TIME-DOMAIN. DIGITAL SEGMENTATION OF CONNECTED NATURAL SPEECH

Wolfgang J. Hess

Institut fur Informationstechnik, Technische Universitat MQnchen
Munich, Fed.Rep. of Germany

Abstract

The digital segmentation algorithm described in this paper subdivides speech signals into discrete sections which permit to localize most of the spoken phonemes in natural speech. Two pre-segmentation steps separate pauses and voiceless parts from the (voiced) rest of the signal. The subsequent main segmentati on step tries to describe the speed of articulation in the vocal tract according to some global speech parameters. Since, during an utterance, the vocal tract does not move at constant speed, but attempts to realize the articulatory target position associated with each phoneme, sections with relatively low changes of vocal tract position ("stationary" segments) and sections with greater changes ("dynamic" segments) can be separated. The dynamic segments can be further characterized when the direction of change in the course of the parameters is regarded.

1. Introduction

This paper describes a segmentation algorithm which forms part of a recognition system for natural speech on the basis of phonemes and phoneme-like elements (Fig. I). The topic of this paper will be confined to the segmentation steps; the remaining steps of the system have been described elsewhere /Hess 1974.1 ,1974.2 ,1972/.

The extreme difference between the information content of the acoustic speech signal and its written counterpart forms one principal problem of any automatic speech recognition system. The classifier itself which classifies the signal into the desired output classes (such as words, phonemes etc.) usually cannot cope with an information content too large. Therefore,it needs a preprocessor which reduces the great redundancy of the signal. For this, one has to extract a series of significant parameters which, whatever they are, maintain most of the (phonetic) information significant for the classifier, but throw off a great part of the signal redundancy. Another problem in recognition of continuous speech is given by the fact that the output of the recognizer has to be discrete in time, and that the classifier can process a limited number of output classes only. Thus, it is obvious that the output of the classifier in a recognition system for continuous speech cannot be words and even not be syllables /Olson, 1967/. Recognition on the basis of phonemes or similar elementary units, however, requires time localization of these units in the speech signal. That means, it poses the problem of segmentation. It is left to the processing strategy whether segmentation is done together with classification (recognition) or during the preprocessing step /Paulus 1974/. In the system described here, segmentation is performed as early as possible.

The first approach to segmentation is already found in vocoder speech transmission systems. This step which, in the following, is labeled "pre-segmentation", determines two binary features of the speech signal, both of which are concerned with the voice source:

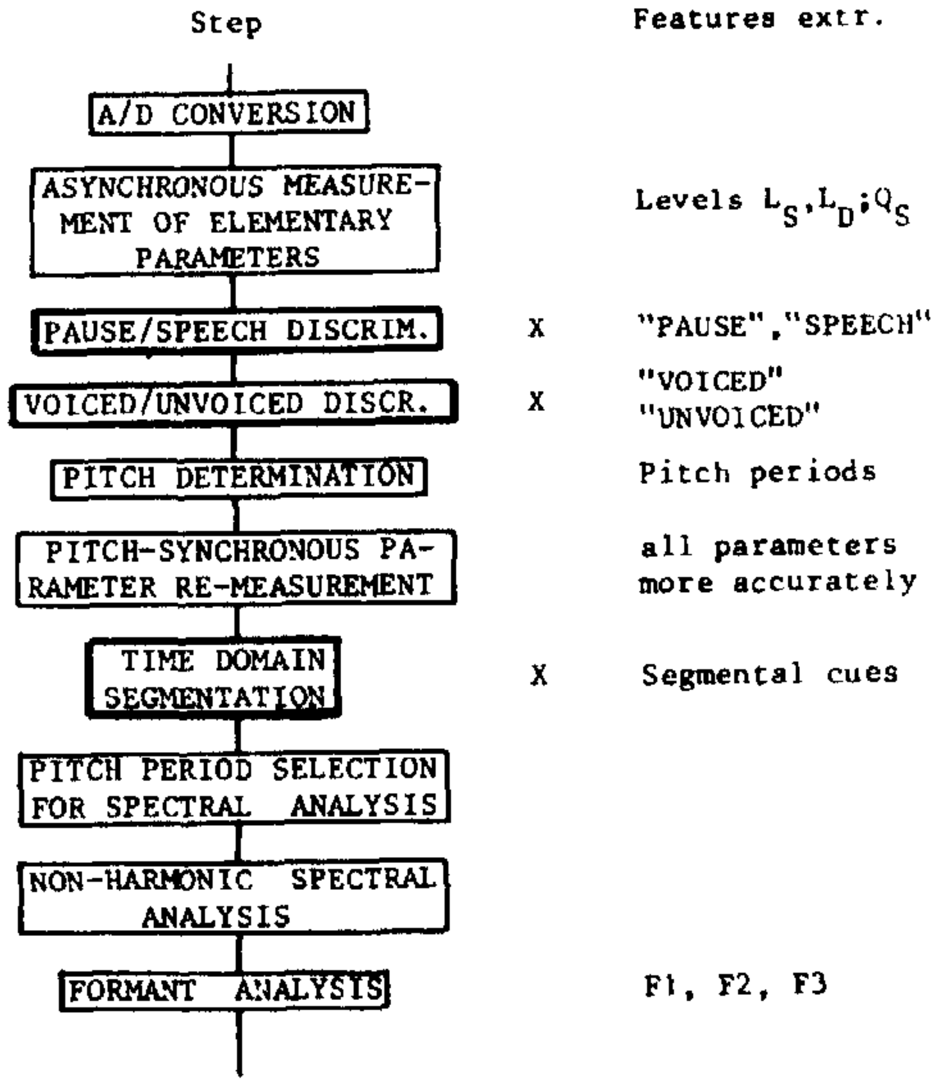


Fig. 1: Block Diagram of the Pitch-Synchronous,Digital Feature Extraction System (PDFES) for Speech Signals (Taken from /Hess, 1974.2/)

X: This part of the system is described in detail in this paper.

1) Discrimination "Pause/Speech".

2) Discrimination "Voiced/Voiceless".

Regarding the vocal tract instead of the voice source will lead to a different kind of segmentation. The vocal tract tries to realize the appropriate target position for every spoken phoneme /Fant and Lindblom, 1961; Flanagan, 1972/. For the further investigations, one has to issue from the following assumptions /Bhimani, 1963/:

a) When realizing a spoken phoneme, the vocal tract first will adjust itself to a target position, then will remain there for a certain time, and finally, will move to the target position associated with the next phoneme. For this reason, fast movements of the vocal tract (leading to "dynamic" or "transitional" segments in the signal) and periods of little change in the vocal tract position ("stationary" segments) will alternate.

