Lawrence R. Rabiner (S'62-M'67-SM'75-F'75), for a photograph and biography, see p. 726 of the June 1983 issue of the TRANSACTIONS.

Thomas B. Martin was born in Burlington, IA, on June 23, 1935. He received the B.S.E.E. degree from the University of Notre Dame, South Bend, IN, in June 1957 and the M.S.E.E. and Ph.D. degrees from the University of Pennsylvania, Philadelphia, in 1960 and 1970, respectively.

In 1970 he cofounded Threshold Technology Inc., which developed the first practical automatic speech recognition systems to be used in industry. In March 1982, he joined Bell Laboratories, Murray Hill, NJ, and has continued to work in speech recognition.

On Temporal Alignment of Sentences of Natural and Synthetic Speech

HANS D. HÖHNE, CECIL COKER, STEPHEN E. LEVINSON, SENIOR MEMBER, IEEE, AND LAWRENCE R. RABINER, FELLOW, IEEE

Abstract—One way to improve the quality of synthetic speech, and to learn about temporal aspects of speech recognition, is to study the problem of time aligning pairs of spoken sentences. For example, one could evaluate various sets of duration rules for synthesis by comparing the time alignments of speech sounds within synthetic sentences to those of naturally spoken sentences. In this manner, an improved set of sound duration rules could be obtained by applying some objective measure to the alignment scores. For speech recognition applications, one could obtain automatic labeling of continuous speech from a hand-marked prototype to obtain models and/or statistical data on sounds within sentences. A key question in the use of automatic alignment of sentence length utterances is whether the time warping methods, developed for isolated word recognition, could be extended to the problem of time aligning sentence length utterances (up to several seconds long). A second key question is the reliability and accuracy of such an alignment. In this paper we investigate these questions.

It is shown that, with some simple modifications, the dynamic time warping procedures used for isolated word recognition apply almost as well to alignment of sentence length utterances. It is also shown that, on the average, the uncertainty in the location of significant events within the sentence is much smaller than the event durations although the largest errors are longer than some event durations. Hence, one must apply caution in using the time alignment contour for synthesis or recognition applications.

Manuscript received May 5, 1982; revised November 24, 1982.

H. D. Höhne was on leave at Bell Laboratories, Murray Hill, NJ 07974. He is with the Technische Universitat, Berlin, West Germany.

C. Coker, S. E. Levinson and L. R. Rabiner are with Bell Laboratories, Murray Hill, NJ 07974.

I. INTRODUCTION

THE state of the art in speech synthesis by rule is that one can synthesize (from either printed text or from a phonetically based set of input symbols) speech whose intelligibility is quite high but whose naturalness is often poor. One reason for the unnatural quality is the rudimentary state of knowledge as to how to properly control pitch and duration of sounds within a sentence. In order to make improvements in the pitch and duration rules used for synthesis, it would be helpful to be able to compare rule generated synthesis of sentence length material to natural productions of the same sentences. By time aligning events within the sentence, one could modify the duration rules of the synthesizer to improve the quality of the match. By experimenting with a number of sentences (and talkers), one could, hopefully, make major improvements in the duration rules of the synthesizer.

Another area that would benefit from the ability to time align a spoken sentence with another spoken version of the same sentence is speech recognition. One of the most difficult and time consuming problems in building a speech recognizer is collecting data for modeling (statistically or otherwise) the properties of speech sounds. By carefully hand labeling a set of test sentences, one could, in theory, automatically obtain a good set of labels on repetitions of the test sentences by using a time alignment procedure. The time aligned events of

0096-3518/83/0800-0807$01.00

the sentences then constitute a training set of data for the recognizer.

The key point in both the above applications is having the ability to take a naturally spoken sentence and time align it with either a synthetically generated sentence (to evaluate duration rules), or another naturally spoken sentence (to time align events within the sentence). Thus, the problem at hand is how to time align two sentence length utterances and, having performed this alignment, how to objectively evaluate how well the alignment procedure worked.

