n=10}^{50} G(n)E(n)/1000. \tag{3.12}$$

The factor 1000 is only meant to bring the values 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. 3.7, a representative example can be seen of the decompositions of the utterance /dəpɑlə/, using four different parameter sets: RC, LAR, A and LA. Like Fig. 3.4, this figure shows the differences in number, location and form of the target functions. Next to the target functions the reconstruction error $E(n)$ is shown for each frame. The vertical bars under the error signal of the A coefficients 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.

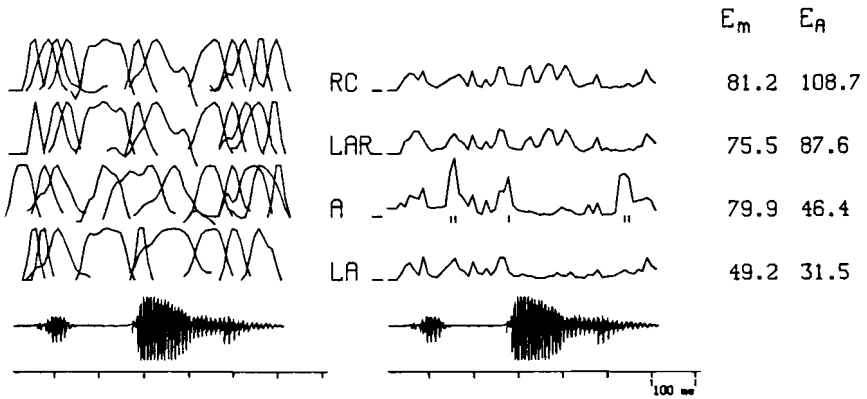

Figure 3.7: Target functions (left) belonging to the reflection coefficients (RC), the log-area ratios (LAR), the areas (A) and the log areas (LA), and the reconstruction error signal (right) of the CVC utterance /dəpɑlə/ for each of the parameter sets. A further explanation of this figure is given in the text.

Temporal decomposition attempts to describe the 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 of Fig. 3.7 it can be seen that the error signal of the A coefficients shows considerable

peaks, confirming once more that the A coefficients are not very convenient parameters for temporal decomposition. Although none of the other parameter sets produces a consistently flat error signal, the achievements of the LA parameters are most 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, the error signals of the RC and the LAR are very much alike in this example.

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 reconstructed LAR always describe the speech signal better than the RC, and of the four sets the A coefficients usually perform worst (although this is not visible in the particular example of Fig. 3.7).

3.6.2 Reconstruction using mixed parameter spaces

In the previous section error signals of reconstructed speech parameters were compared. In all cases, the decompositions (i.e. the number and location of the target functions) were different. The LA parameters nearly always yielded the smallest reconstruction error. However, our temporal decomposition method was optimized using LA parameters, which might have influenced the results. Since the target functions are dimensionless, it is possible to use them in another parameter space than the one in which they have been determined. This strategy makes it possible to investigate whether the target functions determined in LA space are also suitable for use in other parameter spaces. In this way, the comparison is more direct, irrespective of the optimization of the temporal decomposition method. On the other hand, it also offers the opportunity to carry out this procedure the other way round: using target functions, determined in other parameter spaces, for reconstruction in the LA space. Since the other parameter sets mostly yielded more target functions, it is interesting to investigate whether the LA reconstruction error further reduces if more target functions (determined in another parameter space) are used.

An example of the first procedure can be seen in Fig. 3.8A. The LA target functions have been used to determine the target vectors in the various parameter spaces. Subsequently, the speech parameters have been reconstructed and transformed to the LA space. The resulting reconstruction error signals have been plotted for each parameter set.

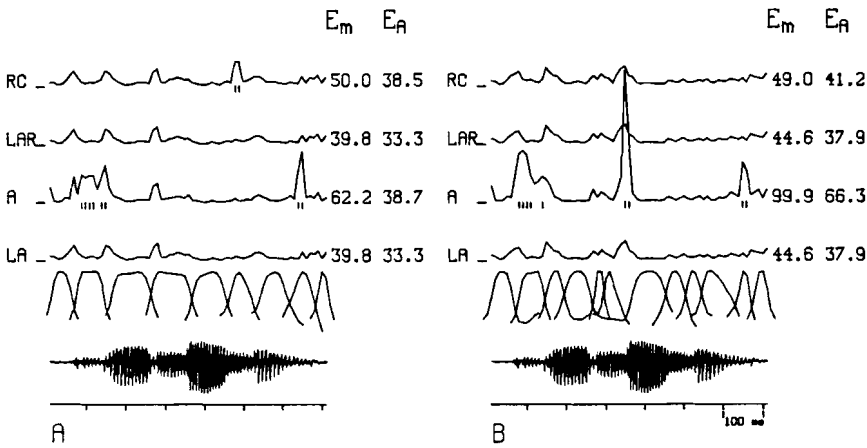

Figure 3.8: A. Target functions of the utterance /dəlɔlə/ determined in the LA space. Using these functions, the optimal target vectors are computed in the RC, LAR, A and LA space, yielding the plotted error signals. B. Same as A., only in this case the target functions are determined in the RC space.

The fact that the error signals of the LA parameters and the LAR are identical is 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 for the computation and interpretation of the target vectors these two spaces are equivalent. Moreover, in the previous section it has already been said that the error signal of the LAR is nearly always larger than that of the LA, even if the description in the LA space consists of more target functions. It follows that for the LAR a better reconstruction of the speech parameters can be obtained with a smaller number of target functions.

Since in all cases in Fig. 3.8A identical target functions are used, it is possible to compare directly the error signals of the RC and the LAR. Again, the error of the RC is the larger of the two. This is mainly due to the occurrence of unphysical values. Also, the fact that these coefficients differ significantly in the region $0.8 \leq |RC_i| < 1$ (see Fig. 3.2) plays a role here.

Another example of the decomposition of the same utterance /dəlɔlə/ following the second procedure is given in Fig. 3.8B. This time the target functions shown are derived in the RC space, yielding considerably more target

functions than in Fig. 3.8A. Again the computation of the target vectors and the reconstruction of the speech parameters have taken place in the various parameter spaces. Now, the error signals in Fig. 3.8A and B can be compared. For the LA parameters it can be seen that the error measures E_m and E_A are higher if the RC functions are used. Since the LA and LAR error signals are identical, this also holds for the LAR. Also the error signal of the A coefficients is smaller when the LA functions are used. In this particular example the RC error signal is smaller if RC functions are used, but quite often the opposite is true, which signifies that even in RC space the RC functions are not always optimal.

These examples show that indeed the optimization of the temporal decomposition method for the LA parameters has its impact on the performance of the other parameter sets. Still, it can be concluded that the LA parameters are more suitable for temporal decomposition than RC or A coefficients, since the error signal is always smallest for the LA parameters. The LAR, on the other hand, must have, in principle, the same possibilities as the LA parameters.

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 to also achieve better results here, a final option we have examined is the use of BF target functions for reconstruction in the LA space. Although this indeed sometimes led to better descriptions of the original speech signal, more often the error signals obtained were significantly larger. As it is impossible to transform the reconstructed BF parameters to the LA space we have not been able to compare the error signals of both spaces.

3.7 Discussion and conclusions

In this chapter it has become clear that every speech utterance can be decomposed in a large number of ways by simply varying the choice of input parameters (see e.g. Fig. 3.4). Variation of the other parameter settings of the method will lead to even more possible decompositions. The criterion for determining the best possible decomposition when optimizing the method was its phonetic relevance. Here, it was shown that speech parameters which yield a phonetically relevant decomposition also give the best reconstruction results, even compared to decompositions which yielded more target functions. Thus, it can be concluded that indeed the optimized temporal decom-

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

A possible explanation of the differences between the results of the LA parameters and the BF parameters lies in the fact that in the latter set amplitude information is integrated in the parameters. In the LA parameters and the other LPC-derived parameters the amplitude 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 parameters were weighted with the gain factor to obtain a situation comparable with the BF parameters, did not yield any better LA results.

The three sets of spectral coefficients were included 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 precisely one target function, formed a cluster of points which, compared with 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. Nor, moreover, do they represent the actually realized speech event, 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 a 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 with the other sets the reconstruction of the signal was achieved with the fewest number of target functions and target vectors. Better results were also 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 parameters. 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 coefficients performed worse than the LA parameters, also when using the LA target functions. This can be understood from their logarithmic relationship; if LA parameters are suitable for linear modelling, as a consequence the A coefficients will not be suitable.

In recent literature temporal decomposition results are reported using the LAR (Ahlbom et al., 1987; Bimbot et al., 1987; Chollet et al., 1986; Marteau et al., 1988; Niranjan et al., 1987). Although this is not in direct accordance with our results (Table 3.1), the resynthesis experiments of section 3.6.2 have made it clear that, in principle, the same results can be obtained with the LAR as with the LA parameters. However, in their experiments the target vectors are assumed to be known, leaving only the target functions to be determined. As we have shown 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, using temporal decomposition in this way, the LAR and the LA parameters will perform equally well, independent of the way in which the temporal decomposition method is optimized.

Appendix

In this appendix it will be proved that, using a fixed set of target functions and the same speech utterance, 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

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

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

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

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

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

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

$$\sum_{k=1}^{K} a'_{ik} \sum_n \phi_k(n)\phi_r(n) = \sum_n (y_i(n) - y_{i+1}(n))\phi_r(n) \tag{A6}$$

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

Substitution of Eq. (A4) in the right-hand terms of Eq. (A7) gives:

$$\sum_{k=1}^{K} a'_{ik} \sum_n \phi_k(n)\phi_r(n) = \sum_{k=1}^{K} a_{ik} \sum_n \phi_k(n)\phi_r(n) - \sum_{k=1}^{K} a_{(i+1)k} \sum_n \phi_k(n)\phi_r(n) \tag{A8}$$

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

which subsequently leads to:

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

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.

Chapter 4

Some further explorations of temporal decomposition

4.1 Introduction

The exploratory experiments described in this thesis are aimed at getting an impression of the possibilities of temporal decomposition with respect to the derivation of phonetic information directly from the acoustic speech signal, without making use of phonetic information. In the previous chapters the optimization of the method has been the central theme; in this chapter, the optimized temporal decomposition method thus obtained will be further explored.

An important aspect which received only little attention in Chapter 3 is the speech quality of the resynthesis. Although we do not aim to use temporal decomposition for speech coding, it is still useful to investigate how well phonetic information is preserved after temporal decomposition. Phonetic information that gets lost in the process is apparently not modelled correctly by the target functions and target vectors. Such knowledge increases the insight into what kind of phonetic information might be derived from the acoustic speech signal by means of temporal decomposition. In Chapter 3 it was already shown that reconstruction errors do occur, and thus the speech signal cannot be reconstructed perfectly. However, an intelligibility experiment is needed to determine what impact such errors have on the phonetic content of the speech signal. A perception experiment will be described in which the intelligibility of temporally decomposed and resynthesized CVC utterances is compared with that of LPC utterances.