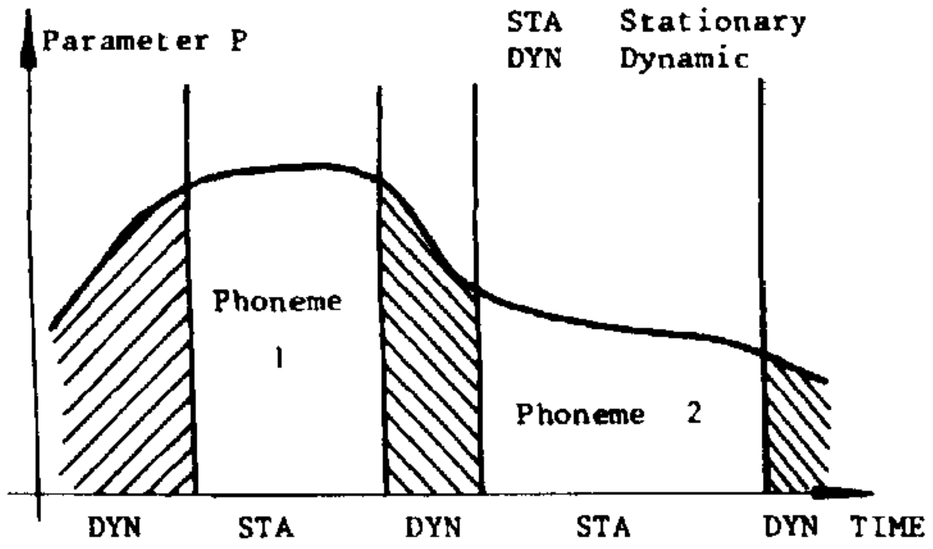


Fig. 2: Schematic Diagram of the Parameter Course and the Segmentation for the Acoustic Realization of two Adjacent Phonemes.

b) Every significant articulation change will cause an essential change in shape of the speech signal as wel1 as its main parameters and vice versa.

A spoken phoneme thus consists of a dynamic beginning, a stationary middle part and a dynamic end which is usually joined by the beginning of the next phoneme to form one dynamic transition (Fig. 2).

There are exceptions: diphthongs may contain several stationary segments, whereas stops may not contain any stationary part at all. Thus, if one succeeds to locate these stationary and dynamic segments, he will be able to locate the spoken phonemes in the speech signal. The third and main step of segmentation therefore will be described as follows:

3) Separate stationary and dynamic segments in the voiced sections of the signal (and, apart from that, also in the longer voice1ess parts whir.h may contain more than one phoneme).

In the following, these steps, to which a correction step is added, are discussed in some detail.

2. Pre-segmentation

This step deals with properties of the voice source. It subdivides the speech signal into voiced sections, voiceless sections, and pauses. The methods applied for this step depend on various environmental conditions, such as the signal-to-noise ratio or the purpose for which the algorithm is to be used. For the procedure described here, the results are used in a speech recognition system, so that a fairly accurate knowledge about pauses and speech sections is required. Furthermore, noise level can be assumed to vary slowly with time. For this reason, a fixed level threshold, as applied e.g. by REDDY /Reddy, 1966/, did not prove sufficient for an accurate pause-speech discrimination. Instead, the level distribution of the signal is taken in form of a histogram during preprocessing (Fig. 3). At the noise level, this histogram shows a distinct peak. Since the signal level during speech sections is subject to much greater changes than the noise level, a threshold L_„ can readily be derived from this peak; thus, pause and speech sections are separated. To perform a better discrimination of the weak fricatives, this procedure is also applied to the level L_D of the digitally differenced speech signal. By this, a second threshold L_{PD} is derived. Hence, a signal section is classified as "pause" when the levels measured are situated below both of these thresholds:

$$\text{"PAUSE"} = (L_S < L_{PS}) \;\&\; (L_D < L_{PD}) \qquad (1)$$

Since the level parameters L_S and L_D are computed as the absolute average of the signal a_n resp. the differenced signal:

$$\begin{aligned} L_S &:= \frac{1}{N} \sum_{i=n+1}^{n+N} |a_i| \\ L_D &:= \frac{1}{N} \sum_{i=n+1}^{n+N} |a_{i+1} - a_i| \end{aligned} \qquad (2)$$

N covers a period of 25 msec resp. - after pitch determination - one pitch period.The level distribution is continuously updated; the influence of more recent values is emphasized by multiplying the whole histogram by a constant less than one from time to time.

Separation of voiced and voiceless sections is done with regard to the ratio Q_S of the levels L_D and L_S:

$$Q_S = \frac{L_D}{2 \cdot L_S} \qquad (3)$$

This parameter gives a crude estimate of the behaviour of the signal in frequency domain; by fixed thresholds the signal is divided into sections labeled "voiced", "voiceless", and "may-be-voiceless". The latter sections will be further classified in the subsequent pitch determination step /Hess, 1974.1/.

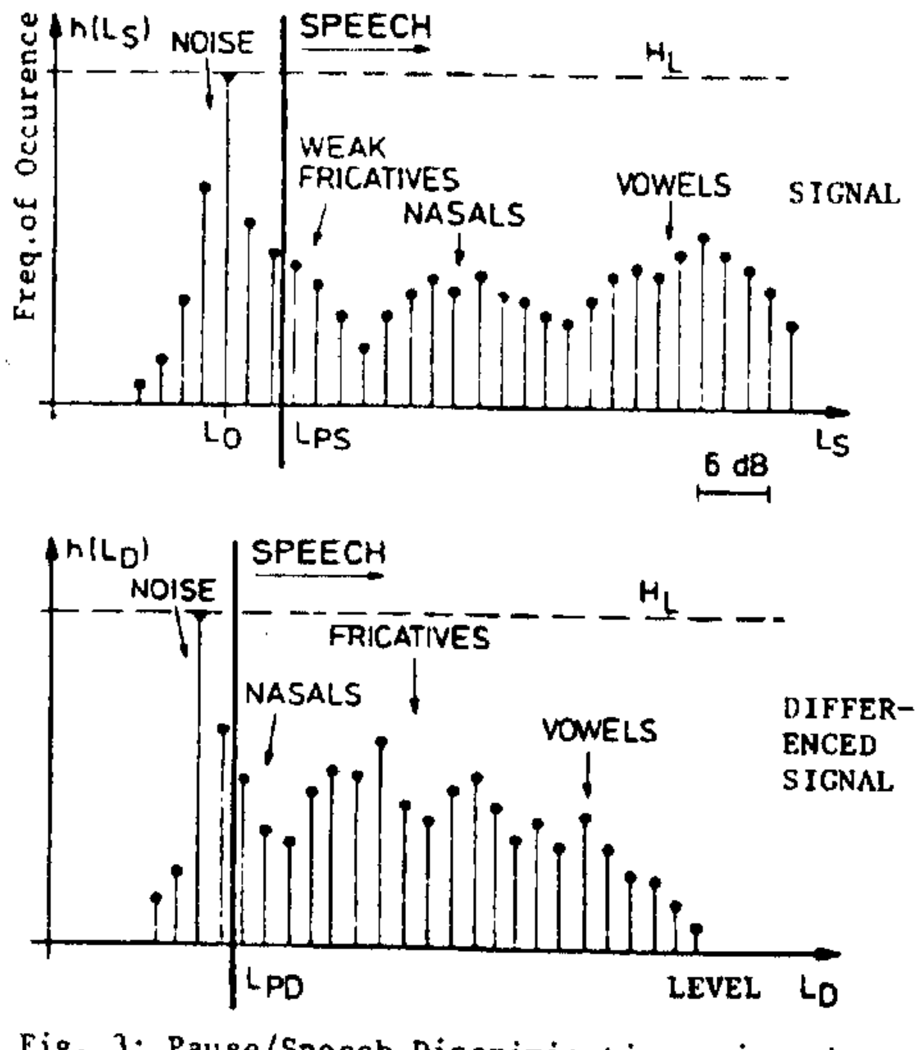


Fig. 3: Pause/Speech Discrimination using the Level Distribution.

