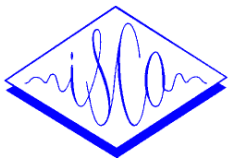

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

A GEOMETRICAL ARGUMENT FOR IMPOSING AN ADDITIONAL CONSTRAINT ON TEMPORAL DECOMPOSITION

P.J.Dix, G.Bloothooft and E.J.M. van Mierlo,

Research Institute for Language and Speech,
University of Utrecht, The Netherlands

Abstract

Temporal decomposition (TD), an analysis procedure introduced by Atal in 1983, yields a linear approximation of speech parameters in terms of a series of time-overlapping interpolation functions and an associated series of data vectors. Essentially, TD is based on a linear model of the effects of coarticulation, the coarticulation effects being modelled by the overlap of neighboring interpolation functions. This method does not make use of any specific phonetic knowledge and can be applied to any time-sequence of data vectors. In this paper we will discuss some of the constraints which must be imposed on such a linear approximation and describe a new and improved method for constructing the interpolation functions.

1.Introduction

Speech can be described as a sequence of distinct articulatory gestures towards and away from articulatory targets, resulting in a sequence of speech events. Since the articulatory system is rather slow, movements towards and away from adjacent targets overlap in time. In 1983 Atal [1] introduced an analysis procedure called temporal decomposition (TD) which models this overlap by decomposing a time-sequence of speech parameters into a series of time-overlapping interpolation functions and an associated series of data vectors. TD is based on the assumption that the articulatory movements from one articulatory target towards the next are sufficiently slow to be *linearly* approximated. In line with this assumption, Atal used log area parameters, which model the cross-sectional areas of an acoustic tube. In principle the method can be applied to any suitable set of speech parameters however, and in fact Van Dijk-Kappers [2] found that somewhat better results were obtained using filterbank output parameters instead of log area parameters. Originally, TD was developed as a low bit-rate coding procedure, but subsequent research has been focused on its acoustic-phonetic interpretation, using it as a model for coarticulation effects.

Following Atal's notation, a sequence of speech parameters $\mathbf{y}(n)$ is linearly approximated by a series of time-overlapping so called phi functions $\phi_k(n)$ and an associated series of so called target vectors $\mathbf{a}_k$:

$$\mathbf{y}(n) \approx \mathbf{y}^*(n) \equiv \sum_{k=1}^{K} \mathbf{a}_k \, \phi_k(n) \quad 1 \le n \le N \qquad (1)$$

In this formula N equals the number of frames in the given utterance, K is equal to the number of speech events and $\mathbf{y}^*(n)$ denotes the approximation of the observed speech parameter vector $\mathbf{y}(n)$.

In TD one starts with the determination of all phi functions. These phi functions are constructed in a left to right traversal of the parameterized speech signal, using two basic assumptions. The first one is that these functions can be derived as linear combinations of the speech parameters $\mathbf{y}(n)$; this assumption basically amounts to the inversion of Eq. (1) in which the speech parameters $\mathbf{y}(n)$ are approximated as a linear combination of the phi functions $\phi_k(n)$. The second assumption is that all phi functions are compact in time, i.e. they are non-zero only in a limited time-interval; using some measure for this compactness, an optimal phi function is constructed out of possible linear combinations of the speech parameters $\mathbf{y}(n)$.

After the determination of all phi functions in a given utterance, all target vectors $\mathbf{a}_k$ are computed in one single step by minimizing the LMS distance between the original speech parameters $\mathbf{y}(n)$ and the linear approximation $\mathbf{y}^*(n)$, i.e. by minimizing:

$$\| \mathbf{y}(n) - \sum_{k=1}^{K} \mathbf{a}_k \, \phi_k(n) \| \qquad (2)$$

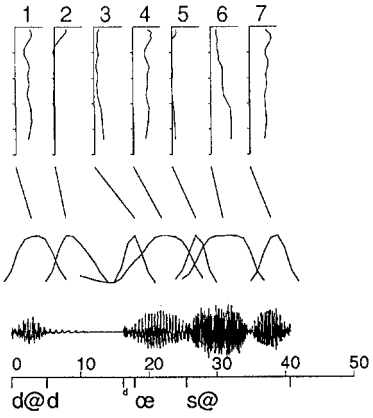

Fig.1 Results of TD on the nonsense word / d@dœs@ /. The overlapping curves depict the phi functions. The target vectors, in this case filterbank output spectra, are drawn at the top.