In the previous chapters the phonetic relevance of the target functions has already been under discussion. In Chapter 2 a small database was used to optimize the parameter values of the method and to compare the performance

of our weighting factor with the original weighting factor of Atal. The object was always to achieve the best possible correspondence between target functions and phonemes. Except for the plosives, phonemes were supposed to be realized as single speech events, which is, of course, a simplified view of reality. As a performance criterion, however, it sufficed. In this chapter, the phonetic relevance of the target functions was investigated, using a much larger, more realistic database consisting of 100 phonologically balanced sentences. Possibly, the experiment may lead to suggestions for further improvement of the method.

The phonetic relevance of the target functions was judged in two ways: phonologically and phonetically. The phonological approach starts from a phonetic transcription of the sentences. Subsequently, for each phoneme it is determined how many target functions are associated with it. The ideal outcome would be one target function for each phoneme. A disadvantage of this approach is that in fluent speech not all phonemes will be actually realized in the acoustic speech signal. In such cases, it cannot be expected that a separate target function will be found. However, since recognition or segmentation techniques often concentrate on the phoneme level, it is important to examine to what extent temporal decomposition could be used as a phoneme detector.

In previous chapters it has become clear that the "ideal" result is far from being reached; phonemes are often associated with more than one target function. The phonetic approach of this experiment is directed towards the investigation of the question as to why a phoneme is associated with a particular number of target functions. The expectation is that each speech segment that is perceptually distinct from neighbouring speech segments is associated with a separate target function. A perceptual analysis is set up to verify this expectation.

In the experiments described in this chapter, the temporal decomposition method as proposed in Chapter 2 was used. To be able to combine the results of the experiments, log-area parameters were used as input. The filter bank output parameters, which gave the best results with respect to the phonetic relevance of the target functions, could not be used since a suitable speech synthesizer was not available. The results of the experiments should give an impression as to whether applications of temporal decomposition for speech segmentation or speech recognition might be feasible.

4.2 Intelligibility of temporally decomposed and resynthesized CVC utterances

4.2.1 Introduction

The intelligibility experiments are set up to investigate the influence which temporal decomposition has on the phonetic content of speech utterances. This investigation was done in two ways. In the first place, the intelligibility of temporally decomposed and resynthesized speech, the so-called TD speech, was compared with that of LPC speech. By comparing these two speech types, the influence of temporal decomposition on the intelligibility is measured directly, since LPC-derived coefficients (log-area parameters) are used as input for the method. LPC speech represents the speech quality immediately before temporal decomposition. The LPC results were taken from an earlier but otherwise identical experiment; that is, the same subjects, the same kind of stimuli and the same tasks for the subjects (Eggen, 1987b).

In the second place, the intelligibility of three different types of TD speech (TD1, TD2 and TD3) was compared. These three speech types differ in the accuracy with which the target functions and target vectors are described. In the temporal decomposition process much effort goes into the accurate determination of target functions. It is, however, unknown how much accuracy is needed for correct modelling of phonetic information. In TD1 the description of target functions and target vectors is very accurate. In TD2, both target functions and target vectors are quantized. In TD3 the target functions are replaced by stylized functions, while the target vectors are further quantized. The intelligibility of TD2 was also measured in the experiment from which the LPC results were taken (Eggen, 1987b), so that the results of both experiments could be related.

4.2.2 Method

4.2.2.1 Speech material and stimulus preparation

In the experiments the stimuli consisted of CVC utterances. CVC combinations are often used for such experiments, since the intelligibility is then measured at a segmental level, excluding syntactic or semantic influences. An extra advantage of such stimuli is that they allow automatic processing of the responses of the subjects.

The words consisted of CVC sequences which are phonotactically possible

in Dutch (e.g. Cohen, Ebeling, Fokkema and Van Holk, 1961; Moulton, 1962). All words were easy to pronounce and sounded natural. The initial consonants were chosen from the set (/p/, /b/, /t/, /d/, /k/, /f/, /v/, /s/, /z/, /x/, /m/, /n/, /l/, /r/, /j/, /w/, /h/), the vowels from the set (/ɑ/, /ɛ/, /I/, /ɔ/, /œ/, /a/, /e/, /i/, /o/, /y/, /u/, /ø/, /ɛɪ/, /ʌy/, /au/) and the final consonants from the set (/p/, /t/, /k/, /f/, /s/, /x/, /m/, /n/, /ŋ/, /l/, /r/, /j/, /w/) (IPA notation).

Stimulus lists consisting of 50 CVC words were generated by a computer program. Each phoneme allowed in a certain position appeared approximately an equal number of times in the CVC words. This was realized by drawing at random, without replacement, a phoneme from the appropriate set. The complete set was replaced when all the phonemes of a set were drawn. The so-obtained lists included words as well as non-words. Excluding the meaningful CVC words from the test sets would have provided the subjects with an additional cue and would have restricted the open response set. In a similar experiment Pols and Olive (1983) did not find any indication that the identification scores for words deviated from those of non-words. Eggen (1987b) has found that the different lists are equivalent, that is, the overall results do not depend on the list.

For each of the speech types a different 50-word stimulus list was generated. These lists were recorded and the four different speech types were created in accordance with the principles described in the next section. Additional lists were prepared for training purposes.

The CVC words used were all spoken by one male Dutch speaker. The quality of his speech after LPC analysis/resynthesis was judged to be good. To ensure that the words were recorded under similar conditions, the words were spoken in isolation at the end of a neutral carrier sentence (e.g. Het woord is ... "bak" (The word is ... "bak"), where the dots indicate a short silence). From our experience, there was no reason to expect that the temporal decomposition results would be speaker-dependent (see also section 4.4). The words were recorded on a digital audio recorder and, after low-pass filtering at 5 kHz, stored on disk using a 12-bit AD converter and a sampling rate of 10 kHz (Eggen, 1987c). Each word was stored in a separate file.

4.2.2.2 Speech types used in the experiment

LPC The LPC speech was synthesized directly from the ten coefficients of a linear predictive analysis. Pitch and voiced/unvoiced parameters were

estimated using the pitch detection algorithm developed by Hermes (1988). The bit rate of this speech type was approximately 12 kbit/s, viz. $\simeq$ 10 kbit/s for the filter parameters and $\simeq$ 2 kbit/s for the source parameters.

TD1 The ten prediction coefficients from the LPC analysis were transformed into log-area parameters, which were subsequently used as input for the temporal decomposition method. The resulting target functions and target vectors were filed with a sufficient accuracy (both with three decimal places), together with their locations in time. For resynthesis these target functions and target vectors were recombined using equation (2.1) and the so-obtained speech parameters were transformed back again into prediction coefficients. Resynthesis of the filter coefficients together with the original source parameters (pitch, gain and voiced/unvoiced parameter) yielded the TD1 speech. Since the number and lengths of the target functions differ for each speech utterance, the bit rate of this speech type can only be indicated as an average. Based on the analysis of the utterances used in this experiment, the bit rate was on the average $\simeq$ 6.7 kbit/s, viz. $\simeq$ 2.0 kbit/s for the source parameters and $\simeq$ 4.7 kbit/s for the filter coefficients. A more detailed overview of the distribution of the bits is given in Table 4.1.

TD2 The procedure for preparing the TD2 speech was much the same as for the TD1 speech. In this case, however, an additional step was executed before filing. Low-valued (< 0.1) sidelobes of the target functions $\phi(n)$ were truncated and, as a consequence, the average length of the functions was shortened. The values of $\phi(n)$ ($0.1 \leq \phi(n) \leq 1.0$) were specified with one decimal place only. Furthermore, the accuracy with which the individual parameters of the target vectors were stored ($-4.00 \leq a_i \leq 6.00$) was decreased to two decimal places (see also Table 4.1). For resynthesis the original source parameters were used again, resulting in the $\simeq$ 4.4 kbit/s TD2 speech.

TD3 The TD3 speech was the result of an attempt to find an efficient stylization of the target functions. It was found that the target functions could be approximated fairly well with an exponential (i.e. Gaussian) function:

$$\phi(n) = \mathrm{e}^{\frac{-\frac{1}{2}(n-\mu)^2}{\sigma^2}} \tag{A1}$$

where μ represents the location of the maximum of the target function and σ is a measure of its width. It also turned out that 8 different values of σ were sufficient to cover the range of all possible target functions, and thus

the number of bits required to encode each target function was reduced to only 3. The parameters of the target vectors were separately quantized to 32 levels, that is 5 bits for each parameter. Finally, the source parameters were logarithmically quantized to 16 levels. This resulted in a bit rate of $\simeq$ 1.8 kbit/s for the TD3 speech (see also Table 4.1).

Table 4.1: Overview of the estimated bit distribution in kbit/s over source and filter parameters for LPC speech and the three TD-speech types.

		LPC	TD1	TD2	TD3
Source	Gain	1.2	1.2	1.2	0.4
	F_0 and V/UV	0.8	0.8	0.8	0.5
Filter	Filter coefficients	10.0			
	Target function		2.4	0.7	0.05
	Target vector		2.6	1.6	0.8
	Location		0.1	0.1	0.1
Total		12.0	6.7	4.4	1.8

4.2.2.3 Listening experiment

The above-mentioned CVC utterances were used in an identification experiment. Eight subjects, all native speakers of Dutch, participated as subjects in this experiment. All reported normal hearing and were familiar with listening experiments and with the kind of speech stimuli used in this particular experiment. The stimuli were presented to the subject over headphones. The subjects, one at a time, had to respond by typing the perceived CVC combination followed by the return key on the keyboard of a computer terminal. As the response time was unlimited, the subject was not bothered by any time constraints. After the return key was pressed, the next stimulus was presented. The response of the subject was filed and evaluated automatically by the software especially designed for these experiments (Eggen, 1987a).

The experimental sessions always started with a set of training stimuli to make the subjects familiar with their specific task and to test whether they had understood the written instructions. Next, three lists with CVC words of TD speech were presented in varying order to the subjects. A complete session took about half an hour. The LPC speech had been tested in an identical experiment performed about six months earlier by Eggen (1987a, 1987b). TD2 speech was tested in both Eggen's and in the present experiment.

4.2.3 Results

A one-way analysis of variance on the arcsine-transformed percentages (Studebaker, 1985) showed no significant differences in terms of percentages correct phoneme identification between the results of the two experiments in which the TD2 speech was tested ($F_{(1,14)} = 3.33$; $p > 0.05$). This is not surprising, since in both cases the experimental conditions, including the subjects, were identical. We concluded that we can relate the scores of the second experiment directly to the LPC scores of the first experiment.

4.2.3.1 Overall identification scores

In Fig. 4.1 the percentages of the phonemes correctly identified are shown for the four different speech types averaged over all eight subjects. A distinction is made between the scores of initial consonants, vowels, and final consonants, since Pols and Olive (1983) showed that for some consonants these scores differed significantly. A two-way analysis of variance was performed on the arcsine-transformed percentages. There was a significant effect of speech type ($F_{(3,21)} = 19.98$; $p < 0.001$), and of phoneme type (initial consonants, vowels and final consonants) ($F_{(2,14)} = 67.14$; $p < 0.001$). Also, there was a significant interaction between speech type and phoneme type ($F_{(6,42)} = 13.22$; $p < 0.001$).