One very simple solution to the time alignment problem for sentences is to use standard isolated word dynamic time warping (DTW) algorithms, and modify them appropriately for sentence inputs. This approach has been shown to be successful for aligning sequences of demisyllables to form words [1], [2] and for sequences of digits and letters to form connected digit strings and name strings [3]-[6].

An alternative method of automatically obtaining label positions in speech was described by Wagner [7]. Wagner used the features "silence," "fricative," and "voiced," extracted from the speech signal by means of signal processing techniques, to align the sequence of extracted features with the phonetic transcription of the speech using a dynamic programming algorithm. The local distances between the extracted features and the element on the phonetic transcription were defined in accordance with acoustic-phonetic rules.

The approach proposed here for aligning sentences on the basis of their parametric representations can be used for the same purpose as Wagner's by carefully hand labeling a first production of a sentence and then automatically getting labels on all reproductions of the sentence. Since the DTW algorithm can (as will be shown here) have an average error in alignment of less than one frame, this method would have a high accuracy in labeling.

The main objective of the method presented in this paper, however, is to study a tool for aiding speech synthesis and recognition, e.g., for obtaining improved duration rules for synthesis or for automatic training for recognition. As such, we will be primarily concerned with the ability of our algorithms to time align pairs of sentences representing the same acoustic events.

In this paper we discuss the issues involved in using DTW time alignment algorithms on sentence length utterances. In Section II we review the DTW algorithm and describe the system used for time alignment. In Section III we describe an evaluation of the sentence alignment procedure using both synthetic and natural speech utterances. In this section we discuss the effects of the DTW parameters on the performance scores. Finally in Section IV we discuss the results and their implications for synthesis and recognition applications.

II. System Used to Time Align Sentences

It is assumed that, in order to solve the time alignment problem, we work with parameterized representations of the two sentences being aligned. We call one sentence the test T and denote its parameterization as

$$T = \{T(1), T(2), \cdots, T(N)\} \tag{1}$$

where $T(n)$, $n = 1, 2, \cdots, N$ are feature vectors of each frame of the test sentence. For our applications $T(n)$ represents an eighth-order linear predictive coding (LPC) analysis of each 45 ms frame of speech, with consecutive frames spaced 15 ms apart.

Similarly, we call the second sentence (the one to be aligned with T) the reference R and denote its parameterization as

$$R = \{R(1), R(2), \cdots, R(M)\}. \tag{2}$$

For a typical 4-6 s sentence, values of M and N in the range 300-400 are obtained.

A time alignment of R to T is a mapping of the form

$$m = w(n) \tag{3}$$

where m represents the reference frame index, and n represents the test frame index. A wide class of algorithms, called dynamic time warping algorithms, can be used to find the optimal function of the form of (3) which minimizes an accumulated distance along the warping path [8], [9]. For LPC feature sets, the local distance (for comparing a given test frame to a given reference frame) is the log likelihood ratio as proposed by Itakura [9] of the form

$$d(T, R) = \log \left[\frac{a_R V_T a_R^t}{a_T V_T a_T^t} \right] \tag{4}$$

where a_R, a_T are the LPC coefficient vectors of the reference and test frames, V_T is the (Toeplitz) autocorrelation matrix of the test speech segment, and t denotes vector transpose. It should be noted that the local distance of (4) is asymmetric as only the autocorrelation matrix of the test segment is used. (Another source of asymmetry in the DTW implementation arises from the placement of the test along the abscissa and the reference along the ordinate [10].)

Fig. 1 illustrates a typical time alignment of R to T. Shown in this figure is the warping function $w(n)$. Assume now that the reference sentence has a series of acoustic/linguistic events that have been previously marked (either by careful hand analysis, or as generated by a speech synthesizer). We denote these events as the points marked A, B, C, D, and E in Fig. 1. From the alignment contour we can now segment the test utterance into corresponding events by solving for

$$n = w^{-1}(m) \tag{5}$$

giving event times $\hat{A}$, $\hat{B}$, $\hat{C}$, $\hat{D}$, and $\hat{E}$, as shown schematically in Fig. 1.