3. Choice of Parameters. Segment Length Function, Minimal Segments.

Are the stationary segments sufficiently stationary resp. the dynamic ones sufficiently dynamic to be determined by an automatic classification algorithm with reasonable reliability? What parameters have to be selected to perform this? In this step, the feature "stationary" or "dynamic" is not a physically determinable characteristic of the speech signal as e.g. the existence of pitch. For this reason, the parameters selected for segmentation have be treated in a nonlinear way to discriminate the signal into the desired segment classes; this can be performed by an adaptive or by a rigid algorithm. As former experiments show, especially by REDDY and VICENS /Reddy and Vicens, 1968/, the segmentation of the signal into stationary and dynamic segments seems possible using few global speech signal parameters and a rigid, non-adaptive decision-tree algorithm.

In the following, the three parameters defined in eqs. 2 and 3 together with the binary features extracted in the pre-segmentation step are used. These parameters are measured pitch-synchronously /Hess, 1972/ and then interpolated at constant intervals ("microsegment interval"). If pitch is not available, these parameters are measured for a period of 25 ms before interpolation. The microsegment interval has been set to 2.5 to 5 ms for the speech material used in these investigations. This short interval was selected in order to process even fast transitions in a correct way; this seems justified by the accuracy of the pitch-synchronous parameter measurement.

Concerning the strategy of segmentation, it proves advisable to give priority to the determination of the dynamic segments. As experiments e.g. by 5HMAN /Ohman, 1962/ show, dynamic segments are more important for speech perception than stationary ones. Secondly, in a recognition system using the results of segmentation, undetected phoneme boundaries will cause irreparable recognition errors. For this reason, the following procedures deal with the dynamic segments primarily.

For further processing, assume the signal to be subdivided into microsegments of fixed length. For each parameter,the relative change between two - not necessarily adjacent - microsegments Mj and M^ is determined as follows:

$$r_{ik} = r_{P_{i,k}} := \frac{P_k - P_i}{P_k + P_i + K} \qquad (4)$$

$$P \in (L_S, L_D, Q_S)$$

In this equation,P represents the segmentation parameter regarded. Mote that nil the parameters Lg, L_D.and Q_s can take positive values only.K is a correction factor whose influence will be discussed later.The relative change r-, is defined to be a "major" change when its absolute value exceeds a given threshold q :

$$\text{"MAJOR CHANGE"} := |r_{P_{i,k}}| > q_P \qquad (5)$$

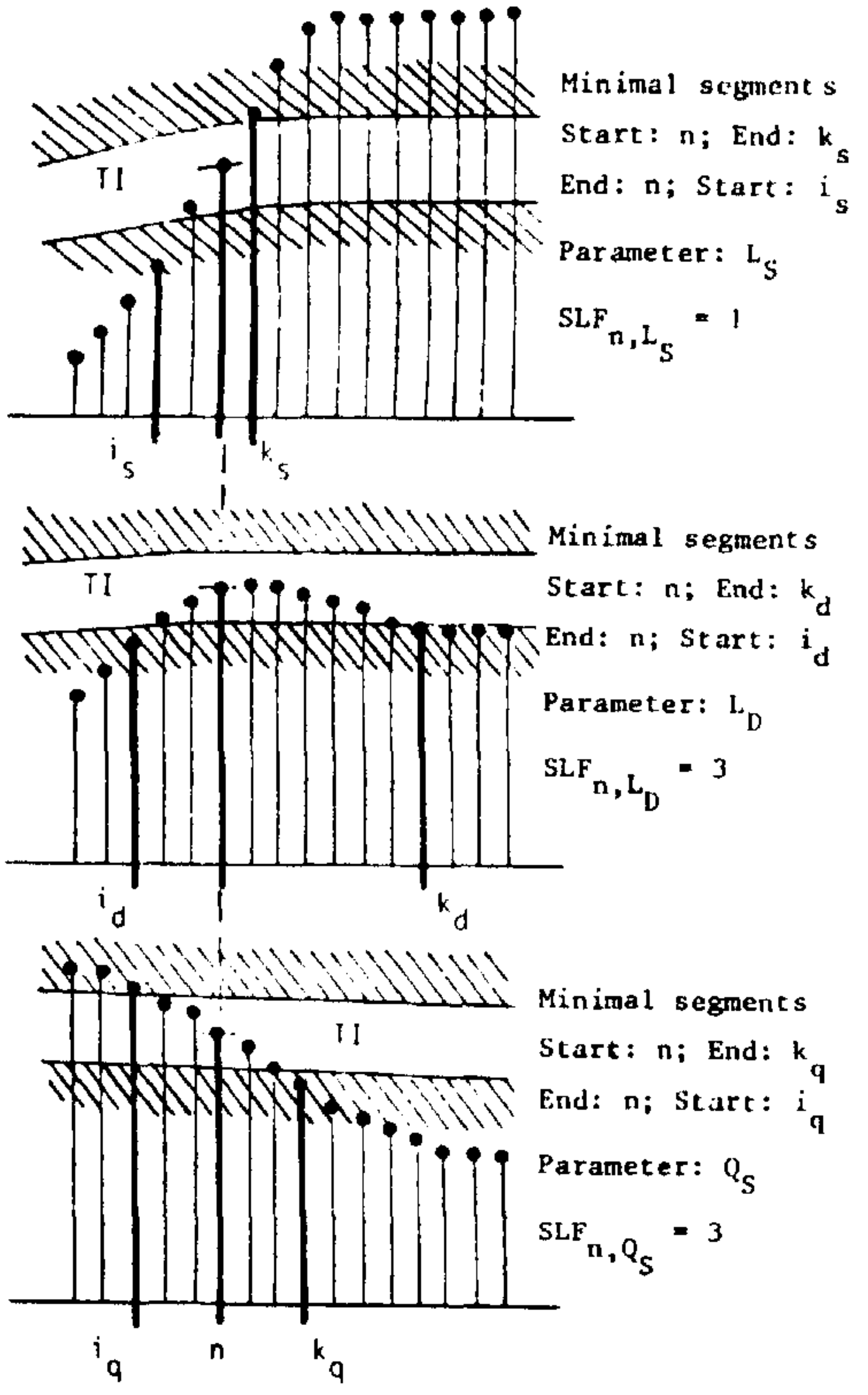


Fig. 4: Definition of Segment Length Function SLF and Minimal Segment.