As the LPC coefficients are used as input for temporal decomposition, it was assumed that the LPC scores form an upper limit for the scores of the other speech types. Indeed, a post hoc Student-Newman-Keuls multiple range test (SNK test) with a 0.05 level of significance revealed that if scores of TD-speech types were higher than the equivalent LPC-speech scores (in Fig. 4.1), these differences were not significant.

The speech types are most clearly distinguished in the case of the initial consonants. Here, the SNK test yielded three significantly different subsets: 1 LPC, 2 TD1 and 3 TD2, TD3. For the vowels the TD3 score was significantly less than the LPC score. Also, TD2 and TD3 scored significantly less than TD1. For the final consonants the distinctions are less clear: only the difference between TD2 and TD3 is significant; it should be noted, however, that the score of TD3 is higher than that of TD2!

For three speech types, LPC, TD1, and TD2, vowels are much better identified than both initial and final consonants. Only in the case of TD3 speech is there no significant difference between the vowel and final consonant scores. For LPC stimuli Van Bezooijen and Pols (1987) reported similar results.

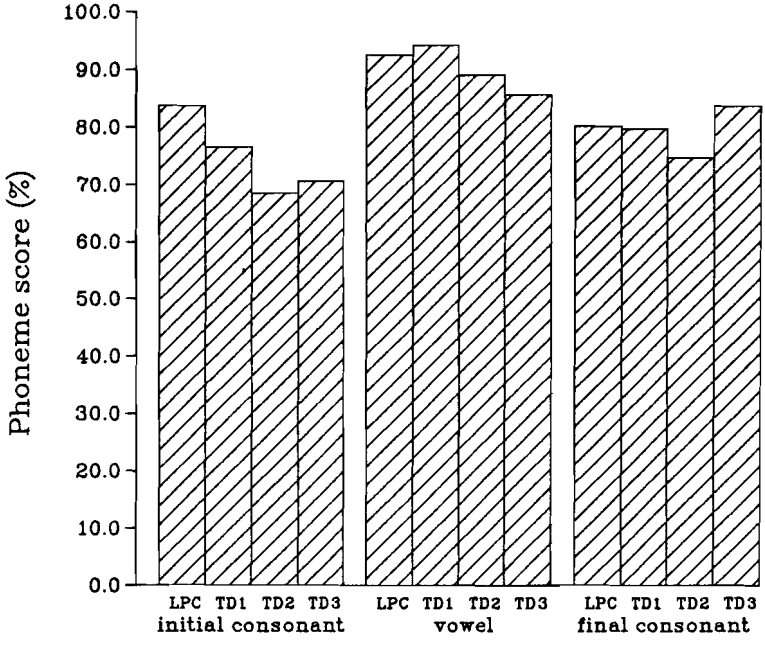

Figure 4.1: Percentages correct phoneme identification for the four different speech types averaged over all subjects. The results are given for the initial consonants, the vowels and the final consonants, respectively.

These results are very reassuring, since only in the case of the initial consonants should the decrease in intelligibility results be attributed to temporal decomposition. In both the other cases, the TD1 score does not differ significantly from the LPC scores and thus the lower recognition percentages of the other TD-speech types have to be a result of the less accurate description of the target functions and target vectors rather than of the temporal decomposition itself.

4.2.3.2 Identification scores for phoneme classes

In the previous section it was shown that only in the case of initial consonants does the TD1 score differ significantly from the LPC score. Since this experiment is aimed at investigating the influence temporal decomposition which has on the phonetic content of the speech signal, it is useful to split up the

results for several phoneme classes that share important phonetic features. It might well be that the decrease in intelligibility can be attributed to only a subset of the phonemes. For the same reason, the identification scores of TD2 and TD3 will also be presented in terms of phoneme classes.

In Table 4.2 the identification scores of initial consonants are given, averaged over phoneme classes, namely plosives, fricatives, nasals and a class consisting of the remaining consonants /h,j,l,r,w/, for each of the four speech types. The numbers between brackets indicate the phoneme class identification score, that is the score when confusions within the class are counted as correct. For instance, a /b/ perceived as /p/ is considered correct in the case of phoneme class identification score.

Table 4.2: Percentages correct identification of initial consonants averaged over phoneme classes for the four different speech types. The scores between brackets give the scores if confusions within the class are counted correct.

	LPC	TD1	TD2	TD3
Plosives (b,d,k,p,t)	73 (*86*)	63 (*92*)	53 (*88*)	60 (*91*)
Fricatives (f,s,v,x,z)	87 (*95*)	95 (*100*)	87 (*99*)	89 (*98*)
Nasals (m,n)	92 (*98*)	54 (*67*)	42 (*46*)	31 (*42*)
Remainder (h,j,l,r,w)	88 (*90*)	80 (*88*)	77 (*87*)	78 (*90*)
Total	84 (*91*)	77 (*90*)	69 (*86*)	71 (*87*)

By comparing the LPC and TD1 scores of Table 4.2 it can be seen that especially the nasals are strongly affected by temporal decomposition. Quantization and stylization of the target functions and target vectors impair the intelligibility of the nasals even further. From the phoneme class identification scores it can be derived that these consonants probably lose their nasality as well, since they are confused with all kinds of other consonants. Also the plosives are more affected than other consonants, be it less strongly than the nasals. However, it follows from the phoneme class scores that the characteristic features of the plosives remain intact. The fricative scores are relatively high, and they are only seldom confused with consonants other than fricatives. This also holds for the quantized and stylized speech types TD2 and TD3.

The recognition rates for most vowels are very high, approaching the values maximally reachable by high-quality speech (Eggen, 1987c). As a consequence of this low confusion rate, subdivision into special classes would not provide useful information.

For the sake of completeness, in Table 4.3 the identification scores of the

final consonants are given, again in terms of phoneme classes. These classes differ somewhat from those in Table 4.2, since in Dutch the voiced plosives and fricatives, and the phoneme /h/, cannot occur in word final position. On the other hand, the class of the nasals is extended with the phoneme /η/.

Table 4.3: Percentages correct identification of the final consonants averaged over phoneme classes for the four different speech types. The scores between brackets give the scores if confusions within the class are counted correct.

	LPC	TD1	TD2	TD3
Plosives (k,p,t)	95 (*100*)	92 (*100*)	91 (*100*)	93 (*100*)
Fricatives (f,s,x)	91 (*97*)	92 (*98*)	87 (*100*)	99 (*99*)
Nasals (m,n,η)	45 (*67*)	43 (*74*)	23 (*68*)	45 (*74*)
Remainder (j,l,r,w)	88 (*99*)	87 (*99*)	83 (*99*)	96 (*100*)
Total	80 (*91*)	80 (*93*)	75 (*95*)	84 (*93*)

For all speech types, the final plosives are much better perceived than the initial plosives. On the other hand, the identification scores of final nasals are much lower than the equivalent scores of initial nasals. In this case, the intelligibility of the LPC nasals is decreased as well. Similar results are reported by Pols and Olive (1983). but they also report that this does not occur for high-quality speech. The very low nasals score of TD2 is apparently the main reason for the significant difference with the TD3 intelligibility. Except for the nasals, confusions take place almost exclusively within the classes. Thus, for all speech types the important phonetic features remain intact.

4.2.4 Discussion

4.2.4.1 Influence of temporal decomposition on intelligibility

The major aim of this perception experiment was to investigate whether the phonetic content of the speech signal is affected by temporal decomposition. Comparison of LPC and TD1 utterances revealed that only TD1 initial consonants have significantly lower scores. More specifically, only initial nasals and to a lesser extent initial plosives too can be held responsible for this decrease in intelligibility. However, the phonetic features of the plosives remained intact. This followed from the fact that wrongly perceived plosives were mainly confused with other plosives. The nasals, on the other hand, were affected more seriously, the phoneme class score also being relatively low. Of course, compared to other phoneme classes, the nasal class is very

small, which restricts the within-class confusion possibilities. Still, it can be concluded that apparently nasality, the feature of this phoneme class, is not modelled correctly by the target functions and target vectors.

It follows that only in the case of initial /m/ and /n/ does temporal decomposition have a negative influence on phoneme features; for all other phoneme classes, the important phoneme features are not affected. Thus, CVC words can be modelled quite adequately by means of target functions and target vectors. This result in itself does not give any guarantee that it is possible to use temporal decomposition for the derivation of phonetic information from the acoustic speech signal. However, what is of more importance here is that a fair impression is obtained of the kind of phonetic information that could be extracted by means of temporal decomposition.

It remains to be seen whether the intelligibility results, and the conclusion that CVC words can be modelled by means of target functions and target vectors also hold for fluent speech. In general, contextual influences make the identification of fluent speech easier than that of isolated CVC words, so probably the results may be extended to fluent speech.

4.2.4.2 Influence of quantization and stylization

The perception experiment showed that quantization or stylization of the target functions and target vectors diminishes the intelligibility of initial consonants and vowels significantly. This effect is strongest for those phoneme classes where TD1 also differs significantly from LPC, namely nasals and plosives; the other phoneme classes are hardly affected. However, the phoneme class identification scores do not differ widely between the three TD speech types.

This result implies that in the temporal decomposition method the determination of the precise shape of the target functions can be simplified. Possibly the convergence criteria in the temporal decomposition method could be adjusted, although this must not be done at the expense of the determination of the optimal location and length of the target function. In another experiment, an attempt was made to substitute an exponential weighting factor for the rectangular one (Van Dijk-Kappers, 1988). Although it was hoped that better decomposition results could be achieved by using a more realistic measure, the results did not differ significantly.

Eggen (1987b) reported that the intelligibility of TD2 speech (in his terms TDC) is comparable with that of other 4 kbit/s speech types. It was already suggested by Atal (1983), however, that the strength of temporal decom-

position lies in the range of even more economical speech descriptions. Indeed, the intelligibility of TD3 ($\simeq$ 1.8 kbit/s) is comparable with that of TD2 ($\simeq$ 4.4 kbit/s). Thus, the experiments presented here also confirm that temporal decomposition is suitable for economical speech coding. An even more economical speech type could be obtained by using a code book of quantized target vectors.

4.3 Phonetic relevance of the target functions

4.3.1 Introduction

In the previous chapters the phonetic relevance of the target functions has only been used to improve the temporal decomposition method, or to compare the performance of different parameter settings. For these purposes a limited set of pseudo-CVC utterances sufficed. However, the actual performance of the method after optimization must be tested on a much larger database. The experiment described here is intended to get a fair impression of the phonetic relevance of the target functions. The results may lead to suggestions for further improvement of the temporal decomposition method.

For this experiment a database consisting of 100 German sentences was used. In several respects this speech material differs from the material used up till now: it consists of fluent speech instead of isolated CVC words, the speaker was native German instead of native Dutch, and in this database the occurrence of phonemes is phonologically balanced. It will be clear that temporal decomposition gets a chance to prove its robustness.

The phonetic relevance of the target functions was judged both phonologically and phonetically. In the phonological approach it was determined how far temporal decomposition can be seen as a phoneme detector. In this respect, the ideal outcome would be one target function for each phonologically transcribed phoneme. However, given the results of previous chapters, it was to be expected that the actual outcome would probably deviate substantially from this ideal result. In the phonetic approach, attention is paid to the question of why phonemes are associated with zero, one, two or even more target functions. In this approach, perceptual criteria are used.