2.Additional constraints on TD

As already mentioned, TD doesn't make use of any specific phonetic knowledge: it just yields a linear approximation of a sequence of data vectors, resulting in a division of more or less stable regions. Geometrically speaking, one would want such a division to be translation and rotation invariant, since these transformations are, using euclidean metrics, irrelevant. This invariance is set out in the

following scheme, in which R and T denote rotation and translation respectively.

Speech parameters	Approxi-mation	Phi-functions	Target vectors
$\mathbf{y}(n)$ $\mathbf{Ty}(n)$ $\mathbf{Ry}(n)$	$\mathbf{y}^*(n)$ $\mathbf{Ty}^*(n)$ $\mathbf{Ry}^*(n)$	$\phi_k(n)$ $\phi_k(n)$ $\phi_k(n)$	$\mathbf{a}_k$ $\mathbf{Ta}_k$ $\mathbf{Ra}_k$

One can hold different views on the desirability of scale invariance. This invariance is not a result of the use of euclidean metrics, but it seems to be a reasonable demand to a linear method. Scale invariance is set out in the following scheme:

Speech parameters	Approxi-mation	Phi-functions	Target vectors
$\mathbf{y}(n)$ $\lambda\mathbf{y}(n)$	$\mathbf{y}^*(n)$ $\lambda\mathbf{y}^*(n)$	$\phi_k(n)$ $\phi_k(n)$	$\mathbf{a}_k$ $\lambda\mathbf{a}_k$

For most parametrizations of a speech signal you can not attach any straightforward physical interpretation to the above transformations, but in case of filterbank output parameters, scale invariance and translation invariance can be related to independence of overall intensity, and in this case invariance is surely desirable.

The demand of the translation invariance leads to an extra boundary condition for the shapes of the interpolation functions:

$$\mathbf{y}^*(n)+\mathbf{b}=\sum_{k=1}^{K}(\mathbf{a}_k+\mathbf{b})\phi_k(n)=\sum_{k=1}^{K}\mathbf{a}_k\,\phi_k(n)+\mathbf{b}\sum_{k=1}^{K}\phi_k(n)$$

So

$$\sum_{k=1}^{K}\phi_k(n)=1 \qquad (3)$$

Besides implying an extra constraint, Eq. (3) also introduces a limitation of the method: if you make the assumption that at any moment of time only two interpolation functions can overlap, then TD intrinsically yields an approximation of the speech data vectors by means of straight lines. This can be seen as follows: the movement from one event towards the next, say from event k towards event (k+1), is given by a weighted sum of two adjacent target vectors

$$\mathbf{y}^*(n)=\phi_k(n)\mathbf{a}_k+\phi_{k+1}(n)\mathbf{a}_{k+1} \qquad (4)$$

Since the weights $\phi_k(n)$ and $\phi_{k+1}(n)$ sum up to one we have:

$$\mathbf{y}^*(n)=\phi_k(n)(\mathbf{a}_k-\mathbf{a}_{k+1})+\mathbf{a}_{k+1} \qquad (5)$$

which is a parametrization of a straight line through the points $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$. The extra constraint as given by (3) eliminates the following (unwanted) degree of freedom in the linear approximation in terms of phi functions and target vectors: for any set of constants c_k the following equality holds.

$$\sum_{k=1}^{K}\mathbf{a}_k\,\phi_k(n)=\sum_{k=1}^{K}(c_k\,\mathbf{a}_k)(\phi_k(n)/c_k) \qquad (6)$$