L_S, L_D, Q_S	Parameters
i, k	Initial and final points of minimal segments
n	Location for which SLF is computed.
Tl	Tolerance interval for $|r_{nj}| < q_P$
q_P	Sensitivity threshold
r_{nj}	Parameter change, as defined in eq. 4
	In the figure: SLF_n = 1

This equation defines the "minimal segment" as sequence of adjacent microsegments: its length is determined in a way that between its initial and final point the controlling parameter P is subject to one major change (Fig. A).

Direction of processing plays an important role in the computation of the minimal segments, when this computation is performed time - sequentially

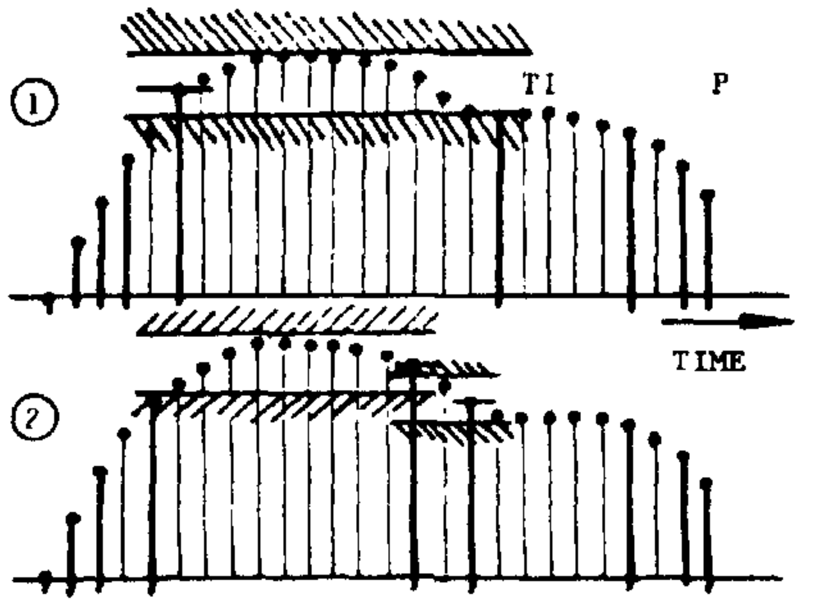


Fig. 5: The Influence of the Direction of Processing when the Minimal Segments are Computed Time-Sequentially out of the Parameter Values Without Using the SLF.

P Parameter
TI Tolerance Interval

1 Forward Processing
2 Backward Processing

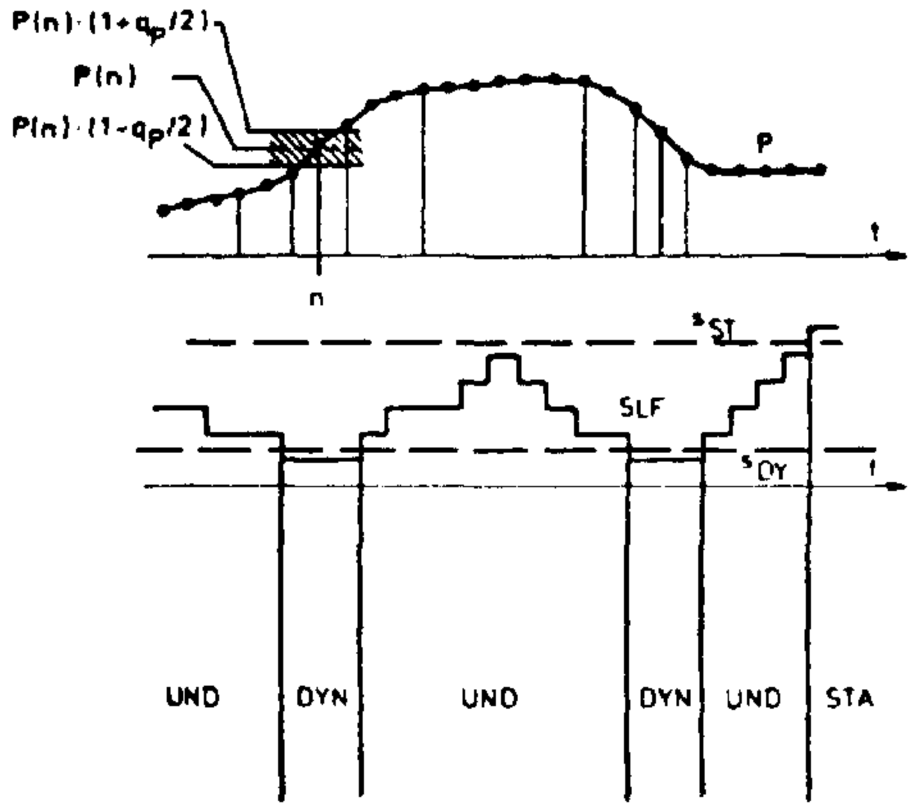


Fig. 6: Primary Segmentation.

STA	Stationary	P	Parameter
DYN	Dynamic	q_p	Sensitivity threshold
UND	Undefined		

s_{ST} SLF Threshold "Stationary"

s_{DY} SLF Threshold "Dynamic"

(see example in Fig. 5). The influence of the processing direction causes significant errors and, for this reason, has to be eliminated. That means that the minimal segments cannot be computed time-sequentially. Hence, an intermediate function, the "segment-length-function" SLF is defined in the following way:

$$SLF_n := \min(1_{i,n}, 1_{n,k}) \qquad (6)$$

$$\text{for } |r_{in}| > q_p \text{ and/or } |r_{nk}| > q_p$$

This definition is valid for any of the parameters P. Tt *is* extended to the binary features in a way that any change of these features is regarded a major one. 1. is the length of the minimal segment which ends at point n, whereas $1_{n,k}$ stands for the length of the minimal segment beginning at n. Thus, the SLF describes beginning and end of every significant change in the controlling parameter and, thus, locates all dynamic segments. In eq. 6, the SLF is defined for one parameter at a time only. If more parameters have to be regarded, a combined SLF is defined as:

$$SLF_n = \min(SLF_{n,P_i}) \qquad i = 1\ (1)\ k \qquad (7)$$

In this equation, SLF_{n,P_i} is the SLF as defined for a single parameter P. according to eq. 6. The combined SLF is computed as the minimum of the SLF values for the individual parameters P.. This definition again emphasizes the dynamic segments, since the value of the SLF is determined by a single major change of one of the parameters already.

After computing the combined SLF of the whole signal, the minimal segments are determined in a way that first all microsegments with SLF-1 (microsegments) are combined to minimal segments,after that the microsegments with SFL=2 etc. (Fig. 4 and 6).

Before that, the SLF has been smoothed in a way that the difference between adjacent SLF values cannot be greater than one. In order to maintain the exact localization of the dynamic segments, the SLF must not increase its value at any point during the smoothing procedure. This step from the microsegment to the SLF reap. to the minimal segment provides minimum dependence of the further segmentation steps on the parameters applied, and - by means of the thresholds q - also minimum dependence on the individual speaker. *Since it* can influence the magnitude of the SLF but not its structure, especially the situation of its extremes, in an essential way, the individual adjustment of the thresholds q is not too critical (Fig. 6).