The results are averaged over phoneme classes which share important features. These classes are the same as those used in the above-described intelligibility experiments, in the hope that the results of both experiments can be related.

4.3.2 Method

4.3.2.1 Speech material

For our experiments a database was available consisting of 100 German sentences spoken by a native male speaker. The sentences were phonologically balanced, which means that the frequency of occurrence of each phoneme approximated that of spoken German.

4.3.2.2 Analysis

Starting from a phonological transcription, the 100 sentences were phonetically labelled and segmented by hand by an experienced phonetician. Use was made of the original digitized speech signals. The phoneme boundaries (i.e. the perceptually determined best locations for a boundary between the acoustic realizations of two adjacent phonemes) were stored in a file. After that, LPC analysis and temporal decomposition were carried out automatically for all 100 sentences. For each sentence a plot was produced in which the waveform, the target functions (i.e. the decomposition) and the stored phoneme boundaries were displayed.

4.3.2.3 Experimental procedure

In the phonological part of the experiment, a phonetician determined for each target function to which phoneme it was associated. Making use of the indicated phoneme boundaries this was in most cases a rather straightforward process. In a few dubious cases, for instance when a target function was located around a phoneme boundary, most weight was given to the location of the maximum of the target function. Since as yet no attempt is made to identify the speech units with which the target functions are associated, no use was made in this process of phonetic information in the target vectors.

Statistics were collected of the number of times each phoneme was associated with zero, one, two or more target functions. The results were averaged over broad phonetic classes, namely vowels, diphthongs, plosives, nasals, fricatives and a class consisting of /l,r,j,w,h/.

For the phonetic part of the experiment a subset of 30, also phonologically balanced, German sentences was used. In this case the LPC version of the sentences was used rather than the original digizited speech signal, since LPC coefficients are used as input for temporal decomposition. To judge the validity of a decomposition, perceptual criteria were used. In general,

the judgement criterion was that each perceptually distinct speech segment should be associated with a separate target function. In this process, a gating technique was used to quickly compare small portions (10, 30 or 50 ms) of the same phoneme. Such a technique is considered to be an efficient way of listening to phonetic details (e.g. 't Hart and Cohen, 1964). The indicated phoneme boundaries still served as an aid in the process, but they were viewed more critically. This was necessary since these boundaries were placed by listening to the original digitized speech signal (PCM), whereas temporal decomposition works on LPC speech, which is somewhat degraded compared to the original. Because of this degradation some of the boundaries become blurred, and, as a consequence, the decision as to which phoneme a target function should be associated had to be changed sometimes. By careful listening it was decided by one, and in half of the sentences by two, phoneticians how many perceptually distinct events could be distinguished in the acoustic realization of a given phoneme. In dubious cases, formant tracks were used as extra information. The number of acoustic events so obtained was then compared with the number of target functions associated with the same phoneme. Again, the results were averaged over broad phonetic classes.

4.3.3 Results

4.3.3.1 Phonological judgement of the target functions

The results of the analysis of the 100 sentences are presented in Table 4.4. For each phoneme class the percentages of phonemes associated with zero, one, two or more target functions is given. In the far right column the total number of phonemes within a class is given. Also given are the results averaged over all phonemes. Of course, it would be naive to expect only one target function for each phoneme; clear exceptions could be plosives and diphthongs. However, since most segmentation and recognition techniques concentrate on the phoneme level, in presenting these results we do not want to anticipate the subdivision of phonemes into subphonemic events. Thus, Table 4.4 gives a good impression of the achievements of temporal decomposition on the phoneme level.

These results immediately raise some questions. Why is a substantial part of the vowels associated with more than one target function? Why is more than 30 % of the diphthongs associated with only one target function? Furthermore, although plosives can be considered to consist of at least two subphonemic events (viz. occlusion and burst), why are they often not associated

Table 4.4: Percentages of phonemes associated with zero, one, two or more target functions for six phoneme classes in 100 sentences. The far right column gives the total number of phonemes for each class.

Phoneme	No. of target functions				
class	0	1	2	> 2	No.
Vowels	4.6	66.7	25.7	2.9	781
Diphthongs		32.6	55.8	11.6	86
Plosives	25.2	63.5	10.2	1.2	433
Nasals	6.2	73.8	19.5	0.5	210
Fricatives	3.1	71.3	24.1	1.6	320
/l,r,j,w,h/	30.8	68.7	0.5		198
Total	11.3	66.2	20.3	2.2	2028

with any target function at all? The same question can be asked of the phonemes of the /l,r,j,w,h/ group. The next section, in which the phonetic approach of this experiment is described, focusses on these questions.

4.3.3.2 Phonetic judgement of the target functions

For each phoneme of a subset of 30 sentences it was judged whether the number of associated target functions was justifiable on phonetic grounds. The judgement criterion used was that each distinct acoustic event should be associated with a separate target function. The results of this perceptual analysis can be seen in Table 4.5. Just as in Table 4.4, the results are averaged over phoneme classes. In this table the plosives are lacking; they will be dealt with separately. The percentages are given of the number of times phonemes are correctly or incorrectly associated with zero, one, two or more target functions. Also the total number of phonemes and the total percentages correct and incorrect are given for each class.

If the percentages correct and incorrect in Table 4.5 are added for each item, a table similar to Table 4.4 results with, however, slightly different percentages. The main cause of this discrepancy is that target functions are sometimes associated with a different phoneme for reasons described above. Furthermore, due to the fact that only a subset of the 100 sentences is used, the distribution of the phonemes differs somewhat.

It can be seen in Table 4.5 that it is considered to be correct that a number of phonemes is not associated with any target function. There are two

Table 4.5: Percentages of phonemes associated with zero, one, two or more target functions for five phoneme classes. The results are divided into "Correct" and "Incorrect", which has been decided perceptually. The far right column gives the total number of phonemes within a class. For each phoneme class the total percentages correct and incorrect are also given.

Phoneme	Correct					Incorrect					
class	0	1	2	> 2	%	0	1	2	> 2	%	No.
Vowels	1.4	68.7	19.0	0.9	90.0	3.3	2.8	3.8		9.9	211
Diphthongs		35.3	41.2	17.6	94.1				5.9	5.9	17
Nasals	4.5	77.3			81.8	6.1		12.1		18.2	66
Fricatives	1.1	77.3	14.8		93.2	3.4		3.4		6.8	88
/l,r,j,w,h/	9.1	74.5	3.6		87.3	12.7				12.7	55

reasons for this seeming parodox. In the first place it appeared that a number of phonologically transcribed phonemes could not be associated with any clear acoustic event in the speech signal. Secondly, in some cases acoustic realizations of phonemes not very distinctively present in the PCM speech disappeared in the LPC speech.

Table 4.5 shows that a large percentage of the decompositions can be considered to be correct on phonetic grounds. Vowels associated with two target functions nearly always consisted of two perceptually different speech segments. For instance, a vowel phonologically transcribed as /*a*/ could consist of the pair [*a*a]. Other examples are /æ/ → [æI] and /e:/ → [eI]. Similar observations apply to the fricatives; in those cases the manner of articulation remained the same, but a clear timbre shift could be perceived, which was also observable in the formant plot. On the other hand, diphthongs associated with only one target function were often realized as a single acoustic event.

The gating technique was also used to examine the plosives more closely. They were dealt with separately since they can be considered to consist of distinct but short-durational subphonemic events, featuring rapid changes which are typically difficult to describe correctly with temporal decomposition. Subphonemic events which might be distinguished are occlusion, burst and aspiration. These subphonemic events, especially aspiration, are certainly not present in every plosive. It is also possible that both burst and aspiration are present but cannot be discriminated. Also, in some cases, a subdivision into subphonemic events cannot be made at all. Moreover, on some occasions, a burst cannot be distinguished from the realization of the following phoneme,

for instance in the case of /t/ followed by /s/.

In Table 4.6 an overview is given of the perceptual analysis of the plosives. For each plosive it is determined what subphonemic events were present in the acoustic speech signal. This resulted in five categories, which are indicated in the first column. The last category consists of those plosives for which it was impossible to make a subdivision into subphonemic events. All other possible categories, for example plosives consisting of a voiced occlusion and a burst, were not found in our database.

Table 4.6: Number of plosives associated with zero, one, two or three target functions. The first column indicates what subphonemic parts of the plosive are present in the acoustic signal. The category "all other plosives" represents all cases where it was too difficult to make this subdivision.

	No. of target functions			
Present in the speech signal	0	1	2	3
Silent occlusion	6	2		
Voiced occlusion	2			
Silent occlusion + burst	3	21	17	
Occlusion + burst + aspiration		2	4	3
All other plosives	11	38		

Instead of dividing the results of the plosives into correct and incorrect, as we did for the other phoneme classes, we prefer to present the "raw" data. With plosives, the decision whether the decomposition is phonetically relevant is difficult to make. A silent occlusion often resulted in a gap in the sequence of overlapping target functions. In this respect, it should be mentioned that of the 21 plosives of the category 'silent occlusion + burst', associated with precisely one target function, 19 of these target functions are located at the place of the burst. A gap is inherent in the low amplitude of the speech signal, due to which the log-area parameters are not very well defined. As a consequence, it is not always possible to find a target function which fulfils the boundary conditions.

In order to be able to give a percentage "correct" for the plosives, a decision has to be made whether or not it is accepted that silent occlusions are sometimes missed. In the former case the percentage correctly decomposed plosives is 79.8. Otherwise, if gaps are not allowed when silent occlusions occur, this percentage decreases to 55.0.

4.3.4 Discussion

In the first part of this experiment the phonetic relevance of the target functions was evaluated from a phonological point of view. Given the phonetic transcription of a sentence, for each phoneme it was determined how many target functions were associated with it. For many applications, such as for instance automatic speech segmentation or speech recognition, the ideal outcome would be one target function for each phoneme. However, from the results of Chapters 2 and 3 it could already be expected that such an ideal outcome would not be attained. Indeed, we see that only about 66 % of the phonemes was associated with precisely one target function. This result is comparable with the results obtained earlier, although the speech material used in previous experiments differs significantly from that used here.

The maximum score which could be obtained is, however, also limited on fundamental grounds. Some of the phonologically transcribed phonemes are not realized in the acoustic speech signal, while others disappear after the LPC analysis. Furthermore, a substantial part of the phonemes is not realized as a single acoustic event. In such cases, one cannot expect an algorithm like temporal decomposition, which does not use any phonetic knowledge, to detect phonemes correctly.

The second part of the experiment was set up to investigate more closely what temporal decomposition does detect. It was expected and hoped that each perceptually distinct speech segment would be associated with a separate target function.

The phonetic observations made during the gating experiment are, of course, well known. Articulatory coarticulation, voice and place assimilation, reduction, etc. all bring their influence to bear on the acoustic realization of speech. What is new here is that, indeed, the relationship between the temporal decomposition (i.e. the target functions) and perceptually distinct acoustic events is quite close. Averaged over all phoneme classes, including the plosives, 87 % of the decompositions correspond to perceptually distinct speech segments.