In TD each phi function usually is normalized to have a maximum value of 1.0. This practice is defended by arguing that one expects the speech signal to be maximally steady at the instant of time where a phi function reaches its maximum value and one expects the target vector to be close to the speech data vector at this moment. In case of strongly overlapping phi functions this choice is somewhat arbitrary, since in that case there is no part in which the speech signal is steady. We can use the constraint given by Eq. (3) to remove this arbitrariness if we choose the normalization of the phi functions to be such that the constraint (3) is satisfied. If Eq. (3) can not be satisfied exactly then the normalization of the phi functions can be choosen so as to fullfil this equation as good as possible. This can be done by means of a LMS procedure similar to the one that is used in the determination of the target vectors $\mathbf{a}_k$, namely by minimizing the following expression as a function of c_k (cf (2)) :

$$\| I(n)-\sum_{k=1}^{K}c_k\,\phi_k(n)\| \qquad (7)$$

In Eq. (7) I(n) denotes the unit function, c_k denote the normalization constants we want to determine, and $\phi_k(n)$ are the phi functions determined by the TD algorithm, all of them with a maximum equal to 1.

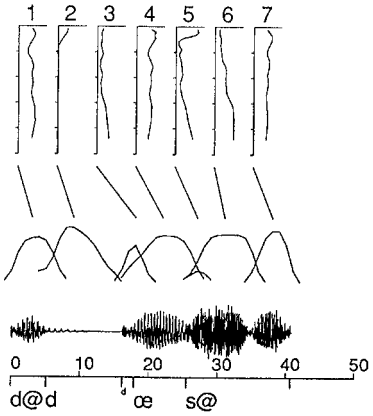

Fig. 2. Results of TD with different normalization factors. The results differ with those of Fig. 1 only in the scaling of the phi functions and the target vectors.

We have investigated the consequences of choosing these normalization factors c_k instead of taking the usual value of 1. In general there were two effects:

- If a phi function has only a limited overlap with its neighboring phi functions then usually the scaling factor c_k is slightly bigger than 1. This result is in agreement with experiments made by Van Dijk-Kappers, who investigated the phonetic relevance of target vectors belonging to several short vowels in different phonemic contexts, and found that her results would improve if the phi functions would have been

normalized to have a maximum that was somewhat bigger than 1

-Phi functions having a strong overlap with both neighboring phi functions tend to be suppressed, i.e. they get a scaling factor considerably smaller than 1. In our opinion such phi functions do not describe any real phonetic phenomena and should be discarded. Remember that you end up with the normalization constant either in the phi function or in the corresponding target vector, so you either have an aberrant phi function or an aberrant target vector.

3.Interpretation of TD as a break-point analysis procedure

The normalization procedure described in the previous section is an ad hoc one, since TD is rotation and scale invariant, but it is not translation invariant, basically because it uses singular value decomposition, which is not invariant for this transformation. Furthermore, Eq. (3) was replaced with only an approximate equality. From a mathematical point of view this procedure is not very satisfactory, however. In this section we will describe TD as a break-point analysis procedure in multidimensional space, break-points being connected by straight lines. We will assume that at any moment of time only two phi functions, which are adjacent in time, are non-zero. This assumption basically amounts to presuming the coarticulation effects to be limited to adjacent events. As we explained earlier, under this assumption one wants TD to yield an approximation by means of straight lines, and in this case the transition from event event k towards event k+1 is given by Eq. (5), which is repeated underneath.

$$\mathbf{y}^*(n)=\phi_k(n)(\mathbf{a}_k-\mathbf{a}_{k+1})+\mathbf{a}_{k+1} \quad (5)$$

Lets assume that the transition takes place in the interval [n,m]. Our algorithm first determines the position of the break-points n and m and it takes $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$ equal to $\mathbf{y}(n)$ and $\mathbf{y}(m)$. These break-points are equal to the middle frames of the k-th and (k+1)-th segment yielded by a segmentation algorithm which is scale, rotation and translation invariant [4].