4. Primary Segmentation.

The following steps labels all the minimal segments "dynamic", "stationary", or "undefined". Adjacent minimal segments with equal labeling are grouped together to "primary" segments. A minimal segment certainly can be labeled dynamic when, during its course, there is a major change between adjacent microsegments. A minimal segment certainly is stationary when there is no major change in any parameter during a wider neighbourhood (30 ms or more). The rest of the minimal segments not belonging to one of these categories is labeled "undefined" (Fig. 6).

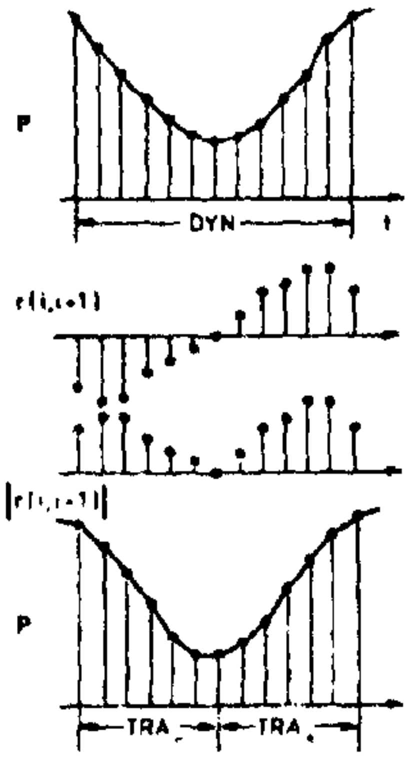
5. From the Dynamic to the Transitional Segment.

A dynamic section in the course of the speech signal always represents the transition from one position of the vocal tract to the subsequent one. This transition can be characterized more accurately, if one succeeds *to* find a measure or a statement for the direction of the transition, and, especially, if one succeeds to indicate whether the transition proceeds in a monotonic way or not. The really transitional segment, that means the direct transition from a spoken phoneme to the next one, should reveal a monotonically increasing or decreasing course of the parameters. In some phonemes, such as stops, however, the target position of the vocal tract is not sustained; thus, a stationary segment cannot be expected at that point. In this case, a sequence of adjacent transitions is grouped together to one dynamic segment by the primary segmentation algorithm. To find a measure for the degree of monotony in a transition, the auxiliary values

$$d_m := \frac{1}{L} \Sigma r_{i,i+1} \qquad c_m = d_m/b_m$$

$$b_m := \frac{1}{L} \Sigma |r_{i,i+1}| \qquad (8)$$

where L represents the number of microsegments in the m-th primary segment, are determined for all segments and all parameters. b_m is a criterion for the over-all change of each parameter within a segment, whereas c^ is a criterion for the direction of change as well as for its degree of monotony. If the course of the controlling parameter P within a segment is monotonically increasing or decreasing, the absolute value of c_m will be situated in the vicinity of 1. Otherwise, the value c_m will be situated around 0. After computing these values, the dynamic segments are transformed into "transitional" segments by determining their direction characteristics. Monotonic segments are labeled "transitional-rising" resp. "transitional-failing". Each non-monotonic dynamic segment is subdivided into monotonic fractions if it exceeds a certain length (see Fig. 7). In order to create a unique determination of direction characteristics when several segmentation parameters arc involved, in each dynamic segment, one of the parameters is selected and labeled dominant for the direction characteristic of that particular segment resp. the transitional segment(s) which will result from it /Hess, 1972/.



DYN	Dynamic
TRA	Transitional
+,-	Direction characteristics
r	Parameter changes between adjacent microsegments
P	Parameter (L_S, L_D, Q_S)

Fig. 7: From the dynamic to the transitional segment.

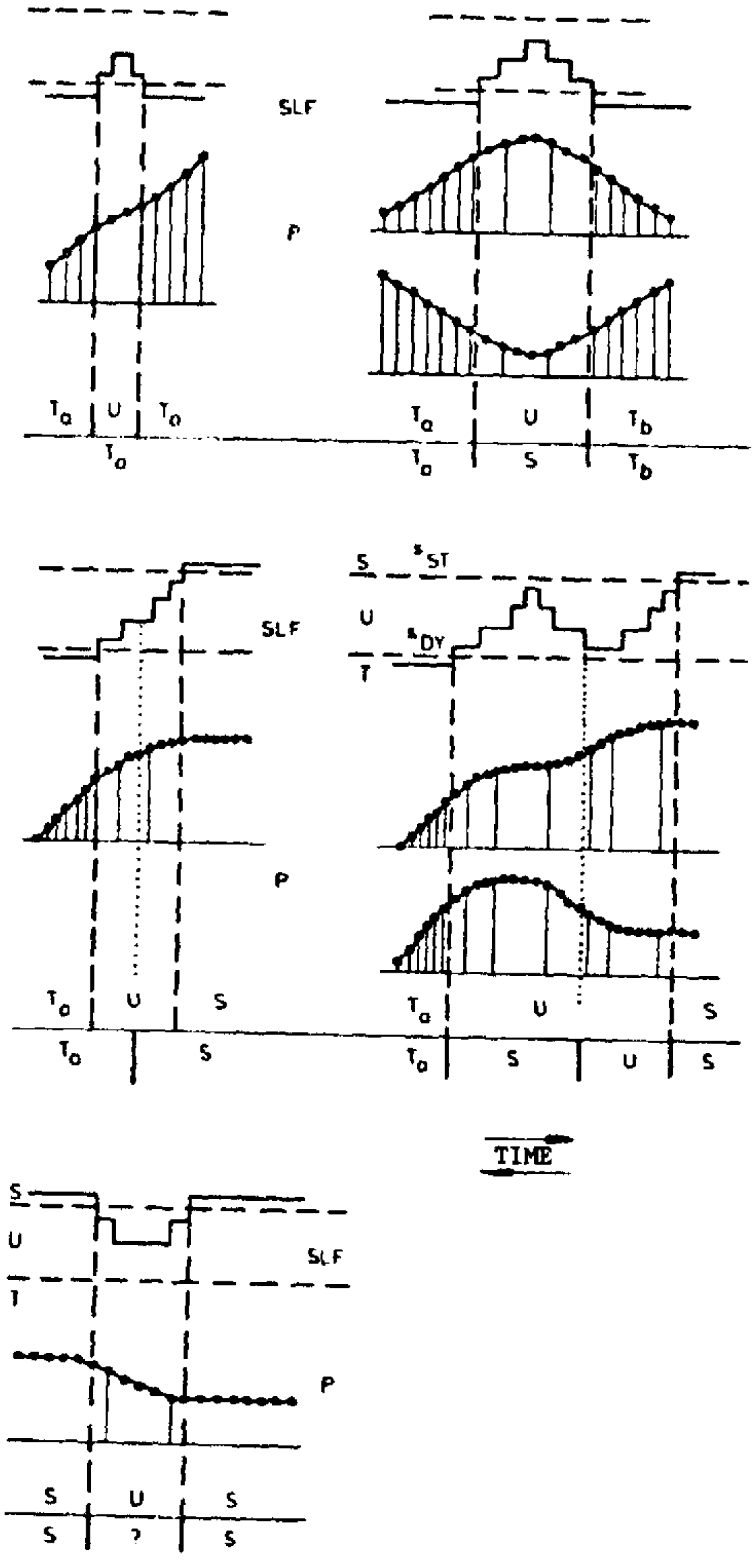


Fig. 8: Processing and Elimination of Short Undefined Segments.