Intentionally, the results of the plosives were not presented in terms of correct or incorrect, since these decisions are difficult to defend. For example, quite often no target function is found for a silent occlusion, which can easily be understood. Due to the low amplitude of the speech signal, the log-area parameters are not very well defined and vary discontinuously. Consequently a target function cannot always be constructed. Possibly, a suitable smoothing

of the log-area parameters in low-amplitude regions would bring about positive effects.

Although the bursts are quite often detected, in principle they are difficult to model with a target function and target vector. A burst is a very short speech event, whereas the target functions have a built-in minimum length of 5 frames (50 ms). Thus, a burst can be "overlooked" quite easily by the temporal decomposition method. Moreover, even if a target function is found for a burst, the question arises whether the target function really models this burst. Indeed, in the above-described perception experiment with temporally decomposed and resynthesized CVC utterances it was found that the identity of some plosives gets lost, although they were still perceived as plosives. This indicated that the bursts, or more specifically the transitions from burst to vowel, are not always modelled correctly.

Some of the incorrect decompositions can be understood in terms of the algorithms of the temporal decomposition method. For instance, one of the first steps in the iterative procedure is the determination of the location of the analysis window. This location is determined by the position of n_{right} of the previously determined target function (see section 2.3.2). Thus, if the length of this target function is large, the initial analysis window is shifted over a relatively long region. This is correct if the long target function belongs to a single acoustic event. However, if the length is caused by a strong coarticulation with following speech events, the long shift is incorrect. As a result, sometimes no separate target function is found for the skipped acoustic event. It is also possible to find a target function which overlaps another target function almost 100 %. Due to a difference in length, one of the two is a "subfunction" of the other. The algorithm to test the similarity of two target functions (Eq. (2.11)) allows this overlap and due to the difference in length the functions are considered to model different speech events, which is clearly arguable. Furthermore, some of the functions clearly show two peaks. In those cases, phonetic evidence for two target functions is nearly always present. It should be possible to incorporate in the method a procedure to test for the presence of peaks and, if present, to split the target function into two target functions. It remains to be seen whether these shortcomings can be overcome in a new version of the temporal decomposition method.

4.4 Discussion and conclusions

In this chapter, two aspects of temporal decomposition have been explored. In the first place, a perception experiment was carried out to determine how much of the phonetic content of the speech signal is preserved after temporal decomposition. In the second place, it was investigated what kind of phonetic information could be extracted from the acoustic speech signal by means of temporal decomposition.

The speech material used in the two experiments was rather different. In the intelligibility experiment isolated CVC combinations spoken by a Dutch speaker were used, whereas in the other experiment the material consisted of fluently spoken sentences, in this case German. Both choices were based on the consideration that such speech materials lend themselves fairly well for the purpose of the experiment, although the choice for German also followed from the simple fact that such a database was available. The speech material differed also from the material used for the optimization of the method, which consisted of a British English sentence and some Dutch pseudo-CVC words. Despite these differences, similar results seem to emerge from the various experiments. Thus, an attempt to relate the results seems permitted.

From the intelligibility experiment it followed that of all phoneme classes nasals are affected most by temporal decomposition. Also, in the second experiment it appeared that the results for nasals were relatively low compared with those of other classes (Table 4.5). However, the percentage correctly decomposed nasals was still 81.8, whereas the intelligibility percentage was only 54. Apparently, a target function located at the right place is not always enough to model the speech parameters of nasals correctly.

The intelligibility of plosives was also influenced by temporal decomposition, be it to a lesser extent than that of nasals; the plosive characteristics often remained intact. The study of the phonetic relevance of the decomposition revealed that for silent occlusions often no target functions were found. This could be explained by the low amplitude of the speech signal. Because of this low amplitude such a gap probably cannot be of any consequence for the intelligibility of the speech utterance. The bursts, on the other hand, are quite often detected, but it may be doubted whether these target functions can model the burst and the transition to the following phoneme correctly. This is probably the cause of the decrease in intelligibility.

The results presented here give a fair impression of the achievements of temporal decomposition. As yet, this research is of an exploratory nature. A

step for the near future will be to use these results in a specific application. Since the decomposition results are so closely related to the acoustic speech signal, an obvious possibility is to use temporal decomposition as a help for producing a narrow phonetic transcription. However, our aims with temporal decomposition go further than this. In a manner similar to that reported by Van Hemert (1987), the combined information provided by target functions and target vectors could be used to automatically label the most salient speech segments. Temporal decomposition might also be suitable as a preprocessor for automatic speech recognition.

Chapter 5

Evaluation and applications

5.1 Introduction

In this final chapter the results presented in this thesis will be evaluated with respect to the aims stated in the first chapter. The principal objective was to investigate the possibilities of temporal decomposition as a tool to derive phonetic information from the acoustic speech signal. The use of phonetic knowledge, for instance a phonetic transcription, was excluded explicitly in order to obtain a fair impression of the achievements of the method itself. The investigations have been exploratory in nature; no specific application was worked at. Since many interesting research possibilities have been left unexplored, a selection of suggestions for further research will be discussed. Furthermore, an overview of possible applications will be given. Some of the applications have already been implemented by other researchers, whereas other applications only exist as ideas.

5.2 Evaluation

5.2.1 Improvement of the method

In order to be able to investigate the phonetic relevance of the decomposition it was necessary to improve and extend Atal's original method, since this method suffered from a number of shortcomings. There were two kinds of shortcomings which had to be solved. In the first place, there were problems which had to do directly with the method itself, irrespective of the intended application. The major problem of this kind was caused by Atal's reduction algorithm (Eq. (2.10)), due to which a gap could result in the sequence of overlapping target functions. Also, a target function was not guaranteed to fulfil the boundary conditions of a well-shaped target function. In the second

place, there were shortcomings which had to do with our intention to use temporal decomposition for the derivation of phonetic information. For instance, the original method focussed on target functions which are temporally as compact as possible. For speech coding such a condition might not really matter. For the phonetic relevance of the target functions, on the other hand, this condition was not satisfactory; the lengths of target functions should be related to the durations of speech events, and these can vary considerably.

The modified method has been described in Chapter 2. Although it was not possible to quantify the improvements, they could be made plausible. Important improvements were solutions to the above-mentioned problems. Except for silent occlusions, the number of gaps in the decomposition is almost reduced to zero, making the method more robust. The method does not focus any longer on target functions which are temporally as compact as possible; speech events of longer duration have as much chance of being detected as short speech events. Only speech events as short as bursts are still difficult to detect. Furthermore, the new method has become less sensitive to initial parameter choices. An example is the length of the analysis window, which is now adapted to the speech event found within this window. Other parameter choices have been optimized. The improvement criterion has always been the correspondence of target functions to speech events.

The modified method has been used throughout the succeeding chapters, although it was realized that the method could be improved further. That is inherent in a method which is still under development. It was felt, however, that it would be better to carry out further explorations using the same method, in order to be able to relate the outcome of the various experiments. Also, the relative importance of the remaining shortcomings would become more clear.

Indeed, the experiments have demonstrated that the modified method still suffers from a few shortcomings. These remaining shortcomings, however, are of a different order from those of the original method. Although these will have to be solved, the method is already useful and usable as it is. Here, they will be discussed briefly.

The similarity algorithm (Eq. (2.11)) checks whether two subsequently determined target functions are different; if not, the second one is rejected. However, this algorithm allows for an almost complete overlap of one target function over another; in such cases two target functions differ only in length and one of the functions can be considered as a subfunction of the other (see for example Fig. 3.1). Strictly speaking, these two target functions are indeed

different and possibly for both functions phonetic evidence is present. For instance, the smaller one could indicate a single acoustic event, whereas the other one points to a strong articulation with a neighbouring acoustic event. For practical applications, however, this is an undesirable artefact.

In Chapter 4 it was mentioned that some target functions clearly exhibited two peaks. As yet, the existence of peaks has not been checked, and otherwise such target functions fulfil the conditions for a well-shaped function. It appeared that a decomposition into two separate target functions would nearly always be phonetically more correct. Addition of an extra boundary condition which checks the presence of peaks should be taken into consideration.

A final shortcoming which should be mentioned is the decomposition of the beginning and the end of the utterance. Clearly, the window adaptation procedure cannot work satisfactorily in those regions since the extension possibilities are limited. This sometimes leads to small spurious target functions, which have no phonetic interpretation. In the present method this shortcoming has been neglected, since it did not lead to serious problems. It does, however, deserve attention.

Suggestions for further improvement are given in section 5.3.

5.2.2 Parameter choice

In Chapter 3 the influence of the choice of input parameters has been investigated. Since the temporal decomposition method was optimized using log-area parameters as input, this experiment was somewhat biased towards these parameters. There was no way of preventing this, however. Temporal decomposition is complex to such a degree that optimization for each parameter set separately was impracticable, let alone optimization for each set in the same measure. For the same reason, the suitability for temporal decomposition of a parameter set could not be predicted from the log-area results. As a consequence, the only possibility was to test the performance of a number of representative sets. Again, the main performance criterion was the correspondence of target functions to speech events.

Despite the bias towards log-area parameters, filter bank output parameters were found to give better results. The achievements of log-area ratios, on the other hand, were much worse. Still, it was demonstrated that, in principle, these parameters could give the same results as obtained with log-area parameters. The decompositions of all other parameter sets were phonetically less relevant than those of the log-area parameters. Although this could be

due partially to non-optimal parameter settings, more fundamental objections, such as unphysical parameter values, also played a role here.

Unfortunately, a speech synthesizer for filter bank output parameters was not available for our experiments. Consequently, intelligibility experiments were not possible using this set of speech parameters. Since we preferred to use only one set of speech parameters, log-area parameters were used throughout the remaining experiments described in this thesis. This choice may have influenced the results of Chapter 4 in a slightly negative way, although the performance criteria in Chapters 3 and 4 were not exactly identical. An analysis, similar to that in Chapter 4, of the phonetic relevance of the target functions using filter bank output parameters instead of log-area parameters, may be worth while.

5.2.3 Phonetic relevance of the decomposition

The major aim of this thesis was to investigate whether it is possible to derive phonetic information from the acoustic speech signal in an objective way by means of temporal decomposition. Stated otherwise, the question was: does the decomposition have any phonetic relevance? From our experiments, however, especially those in Chapter 3, it followed that one should not talk about the decomposition; by varying the parameter settings or the choice of input parameters, a wide variety of decompositions of the same utterance could be obtained. Thus, the relevant question would rather be split up into two. On the one hand, it was necessary to investigate how the phonetically most relevant decomposition could be obtained. This led to the optimization of the method. On the other hand, the decomposition had to be interpreted in phonetic terms.

From the foregoing, the impression could be obtained that temporal decomposition is just a matter of arbitrariness: different parameter settings leading to different decompositions. That is, however, only part of the truth. In Chapter 3 it appeared that if temporal decomposition leads to target functions that are phonetically interpretable, the reconstruction error is nearly always smaller than alternative decompositions although the latter decompositions often consisted of a larger number of target functions. This gives a strong indication that the optimized temporal decomposition method models relevant speech units. Moreover, this means that the improvements with respect to the phonetic relevance of the target functions are also of value for applications such as speech coding.