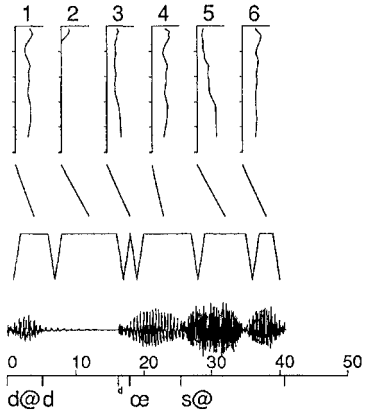

Fig. 3. The break-points are picked as the middle frames of a segmentation algorithm which is scale, rotation and translation invariant. The data vectors belonging to these middle frames are drawn at the top. The segments are indicated by the angular shapes drawn above the oscillogram.

Once that these two break-points n and m are determined, the phi function $\phi_k(n)$ is determined in the interval [n,m] by computing the point on the line segment between $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$ with minimum distance from $\mathbf{y}(n)$. We take $\phi_{k+1}(n)$ equal to $1-\phi_k(n)$. One can easily verify that this procedure also is translation, rotation and scale invariant.

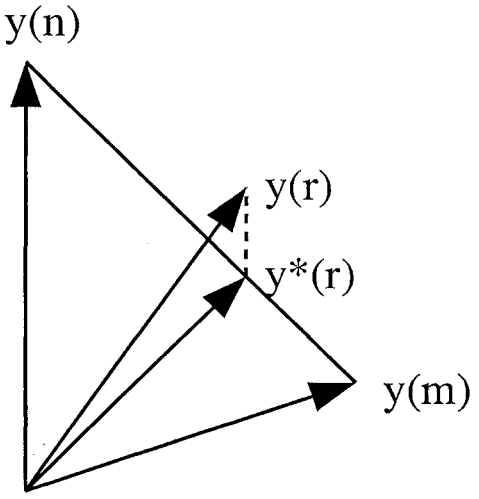

Fig. 4. Construction of the phi functions in the transition interval [n,m]. The point on the line segment with minimum distance from $\mathbf{y}(r)$ is taken as the best approximation.

Clearly the following holds:

$\phi_k(n)=1$,

$\phi_k(m)=0$ and [9]

$0 \le \phi_k(r) \le 1.0$ for r in [n,m]

An appealing result of these properties is that one can interpret the values $\phi_k(n)$ as a kind of activation values of the corresponding event. During the transition from one event towards the next the activation value of the left event decreases from one to zero, while the right event increases its activation value from zero to the value of one.

We have implemented this algorithm and made some comparisons with the results of Atal's analysis. In general we have the impression that our modified algorithm yields phi functions which have a more plausible behaviour, showing little overlap if there is a rapid change and showing much overlap if there is a gradual change from one event towards the next.

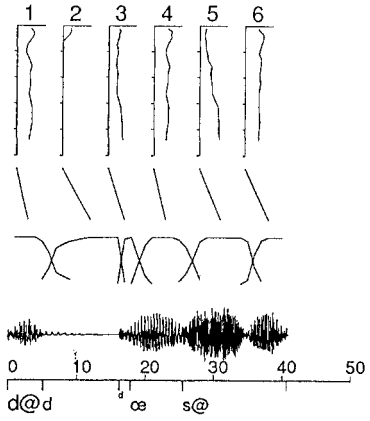

Fig. 5. Results of the modified TD algoritm

Discussion

We have discussed some constraints that the use of Euclidian metrics impose on the temporal decomposition method. We have shown that if at any moment of time only two phi functions overlap TD can be viewed as a break-point analysis procedure in multidimensional space, breakpoints being connected by straight line segments. We showed that TD had to be altered in order to fullfil these constraints and gave a description of a modified and improved algorithm. Finally, we demonstrated that one can interpret the value of a phi functions as a kind of activation values of the corresponding event.

References

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

[2] A.M.L van Dijk-Kappers, (1989), Temporal decomposition of speech and its relation to phonetic information, thesis Technical University Eindhoven

[3] M. Niranjan and F. Fallside (1989), Temporal decomposition: a framework for enhanced speech recognition, Proceedings ICASSP, pp 655-658

[4] P.J. Dix and G. Bloothooft (1990), Segmentation of speech based upon a linear model of the effects of coarticulation, to appear in Proceedings NATO ASI workshop on automatic speech recognition.