T	Transitional
S	Stationary
U	Undefined
?	Decision ambiguous (depends on individual circumstances)
P	Parameter (L_S, L_D, Q_S)
——	Segmentation $\frac{\text{BEFORE}}{\text{AFTER}}$ Procedure
s_{ST}	SLF Threshold "Stationary"
s_{DY}	SLF Threshold "Dynamic"

6. Undefined Segments.

After the primary segmentation, about 30X of the speech signal remain as undefined segments. These segments have heen generated at those points of the speech signal where a clear decision as to stationary or dynamic could not he made immediately.A special algorithm eliminates these undefined segments. For this, it needs the classification of their neighbours. Four main categories of comhinations with undefined segments may occur:

a) **"transitional" - "undefined" - "transitional"**

h) **"stationary" - "undefined" - "stationary"**

c) **"transitional" - "undefined" - "stationary" and vice versa**

d) long undefined segments in any environment.

The performance of the algorithm for the "short" undefined segments (categories a to c) with regard to the SLF and the direction characteristics of the undefined segment itself and its neighbours is shown in Fig,8. Two configurations may give ambiguous results: undefined segments between stationary segments as well as undefined segments between transitional ones.the latter only if the direction characteristics arc equal in all three segments involved.In this case the undefined segment is labeled different from its neighbours if the change rate b in the undefined segment and in the neighbours differ strongly. If there is only a slight difference, the segments are grouped together. If the decision is not clear, the undefined segment is labeled "may-be-transitional" or "may-be-stationary" thus revealing that the result at this point is subject to some unsafety.

As can be seen in Fig. 7, it is not provided by the algorithm to split up an undefined segment into more than two segments. This may prove sufficient for isolated words; in connected speech, however, there exist longer undefined segments which may consist of three and more phonemes. These segments arc treated in a recursive way. They are assumed to contain at least one segment to be labeled "stationary" and one to he labeled "transitional". To 1ocate these segments, the SLF thresholds for labeling a segment stationary resp. dynamic (see Fig. 6) arc decreased resp. increased until at least one stationary and dynamic segment is found. These subsegments then are separated from the rest which keeps its undefined labeling and can be reprocessed according to one of the four categories.

After performing this procedure, no undefined segments will remain in the course of the signal.

7. Corrections.

7.1 Parameter Errors.

Since the segmentation parameters* for better accuracy, are measured pitch-synchronously, one must regard that the speaker may utter irregular signals or that pitch detection may fail temporarily /Hess, 1974.3/. The effects of this type of error are reduced by smoothing the segmentation parameters where such an error is detected. Usually, a slight parameter smoothing is done where pitch is regular; a medium smoothing is performed on voiceless sections. Irregular sections have to he smoothed in a way that parameter extremes due to incorrect measurements are reduced to almost zero. Thus, if an obviously nonsense segment combination is detected which could have been caused by a parameter error, the involved part of the signal parameters will be smoothed as if it were irregular, and, after that, will be reprocessed by the segmentation algorithm.Among the segmentation results which will cause smoothing and reprocessing are the following combinations:

a) Clusters of more than three transitional segments

b) Failure of the algorithm to divide a non-monotonic dynamic segment into monotonic transitional ones

c) Clustering of "may-be"-segments or verv short stationary segments.

7.2 Hidden stationary segments.

Two important special cases have been remaining (Fig. 9). They represent the only errors introduced by the strong emphasis of the dynamic segments in the definition of the SLF and the minimal segments:

a) A local minimum of Q within a transitional segment, especially after a pause or a voiceless segment, points to a short glide or nasal which could not be detected due to a substantial increase in both levels L, and L. . At this point. the algorLthm inserts a "may-be-stationary" segment.

b) The sequence "transitional-rising'V'transitional-fal1ing" in stationary environment or immediately after a pause leads to a vowel being too short to form a stationary segment (e.g. a reduced vowel). Here, the algorithm inserts the missing stationary segment in the middle, if the two transitional segments are long enough.

7.3 Sensitivity of the SLF Algorithm with Respect to the Values of the Parameters.

There is one error associated with eqs . 4 and 5, that means with the definition of the SLF. If the value of the correct!on factor in eq. 4 is high, r., represents a value about prnportional to the absolute parameter change. If the value of K is situated near zero, r.. represents the relative parameter change. In both cases, the resu1ts show that the values of r., (and, subsequentlv, the SLF values) associated with a normal phoneme boundary depend on the absolute value of the signal level at that point. Therefore,besides the threshold q , the correction factor K represents an additiona1 degree of freedom which may balance the response of the system to a phoneme boundary with respect to the values of the parameters at the point regarded.

8. Results

As an example. Fig. 10 shows the segmentation of the German word "Belohigung". For each of the three parameters L_S,L_D, and Q_S , the individual minimal segment sequence together with the course of the individual parameter are depicted; besides that, the figure shows the combined SLF, the combined sequence of minimal segments as well as the final segment sequence. From this figure, the influence of the three parameters also can be seen. The signal level L that means the absolute average of the unprocessed speech signal, shows the

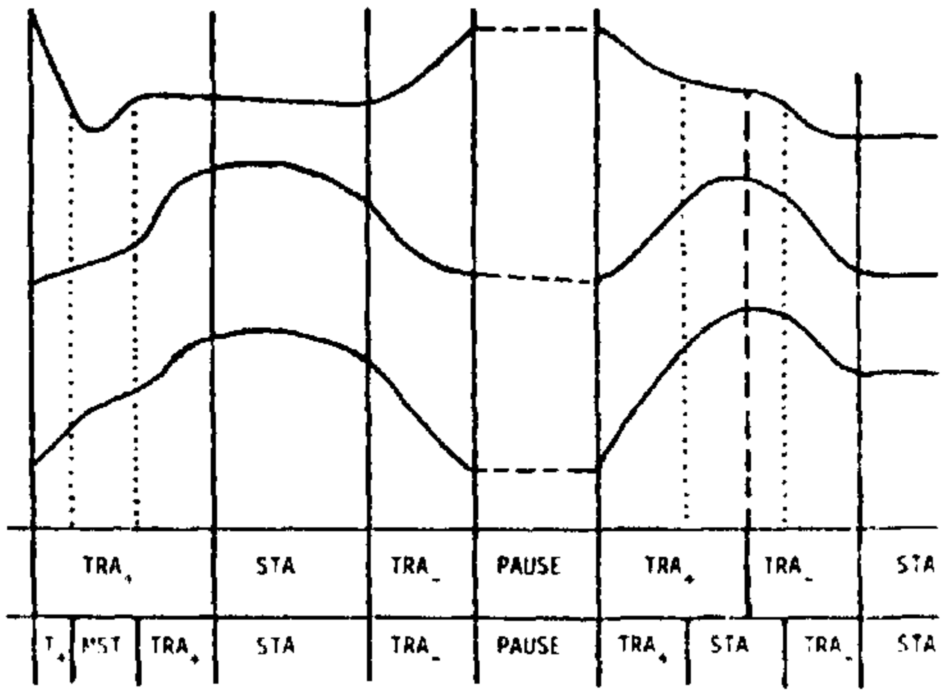


Fig. 9: Hidden stationary segments

TRA Transitional
STA Stationary
DYN Dynamic
MST "May-be"-Stationary
+,- Direction characteristics