In Chapter 1 it was argued that the achievements of most approaches in speech research were limited by the fact that they considered speech as a sequence of temporally distinct speech units. The strength of the temporal decomposition approach is that the speech signal is decomposed into overlapping units of variable lengths. Such a description of speech in terms of overlapping units is in agreement with both the perception and production of speech. The phonetic interpretation of these units has been investigated in Chapter 4. The experiments clearly showed that temporal decomposition does not detect the phonemes prescribed by a phonological transcription. Of course, this could hardly have been expected, since some phonemes are not realized in the acoustic speech signal. Even a very advanced speech analysis technique will never be able to locate phonemes which are not present at all in the acoustic speech signal, without recourse to higher level processing in which language models and lexical rules play an important role. For humans a similar task, making a phonological transcription of speech of an unknown language, is also impossible. On the other hand, making a phonetic transcription of an unknown language is possible for experienced phoneticians. Temporal decomposition should be judged with respect to the latter way of transcribing speech. Such a transcription is closely related to the acoustic contents of the speech utterance. In Chapter 4 it was shown that temporal decomposition segments the speech signal into perceptibly distinct acoustic events, and thus the decomposition can be considered as phonetically relevant.

5.2.4 Validity of the model

Temporal decomposition models speech parameters as linear combinations of a number of target vectors. In most respects this turned out to be an adequate description of the speech parameters, provided that a proper set of input parameters was used. In the introductions of both Chapters 2 and 3 it was mentioned that the model was based on articulatory considerations. The target vectors were supposed to model ideal articulatory targets of which the target functions describe the temporal evolution. The experiments did not, however, yield any evidence that the target vectors can indeed be considered as ideal articulatory positions. Of course, ideal articulatory positions were not available as a reference, and thus the target vectors had to be judged on their own merits. An analysis of log-area target vectors in the F_1-F_2 plane (Chapter 3) demonstrated that the vector points exhibit more spread than would be expected if they really modelled ideal positions. Partially, this spread could be

attributed to a degree of freedom in the temporal decomposition method, due to which the length of the target vectors was not always optimal. However, it was felt that this non-optimal normalization was the cause of the deviation of the target vectors from the actually realized speech frames, rather than the deviation from ideal target positions.

Both Niranjan and Fallside (1987) and Marteau et al. (1988) have contributed to an interpretation of temporal decomposition. The sequence of speech parameters can be interpreted as a point moving through a multi-dimensional space. This point moves towards and away from target points, representing articulatory positions. Assuming for a moment that only two adjacent target functions overlap, this means that the space in which the movements of the point take place is reduced to a two-dimensional subspace. In this way it can be understood why temporal decompostion is suitable for economical speech coding: instead of describing the movement of the point through a multi-dimensional space, it is described throught a sequence of two-dimensional subspaces. For most speech sounds this will be a valid approximation, but for rapid transitions it does not suffice. The fact that both log-area parameters and filter bank output parameters give reasonably satisfactory results indicates that the movement through the multi-dimensional space as described by these parameters is sufficiently slow.

Apart from the articulatory target positions, it has also been assumed that temporal decomposition could model coarticulation, namely by varying the amount of overlap of two adjacent target functions. Indeed, Bimbot et al. (1987) claim that in speech synthesis coarticulation could be varied by manipulating the target functions. Of course, the transition from the acoustic realization of one phoneme to the next is determined by the specific overlap of the target functions. However, in a pilot experiment (not reported in this thesis) it was found that temporal decomposition is not nearly sensitive enough to model differences in coarticulation effects. A similar indication could be obtained from the intelligibility experiments of Chapter 4. Substituting an exponential function for a target function resulted only in minor differences in intelligibility. These results indicate that only little phonetic value should be attached to the exact shape of a target function.

A way to test the validity of the linear model is to consider the course in time of the error signal, defined as the Euclidean distance between the original and resynthesized speech parameters. If a linear model satisfies, this error signal is constantly small and does not exhibit distinct peaks. It was shown in Chapter 3 that the log-area parameters fulfil this criterion to a large

extent, especially when compared with the error signals of other parameters. Another way to test the validity of the model is to judge the quality of resynthesized speech. The perception experiments in Chapter 4 showed that the intelligibility of temporally decomposed and resynthesized speech is almost comparable to LPC speech. Though the overall identification scores are significantly less than the LPC scores, this is only due to a small number of phonemes. Although an all-explaining cause for the decrease in intelligibility of these phonemes has not been found, it turned out that rapid transitions were often affected. It is possible that in those cases linear modelling of the speech parameters does not suffice.

5.3 Suggestions for further research

Since temporal decomposition is a relatively new speech analysis technique, only a few aspects of the method have yet been explored. For the future many interesting research possibilities remain. In the following a few suggestions will be given.

From our experiments, especially those in Chapter 4, it followed that about 87 % of all perceptually distinct acoustic events is detected by temporal decomposition. In this context, detection means that an acoustic event is associated with one target function and one target vector. The percentage of 87 is sufficiently high to proceed with a next step, either labelling or segmentation of the speech signal.

5.3.1 Labelling

Van Erp and Boves (1988) state that instead of segmentation of the speech signal, it seems better to just indicate the most salient parts of the speech signal. In this sense, temporal decomposition could provide a helpful tool in labelling speech, since the locations of most perceptually distinct speech events are detected. However, labelling also entails identification of the speech units. In this process, two approaches are possible. In the first place, explicit phonetic knowledge, for instance a phonetic transcription, could be used to identify the speech units. In that case, the units have to be mapped with a given sequence of phonetic symbols. The second possible approach is more complex. In that case the units have to be identified on the basis of their own information. Probably, the target vectors could supply the most important phonetic information, although the length of the target functions could also

be of importance.

In order to be able to identify a target vector one could create a code book of labelled vectors. Each phonetic element should be represented by at least one, but probably more, code book entries. Identifying a target vector amounts to determining to which code book vector the distance is minimal, subject to a suitable distance measure. Instead of matching the target vector itself to the code book vectors, one of the frames of the phonetic element indicated by the target function could be taken. A possibility is to use that frame which has the smallest Eucledian distance to the target vector (SED frame). The advantage of this approach is that all SED frames are guaranteed to be phonetically interpretable, since they are actually realized. Furthermore, it can be seen in Figure 3.5 that the clusters formed by the middle frames of vowels are slightly more compact than the target vector clusters. If that also holds for the SED frames, using these frames instead of the target vectors will result in fewer confusions.

A convenient cluster technique (see e.g. Anderberg, 1973) should be used to create the code book from either the target vectors or the SED frames. Clearly, using only F_1-F_2 would be a severe limitation to the results maximally reachable, discarding a lot of useful information. The best results can be obtained using the whole space spanned by the parameters. Since the spatial resolution must be as high as possible, special care should be given to the choice of parameter space; different spaces will give different results! Extension of this parameter space with a dimension representing amplitude might be considered.

5.3.2 Segmentation

The speech signal is often segmented prior to or as a part of automatic speech recognition processes (e.g. Hatazaki, Tamura, Kawabata and Shikano, 1988; André-Obrecht and Su, 1988). Although temporal decomposition segments the speech parameters into a sequence of overlapping target functions, in a practical implementation preference usually goes to non-overlapping units. Thus, it is necessary to define some boundary criterion for transforming the overlapping units into a sequence of non-overlapping units. The point of intersection of two adjacent target functions, and the location which divides the overlap area into two equal parts, could be possible criteria. This, however, has not been explored and it is not at all sure that this would yield a useful sequence of non-overlapping units. From the intelligibility experiment

it followed that the speech quality was not sensitive to the shape of a target function. Also, the shape of a target function is sensitive to variation of the parameter settings. This indicates that the phonetic value of the exact shape of a target function is limited. Since the shape of a target function is of direct influence on the boundary placement criteria as proposed above, it follows that the accuracy with which the boundaries can be determined is also limited.

In a procedure described by Van Hemert (1987) segmentation of the speech signal is optimized using explicit phonetic knowledge. In his procedure the advantages of an implicit and an explicit segmentation method are combined. This resulted in a useful method for the automatic preparation of diphone libraries. In a like manner, phonetic knowledge could be used to improve the segmentation results of temporal decomposition.

5.3.3 Amplitude information

If LPC coefficients (e.g. log-area parameters) are used as input for temporal decomposition, the amplitude of the speech signal is not taken into account. In filter bank output parameters, on the other hand, amplitude information is integrated in the parameters. In Chapter 3 it has already been suggested that this might be the cause of the better performance of these parameters. It will be worth while to consider ways to make use of the phonetically relevant information which can be provided by the amplitude when LPC coefficients are used as input.

A rough segmentation by hand of the speech waveform is possible using only amplitude information. This suggests that implementing amplitude information at an early stage of the method might be profitable for the determination of the target functions, since these indicate the locations of phonetic elements. However, we did not manage to find a suitable way to realize this implementation. Weighting the speech parameters with the amplitude did not have the desired effect, although in such a way an integration of amplitude information comparable to that of the filter bank output parameters was created. Neither was the use of the amplitude as an extra (11^{th}) speech parameter effective; the time variation of the amplitude appeared to be too independent of the other parameters, due to which the influence of the amplitude on the determination of the target functions became too strong.

It may perhaps be better to use amplitude information after the determination of both target functions and target vectors. An error criterion as defined

in Chapter 3 could be used to judge the decomposition. A high reconstruction error combined with a high amplitude may indicate an undesired gap in the sequence of target functions. The gap could be filled by a search directed at constructing a target function at that particular location.

A final possibility which has not been verified is to test the course of amplitude between, for instance, n_{left} and n_{right} (see Chapter 2) of the newly determined target function. A certain change in amplitude, which nearly always indicates the transition to another phonetic element, will then be considered as unacceptable. In that case, the target function must be rejected, and a new procedure using a smaller initial analysis window should be started.

5.4 Possible applications

In the following sections some possible applications will be discussed. Some of them have already been implemented by other researchers, be it in a preliminary way. Most of them can be characterized as promising, although the merits of all still have to be proved.

Segmentation of the speech signal by means of temporal decomposition has already been discussed extensively in section 5.3.2, and therefore this application will be omitted here.

5.4.1 Recognition

An obvious application of temporal decomposition lies in the field of automatic speech recognition. In section 5.3.1 it has already been indicated how the target vectors might be applied to identify the phonetic elements with which they are associated. Of course, this identification process alone is not sufficient for recognition. As such, temporal decomposition should be viewed as a preprocessor in the recognition process.

Bimbot et al. (1988) report the results of a preliminary recognition experiment on a small corpus of continuously spelled French surnames. In the training phase target vectors are automatically extracted and manually labelled. In the recognition phase a lattice of the three best candidate phonemes is obtained and searched through, taking into account the lexical constraints of the French alphabet. They claim a recognition score of 70 % on the letter level. Although this identification score may not seem very high, it should be noted that unlike many other recognition approaches the number of words to be recognized is not restricted.

Possibilities to further improve this rather new technique certainly exist. Niranjan and Fallside (1987), for instance, suggested connecting temporal decomposition with Markov modelling. It would certainly be interesting to investigate whether a combination of these two techniques would lead to better recognition results.

5.4.2 Coding

Temporal decomposition was originally proposed for economical speech coding (Atal, 1983). In Chapter 4 we have verified that speech coded by temporal decomposition can indeed be matched to speech with comparably low bit rates. The practical use, however, is restricted to those situations where a time delay is acceptable. This cannot be avoided since the analysis of the speech parameters can only start when a fair number of speech frames is available.

A possible application is the economical coding of a diphone inventory. Houben (1987) investigated this possibility using an approach based on temporal decomposition. He used the first and the last frame of a diphone as target vectors. By weighting these vectors with very stylized target functions, all other frames of the diphone were described. Houben found that for 62 % of the diphones such a simple description sufficed. The remaining diphones needed at least one extra target function and vector to retain the same speech quality. The target functions he used consisted of two straight lines, one of which was horizontal; the results of Chapter 4 suggest that the same bit rate but better results could be obtained if exponential functions were used.

5.4.3 Synthesis

A final application, which has been proposed in literature, is speech synthesis. Synthesis based on temporal decomposition lies somewhere between segmental synthesis (e.g. diphone concatenation) and rule-based synthesis. Ahlbom et al. (1987) and Bimbot et al. (1987) use about 7000 polysons as their basic synthesis units. These diphone-like units are classified according to the structure of their target functions. They state that the temporal patterns of all combinations of, for instance, a vowel and an unvoiced fricative are similar. Thus, all polysons of that type are coded with an identical pattern of target functions. The target vectors still depend on the phonemes. These coded polysons can be used directly for speech synthesis, but by manipulating the target function pattern, it is also possible to introduce variations in speaking

rate, stress, etc. With this synthesis system, they aim to derive better rules for acoustic speech synthesis.

Summary

Temporal decomposition is a speech analysis method, proposed in 1983 by Atal for economical speech coding, which segments the continuously varying acoustic speech signal into a sequence of overlapping units of variable lengths. Each unit is composed of a target function and an associated target vector. The target vector can be considered to model an articulatory target position. The target function describes the temporal evolution of the target. Although the method takes into account some articulatory considerations, no use is made of explicit phonetic knowledge. This study has been set up to investigate the phonetic meaning that can be attributed to these speech units. Such knowledge will not only provide deeper insight into the structure of the speech signal, but may also have applications in speech coding, segmentation, recognition and synthesis.

The original temporal decomposition method suffered from a number of shortcomings. Although these problems were not fundamental to the philosophy of the method, they had to be solved before an extensive investigation of the phonetic relevance of the speech units could be carried out. Chapter 2 focusses on improvement and extension of the method. An important improvement is that the method no longer forces the target functions to be as temporally compact as possible. This property of the original method militated against the search for speech events of longer duration. Furthermore, the new method uses a less arbitrary criterion for the selection of the target functions. Together with other changes this resulted in a more robust temporal decomposition method.

Many representations of the acoustic speech signal can be used as the input for temporal decomposition. However, a pilot experiment demonstrated that the decomposition outcome depends to a great extent on the choice of input parameters. The original method of Atal (1983) as well as de modified method described in Chapter 2 use log-area parameters as input. In Chapter 3 an extensive experiment is described in which the decomposition results of nine commonly used speech parameter sets are compared. The main objective was to investigate whether the results of the optimized method could be further improved by using a different set of input parameters. The main performance criterion was the phonetic relevance of target functions, but also the phonetic relevance of the target vectors and resynthesis of the speech signal were taken into account. With respect to the main criterion, filter bank output parameters gave the best results, followed closely by log-area parameters. A

speech synthesizer for filter bank output parameters was not available, however. As a consequence log-area parameters have been chosen here as the most suitable set of input parameters for temporal decomposition.

Before attempting to interpret the speech units determined by temporal decomposition, it is worth evaluating how closely a description of speech in terms of target functions and target vectors resembles the original speech signal. In Chapter 4 a perception experiment is described in which the intelligibility of temporally decomposed and resynthesized CVC utterances is measured. Although the identification score differed significantly from the LPC score, which can be considered as the maximally reachable score, this turned out to be mainly due to a small number of initial consonants. Especially the intelligibility of nasals, and to a lesser extent that of plosives, was affected. The overall scores, however, justified further research on the phonetic relevance of the decomposition.

The perception experiment included an evaluation of a quantized and a stylized version with very low bit rates. The results showed that temporal decomposition is indeed very suitable for economical speech coding (down to at least 1.8 kbit/s).

Chapter 4 also contains an evaluation of the phonological and the phonetic relevance of the decomposition on a database of 100 phonologically balanced German sentences. The phonological approach starts from a phonetic transcription of the sentences, and for each phoneme the number of target functions associated with it is determined. The phonetic part focusses on the question whether it can be understood on acoustic grounds why a phoneme is associated with a particular number of target functions. It was found that 87 % of the target functions can be related to a perceptibly distinct sound, and thus can be considered as phonetically relevant. Partially, the remaining 13 % can be attributed to plosives of which the occlusion and the burst are not always detected. Another part is due to shortcomings of the method which, however, are of a different order than those of the original method.

In Chapter 5 the results of the previous chapters are evaluated. The improvement of the method, the parameter choice, the phonetic relevance of the decomposition, and the validity of the model, are discussed. Since many interesting research possibilities have been left unexplored, some suggestions for further studies are given. Finally, possible applications of temporal decomposition in the fields of automatic speech recognition, speech coding, speech synthesis and speech segmentation, are discussed.

Samenvatting

Temporele decompositie is een spraakanalysemethode, die door Atal in 1983 werd voorgesteld voor het zuinig koderen van spraak. Deze methode deelt het continu in de tijd variërende akoestische spraaksignaal op in een reeks overlappende eenheden met variabele lengtes. Elke eenheid wordt beschreven door een doelfunktie en een daarbij behorende doelvektor. De doelvektor kan worden beschouwd als een model voor een artikulatorische doelpositie. De doelfunktie beschrijft hoe deze doelpositie in de tijd wordt benaderd. Hoewel de methode rekening houdt met enige artikulatorische overwegingen, wordt er geen gebruik gemaakt van expliciete fonetische kennis.

Dit onderzoek is opgezet om de fonetische betekenis, die mogelijk aan de spraakeenheden kan worden toegeschreven, nader te onderzoeken. Kennis hierover zal niet alleen bijdragen aan een dieper inzicht in de samenstelling van het spraaksignaal, maar kan ook toepassingen hebben op het gebied van spraakkodering, -segmentatie, -herkenning en synthese.

De originele temporele decompositie methode leed aan een aantal tekortkomingen. Hoewel deze problemen niet fundamenteel waren voor de filosofie van de methode, moesten ze wel eerst worden opgelost alvorens een uitgebreid onderzoek naar de fonetische relevantie van de spraakeenheden kon worden uitgevoerd. In Hoofdstuk 2 staat de verbetering en de uitbreiding van de methode centraal. Een belangrijke verbetering is dat de doelfunkties niet langer door de methode worden gedwongen om zo kompakt mogelijk te zijn. Deze eigenschap van de originele methode werkte het zoeken naar langer durende spraakklanken tegen. Verder gebruikt de nieuwe methode een minder arbitrair criterium voor de selektie van de doelfunkties. Samen met enkele andere veranderingen resulteerde dit in een robuustere temporele decompositie methode.

Verschillende representaties van het akoestische spraaksignaal kunnen worden gebruikt als input voor de temporele decompositie methode. Een pilotexperiment liet echter zien dat de temporele decompositie resultaten sterk afhangen van de gekozen input parameters. Zowel de methode van Atal (1983) als de in Hoofdstuk 2 beschreven methode gebruiken log-area-parameters als input. In Hoofdstuk 3 wordt een uitgebreid experiment beschreven waarin de decompositie resultaten van negen vaak gebruikte parameter sets met elkaar worden vergeleken, met als voornaamste doel om na te gaan of er voor de geoptimaliseerde methode parameters zijn die betere resultaten geven dan

de log-area-parameters. Het belangrijkste criterium voor geschiktheid was de fonetische relevantie van de doelfunkties, maar er werd ook rekening gehouden met de fonetische relevantie van de doelvektoren en de resynthese van het spraaksignaal. Met betrekking tot het belangrijkste criterium, de fonetische relevantie van de doelfunkties, gaven bandfilterparameters de beste resultaten, op de voet gevolgd door log-area-parameters. Omdat het echter niet mogelijk was om bandfilterparameters te resynthetiseren, werden log-area-parameters gekozen als meest geschikte parameters om als input voor deze versie van de temporele decompositie methode te dienen.

Voordat geprobeerd werd om de spraakeenheden die met temporele decompositie worden gevonden te interpreteren, was het nuttig om te evalueren hoe goed de beschrijving van het spraaksignaal in termen van doelfunkties en -vektoren op het originele spraaksignaal lijkt. In Hoofdstuk 4 wordt een perceptie-experiment beschreven waarin de verstaanbaarheid van temporeel gedecomponeerde en weer geresynthetiseerde CVC-woorden wordt bepaald. De verstaanbaarheidscore week weliswaar significant af van de LPC-score, die als maximaal haalbare score beschouwd kan worden, maar dit kon worden geweten aan een klein aantal beginconsonanten. Vooral de verstaanbaarheid van nasalen, en in mindere mate ook die van plosieven werd aangetast. Toch rechtvaardigden de verstaanbaarheidscores van de overige fonemen verder onderzoek naar de fonetische relevantie van de decompositie.

In het perceptie-experiment werd tevens de verstaanbaarheid van een gekwantiseerde en een gestileerde spraakversie met hele lage bitrates geëvalueerd. Uit de resultaten volgde dat temporele decompositie inderdaad bijzonder geschikt is voor zeer zuinige spraakkodering (in ieder geval zo laag als 1.8 kbit/s).

Hoofdstuk 4 bevat ook de evaluatie van de fonologische en de fonetische relevantie van de decompositie, waarvoor een bestand van 100 Duitse, fonologisch gebalanceerde, zinnen is gebruikt. In de fonologische benadering wordt uitgegaan van de fonetische transcriptie van de zinnen en er wordt bepaald met hoeveel doelfunkties ieder foneem geassocieerd is. In het fonetische gedeelte gaat de aandacht naar de vraag of op akoestische gronden begrepen kan worden waarom een foneem met een bepaald aantal doelfunkties is geassocieerd. Het bleek dat 87 % van de doelfunkties gerelateerd kan worden aan perceptief verschillende spraakklanken, en dus als fonetisch relevant beschouwd kan worden. De resterende 13 % is voor een deel te wijten aan plosieven, waarvan de occlusie en de plof niet altijd goed worden gedetekteerd. Een ander deel

is toe te schrijven aan enkele aspecten van de methode die nog voor verdere verbetering in aanmerking komen. Deze tekortkomingen zijn echter van een geheel andere orde dan die van de originele methode.