—— Results $\frac{\text{BEFORE}}{\text{AFTER}}$ Correction

greatest changes and thus is most suitable for locating dynamic segments at stops and transitions before and after pauses and voiceless sounds. Transitions between vowels and semivowels, nasals or glides,however,are better detected by the level L_D of the differenced signal. The level ratio Q_S, at last, is most sensitive to changes in the overall frequency behaviour of the signal; for example, at the burst of the phoneme /g/.

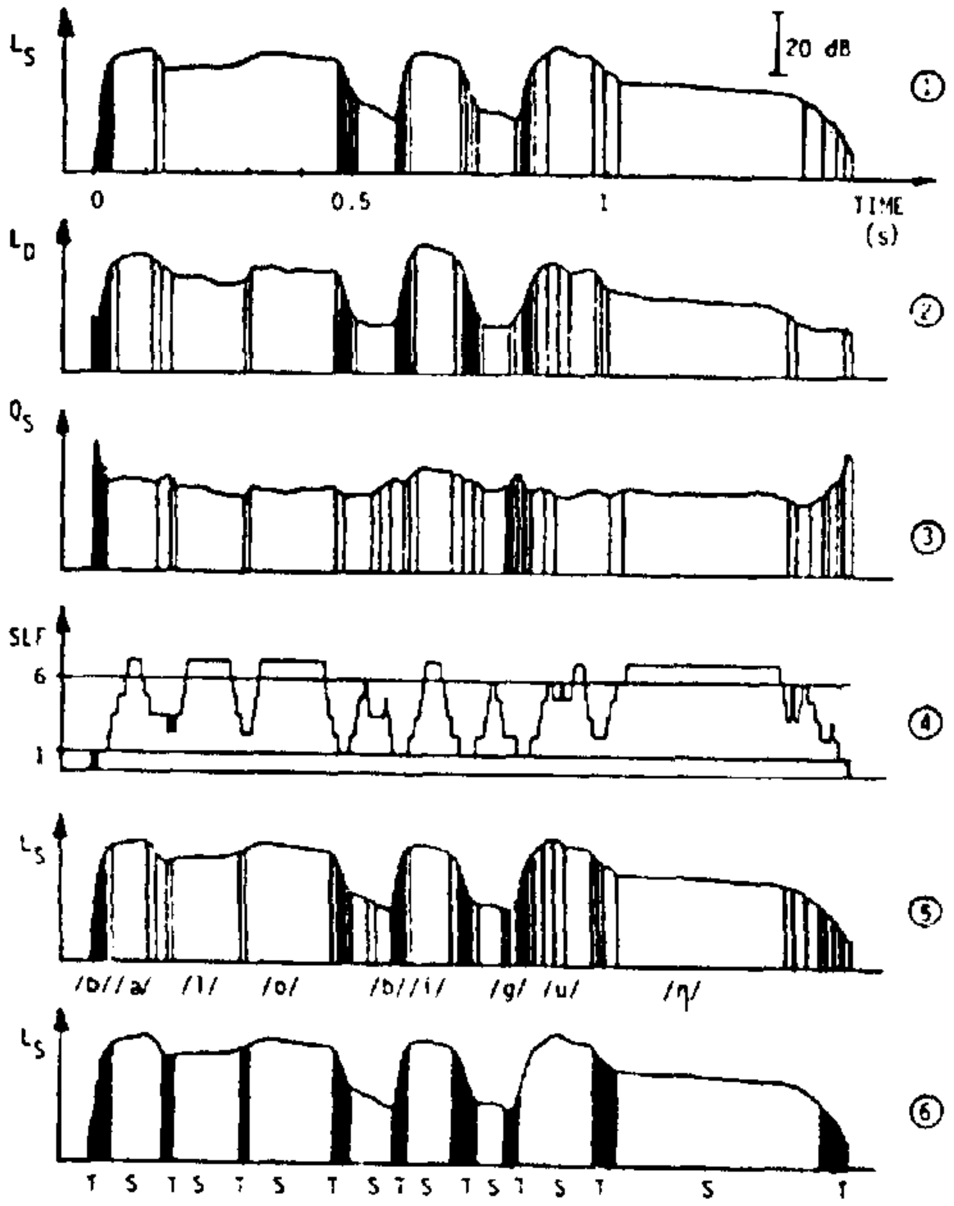


Fig. 10: Performance of the Algorithm on the German word "Belobigung"

1 Minimal segment sequence; L_S only
2 Minimal segment sequence; only L_D
3 Minimal segment sequence; Q_S only
4 Course of SLF; all parameters
5 Minimal segment sequence
6 Final segmentation
S Stationary
T Transitional