In Hoofdstuk 5 worden de resultaten, die in de eerdere hoofdstukken zijn beschreven, geëvalueerd. Hierbij komen de verbetering van de methode, de parameterkeuze, de fonetische relevantie van de decompositie en de geldigheid van het model, uitgebreid aan de orde. Ook worden er enige voorstellen gedaan voor verder onderzoek. Het hoofdstuk eindigt met een bespreking van een aantal mogelijke toepassingsgebieden van temporele decompositie, zoals automatische spraakherkenning, -kodering, -synthese, en -segmentatie.

References

Ahlbom, G., Bimbot, F. & Chollet, G. (1987) Modeling spectral speech transitions using temporal decomposition techniques, *Proceedings ICASSP*, 13-16.

Anderberg, M.R. (1973) *Cluster analysis for applications*, Academic press, New York / London.

André-Obrecht, R. & Su, H.Y. (1988) Three acoustic labellings for phoneme based continuous speech recognition, *Proceedings 7th FASE Symposium*, 943-950.

Applebaum, T.H., Hanson, A.H. & Wakita, H. (1987) Weighted cepstral distances measures in vector quantization based speech recognizers, *Proceedings ICASSP*, 1155-1158.

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

Bimbot, F., Ahlbom, G. & Chollet, G. (1987) From segmental synthesis to acoustic rules using temporal decomposition, *Proceedings 11th ICPHS, 5*, 31-34.

Bimbot, F., Chollet, G., Deleglise, P. & Montacie, C. (1988), Temporal decomposition and acoustic-phonetic decoding of speech, *Proceedings ICASSP*, 445-448.

Bridle, J.S. & Chamberlain, R.M. (1983) Automatic labelling of speech using synthesis-by-rule and non-linear-time-alignment, *Speech Communication, 2*, 187-189.

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

Cohen, A., Ebeling, C.L., Fokkema, K. & Van Holk, A.G.F. (1961) *Fonologie van het Nederlands en het Fries*, second edition, The Hague, Martinus Nijhoff.

Eggen, J.H. (1987a) Software voor het meten van spraakverstaanbaarheid, *IPO report R 609*, Eindhoven.

Eggen, J.H. (1987b) Evaluation of speech communication quality with a Monosyllabic Adaptive Speech Interference Test, *IPO report MS 596*, submitted to *Speech Communication*.

Eggen, J.H. (1987c) Start evaluation of available methods for analysis, manipulation and resynthesis of speech, *IPO report R 612*, Eindhoven.

Flanagan, J.L. (1972) *Speech analysis synthesis and perception*, Springer-Verlag.

Fowler, C.A. (1984) Segmentation of coarticulated speech in perception, *Perception and Psychophysics, 36*, 359-368.

Gerbrands, J.J. (1981) On the relationships between SVD, KLT and PCA, *Pattern Recognition, 14*, 375-381.

Golub, G.H. & Van Loan, C.F. (1983) *Matrix computations*, North Oxford academic, 16-20.

Gray, A.H. & Markel, J.D. (1976) Distance measures for speech processing, *IEEE Transactions on Acoustics, Speech and Signal Processing, 24*, 380-391.

't Hart, J. & Cohen, A. (1964) Gating techniques as an aid in speech analysis, *Language and Speech, 7*, 22-39.

Hatazaki, K., Tamura, S., Kawabata, T. & Shikano, K. (1988) Phoneme segmentation by an expert system based on spectogram reading knowledge, *Proceedings 7th FASE Symposium*, 927-934.

Hermes, D.J. (1988) Measurement of pitch by subharmonic summation, *J. Acoust. Soc. of Am. 83*, 257-264.

Houben, C.G.J. (1987) Zuinige codering van difonen, *IPO report R 575*, Eindhoven.

Lawley, D.N. & Maxwell, A.E. (1971) , *Factor analysis as a statistical method*, Butterworth, London, 79-82.

Lennig, M. (1983) Automatic alignment of natural speech with a corresponding transcription, *Speech Communication, 2*, 190-192.

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

Marcus, S.M. & Van Lieshout, R.A.J.M. (1984) Temporal decomposition, *IPO Annual Progress Report 19*, 25-31.

Markel, J.D. & Gray, A.H. (1976) *Linear prediction of speech*, Springer-Verlag.

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

Moulton, W.G. (1962) The vowels of Dutch: phonetic and distributional classes, *Lingua*, 294-312.

Niranjan, M. & Fallside, F. (1987) On modelling the dynamics of speech patterns, *Proceedings ECOST* , 71-74.

Nocerino, N., Soong, F.K., Rabiner, L.R. & Klatt D.H. (1985) Comparative study of several distortion measures for speech recognition, *Proceedings ICASSP*, 25-28.

Pols, L.C.W. (1977) *Spectral analysis and identification of Dutch vowels in monosyllabic words*, doctoral thesis, Amsterdam.

Pols, L.C.W. & Olive, J.P. (1983) Intelligibility of consonants in CVC utterances produced by dyadic rule synthesis, *Speech Communication 2*, 3-13.

Sekey, A. & Hanson, B.A. (1984) Improved 1-bark bandwidth auditory filter, *J. Acoust. Soc. Am. 75(6)*, 1902-1904.

Studebaker, G.A. (1985) A "rationalized" arcsine transform, *Journal of Hearing and Speech Research, 28*, 455-462.

Van Bezooijen, R. & Pols, L.C.W. (1987) Evaluation of two synthesis by rule systems for Dutch, *Proceedings ECOST, 1*, 183-186.

Van Dijk-Kappers, A.M.L. (1988) Temporal decomposition of speech: compactness measures compared, *Proceedings of the 7th FASE Symposium*, 1343-1350.

Van Dijk-Kappers, A.M.L. (1989) Comparison of parameter sets for temporal decomposition, to appear in *Speech Communication.*

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

Van Dijk-Kappers, A.M.L. & Marcus, S.M. (1989) Temporal decomposition of speech; to appear in *Speech Communication.*

Van Erp, A. & Boves, L. (1988) Manual segmentation and labelling of speech, *Proceedings 7th FASE Symposium*, 1131-1138.

Van Hemert, J.P. (1987) Automatic segmentation of speech into diphones, *Philips Technical Review, 43*, 233-242.

Viswanthan, R. & Makhoul, J. (1975) Quantization properties of transmission parameters in linear predictive systems, *IEEE Transactions on Acoustics, Speech and Signal Processing, 23*, 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,*

21, 34-40.

Willems, L.F. (1987) Robust formant analysis for speech synthesis applications, *Proceedings ECOST*, 250-253.

Curriculum vitae

20 sept. 1959	Geboren te Haarlem.
aug. 1971 - juni 1977	Revius Lyceum te Doorn, Gymnasium B.
sept. 1977 - okt. 1984	RU Utrecht, Experimentele Natuurkunde, afstudeerrichting Medische Fysica.
sept. 1985 - sept. 1988	Onderzoekmedewerkster in dienst van de Stichting Taalwetenschap van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), gedetacheerd bij het Instituut voor Perceptie Onderzoek (IPO), Eindhoven, voor het verrichten van promotieonderzoek naar temporele decompositie van spraak.
sept. 1988 - feb. 1989	Onderzoekmedewerkster verbonden aan de Horen en Spraak groep van het IPO.
maart 1989 - heden	Universitair docente bij de vakgroep Medische en Fysiologische Fysica van de RU Utrecht.

Stellingen

1. Niranjan en Fallside (1987) bestrijden de bewering van Ahlbom, Bimbot en Chollet (1987) dat singular value decomposition (SVD) geen essentiële stap is voor temporele decompositie. In hun argumenten houden zij echter onvoldoende rekening met het feit dat Ahlbom et al. (1987) het aantal input parameters beperken en bovendien een niet nader omschreven decorrelatie-algoritme gebruiken. Uit de bijgevoegde figuren blijkt wel dat de resultaten behaald met en zonder SVD verschillend zijn.

   G. Ahlbom, F. Bimbot, G. Chollet (1987), Modeling spectral speech transitions using temporal decomposition techniques, *Proc. ICASSP*, 13-16.
   M. Niranjan, F. Fallside (1987), On modeling the dynamics of speech patterns, *Proc. ECOST*, 71-74.

2. Uit het feit dat op een recente conferentie door Carey, Harding, Carey, Anderson en Tucker, een sprekeronafhankelijke spraakherkenner werd gedemonstreerd die veertien zorgvuldig gebalanceerde Engelse woorden kon onderscheiden, mag worden afgeleid dat automatische spraakherkenning nog in de kinderschoenen staat.

   M.J. Carey, R.S. Harding, E.J. Carey, A.J. Anderson, R.C.F. Tucker (1988), A speaker-independent speech recogniser, *Proc. 7th FASE Symposium*, Edinburgh, 9-15.

3. De definitie van "close copy stilering" in het intonatieonderzoek zoals gehanteerd door Willems, Collier en 't Hart (1988) leidt in de praktijk niet tot werkelijk eenduidige stileringen. Dit vermindert de bruikbaarheid van close copy stileringen in het intonatieonderzoek.

   N. Willems, R. Collier, J. 't Hart (1988), A synthesis scheme for British English intonation, *J. Acoust. Soc. Am. 84 (4)*, 1250-1261.

4. Uit het feit dat mensen die regelmatig worden blootgesteld aan difoonspraak deze in het algemeen beter verstaan en hoger waarderen dan naïeve luisteraars, blijkt dat er nog heel wat te verbeteren valt aan de kwaliteit van de spraaksynthese.

5. Om te kunnen beoordelen in hoeverre het "alphabet learning" simulatie-experiment van Carpenter en Grossberg (1987) zinnige resultaten oplevert, moet de gegeven informatie op tenminste drie manieren worden uitgebreid: a) Vermelding van de samenstelling van de klassen van letters, b) Het aannemelijk maken dat dit een zinnige indeling is, en c) Vermelding in hoeverre de indeling afhangt van de volgorde waarin de stimuli worden aangeboden.

   G.A. Carpenter, S. Grossberg (1987), Neural dynamics of category learning and recognition: attention, memory consolidation, and amnesia, in J. Davies, R. Newburgh, and E. Wegman (Eds.), *Brain structure, learning, and memory*, AAAS Symposium Series, 239-283.

6. Boeken die in de titel "de Azteken" hebben staan, zijn vaak volop geïllustreerd met foto's van bouwwerken en kunstvoorwerpen die aan andere Oud-Mexicaanse volkeren worden toegeschreven, waardoor de Azteken ten onrechte te veel eer krijgen.

   Bijv.: C. Burland, W. Forman (1987), *De Azteken. Geloof en goden in het Oude Mexico*, (Nederlandse vertaling).
   De Azteken. Kunstschatten uit het Oude Mexico, Zaberndruck, Mainz am Rhein, 1987.

7. Deze stelling is dubieus.

   Geïnspireerd door: D.R. Hofstadter (1986), *Metamagical themas: Questing for the essence of mind and pattern*, Bantam books, 5-48.

8. De tevredenheid over de kwaliteit van het treinvervoer in Nederland neemt toe naarmate men er frequenter gebruik van heeft gemaakt.

Astrid M.L. Van Dijk-Kappers
Eindhoven, 30 mei 1989