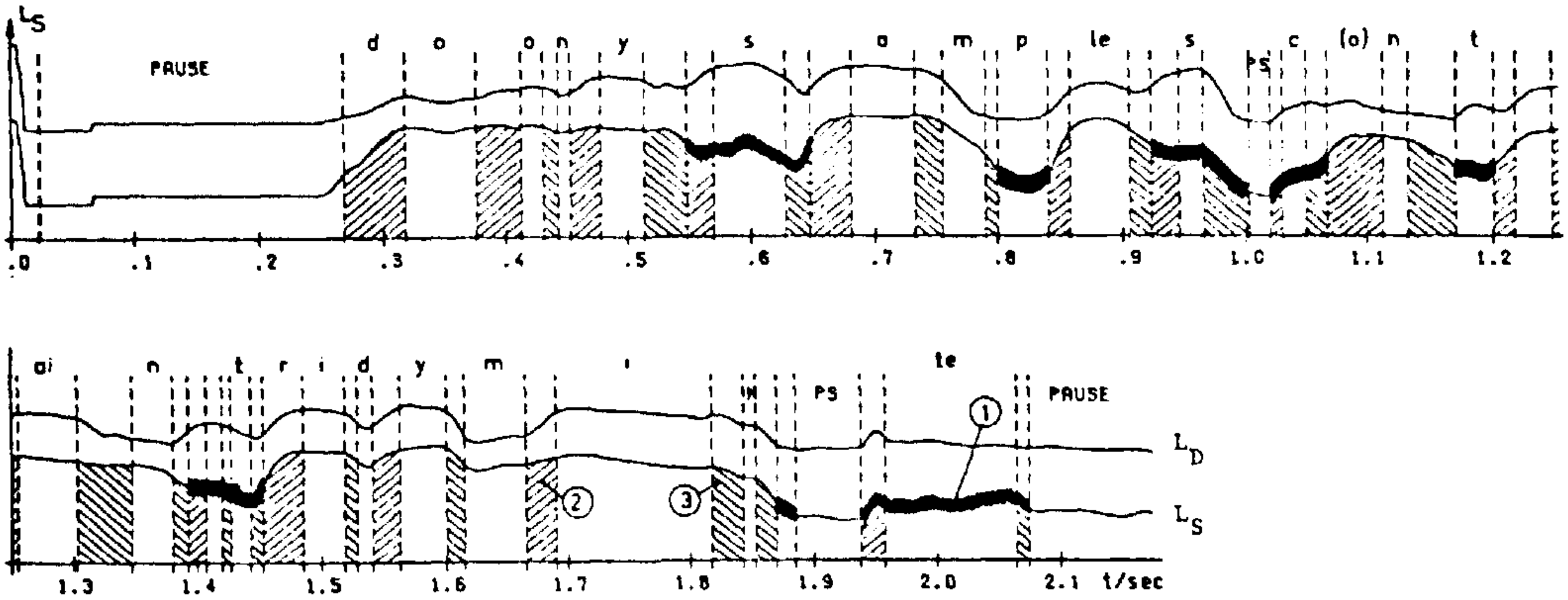


Fig.11: Results of the segmentation algorithm for the utterance "Do any samples contain tridymite?" The click at the beginning was used for synchronization. Speech material taken from /Hess,1973/

1 Voiceless
2 Transitional, rising
3 Transitional, falling

PS, PAUSE Pause
M "May-be"-segment

The segmentation procedure was applied to speech signals consisting of a list of 200 isolated German words, each uttered by seven speakers.It turned out that, from these utterances, the algorithm missed 1.3% of the phoneme boundaries and 0.9% of the stationary segments (especially short glides). The voiceless/voiced discrimination contributed 0.3% to the total error score.and the pause/speech discrimination participated with 0.8Z. A percentage of 2.1Z of the boundaries were inserted additionally thus dividing one phoneme into more than one stationary segment.

Fig. 11 shows the performance of the segmentalion procedure for a whole sentence.

When reviewing the results, one has to check whether resp. to what extent the unique relation between phoneme and stationary or transitional segment arising from the segment definition and the phonetic content leading to a "reference segmentation" is realized by the speaker at all. For example, several allophones had to be admitted for the phoneme /r/ and for the diphthongs. The error rates given above were measured in comparison with the results of a hand segmentation of the actual signals. With respect to this, the results seem promising, and what limits the value of this (and almost any) segmentation procedure is not so much the inaccuracy of the method or of the algorithm, but inaccuracies introduced by the speakers, such as additional spoken phonemes, phoneme reduction, and erroneous pronunciat ion.

Acknowledgement.

The author thanks Dr.-Ing.R.Schrag and Dr.-Ing.H. Seidel for many fruitful discussions, and Mrs. I. Kunert for preparing the manuscript- This work was partially supported by Deutsche Forschungsgemeinschaft (Germen Research Association) within SFB 50 Kybernetik (Special Research group 50: Cybernetics). Parts of the algorithm were run on the TR 440 computer of the I)RF Informatik (National Research Program in Computer Science).

References.

/Bhimani,1963/ B.V.Bhimani: Multidimensional Model for Speech Recognition. Def.Doc.Center Alexandria (Va.), USA, 1963.

/Denes,1969/ P.B.Denes, T. von Keller:Articulatory Segmentation for Automatic recognition of Speech. Proc.6th ICA,Tokyo 1968: American Elsevier Publ.Comp. 1969.

/Delattre,1968/ P.Delattre: From acoustic cues to distinctive features. Phonetica 18 (1968), pp. 198...230

/Fant,1961/ G.Fant and B.Lindblom: Studies of Minimal Speech Sound Units. STL-QPSR 1961 Nr. 2, pp.1...10

/Flanagan,1972/ J.L.Flanagan: Speech Analysis, Synthesis, and Perception. Springer-Verlag, Berlin, Heidelberg, New York 1972 (2nd ed.)

/Hess,1972/ W.Hess: Digitale grundfrequenzsynchrone Analyse von Sprachsignalen als Teil eines automatischen Spracherkennungssystems. Dr.-Ing. thesis, Munich 1972.

/Hess,1973/ W.Hess: Time-domain segmentation of speech signals. Contribution to a workshop on segmentation and labeling of speech signals, held at Carnegie-Mellon University, Pittsburgh, Pa., USA, in July 1973 (proceedings to appear in 1975).

/Hess,1974.1/ W.Hess: A pitch-synchronous, digital feature extraction system for phonemic recognition of speech. In: Proceedings of the IEEE symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, Pa, USA, 15-19 April 1974. Ed. by L.Frman. IEEE Press, New York 1974.

/Hess,1974.2/ W.Hess: A pitch-synchronous, digital feature extraction system for phonemic recognition of speech. In: Proceedings of the Speech Communication Seminar (SCS) Stockholm, Aug. 1-3, 1974. Stockholm 1974: Almqvist and Viksell.

/Hess,1974.3/ W.Hess: On-line, digital pitch period extractor for speech signals. In: Proceedings of the 1974 International Zurich Seminar on Digital Communications, paper A5. 1974: Zurich, Switzerland.

/Hughes,1965/ G.W.Hughes, J.F.Hemdal: Speech Analysis. Purdue Res. Found. Techn. Rept. TREE 65-9, 1965.

/Lea,1972/ W.A.Lea: An Approach to Syntactic Recognition without phonemics. In: Conference Record, 1972 International Conference on Speech Comm. and Processing. AFCRL, Bedford, Mass., USA 1972.

/Olson,1967/ H.F.Olson et al.: Speech Processing Techniques and Applications. IEEE Trsa. AU-15 (1967), pp. 120...126

/Öhman,1962/ S.E.G.Öhman: Perceptual Segments and rate of change of spectrum in connected speech. Proceedings of the Speech Communication Seminar, Stockholm 1962.

/Otten,1964/ K.W.Otten: Segmentation of continuous speech into phonemes. Rept.Nr.RTD-TDR-63-4005, Part 2, U.S.Army, 1964.

/Paulus,1974/ F.Paulus, R.Schrag, and Th. Schotola: Advances towards a four-stage system for automatic speech recognition. In: Proc. of the Speech Communication Seminar, Stockholm 1974.

/Pols,1969/ Pols,L.C.W., J.T. van der Kamp, and R. Plomp: Perceptual and Physical Space of Vowel sounds. J.A.S.A. 46(1969), pp. 458...467

/Reddy,1966.1/ D.R.Reddy: Segmentation of speech sounds. J.A.S.A. 40 (1966), pp. 307...312

/Reddy,1966.2/ D.R.Reddy: Phoneme Grouping for Speech Recognition. J.A.S.A. 41 (1967), pp. 1295...1300

/Reddy and Vicens,1968/ D.R.Reddy and P.Vicens: A Procedure for the Segmentation of connected speech. J.Audio Eng.Soc. 16(1968), pp.404...411.