The example of Fig. 1 illustrates the potential power of being able to time align pairs of sentences. From a single carefully marked token of a sentence, one could conceivably obtain statistical data on sound durations for several repetitions of the sentence by the same or different talkers. The key to the whole process is the ability to time align two sentences, and to verify that the alignment path has physical validity.

The specific DTW algorithm that was used in this investigation is a version of the UELM (unconstrained endpoint, local minimum) algorithm of Rabiner *et al.* [10]. For this algorithm the warping path w satisfies the local path constraint

$$1 \leqslant w(1) \leqslant \delta + 1 \tag{6}$$

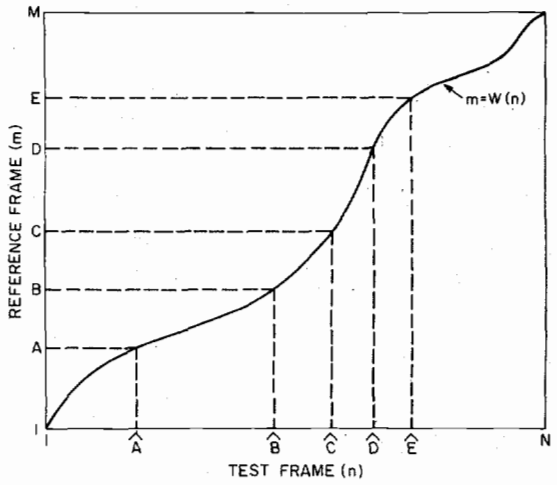

Fig. 1. Example illustrating time alignment of two sentences with internal events A, B, C, D, E aligning to $\hat{A}, \hat{B}, \hat{C}, \hat{D}$, and $\hat{E}$.

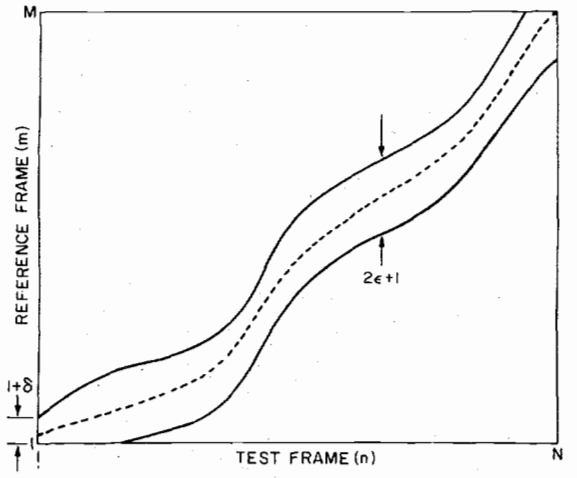

Fig. 2. A typical time alignment using the UELM algorithm showing the parameters δ and ϵ.

at the initial path boundary, i.e., the first reference frame (which aligns with the first test frame) can lie anywhere within a $(\delta + 1)$ frame region at the beginning of the reference pattern. The global path constraints on the region in which the warping path can lie can be expressed in the form

$$m_0 - \epsilon \leqslant m \leqslant m_0 + \epsilon \tag{7}$$

where

$$m_0 = \underset{m}{\text{argmin}} \, [D_A(n-1, m)] \tag{8}$$

and $D_A(n-1, m)$ is the total accumulated distance along the warping path from the initial path point to the grid point $(n-1, m)$, i.e.,

$$D_A(n-1, m) = \sum_{i=1}^{n-1} d(i, w(i)) \Big|_{w(n-1)=m}. \tag{9}$$

The quantities δ [of (6)] and ϵ [of (8)] are parameters of the UELM algorithm, and a choice of values for these parameters will be discussed later.

The reason for using the UELM algorithm is to sharply reduce computation for the case when N and M are large (as for sentences) since the DTW computation for a given value of n is only carried out for a total of $(2\epsilon + 1)$ grid points. When $\epsilon \ll M$ substantial savings in computation results, with essentially no loss in accuracy in finding the best path. Fig. 2 illustrates a typical search using the UELM algorithm. By following the local minimum, within a reasonable range ϵ, the algorithm can generally examine a wide range of paths.

In the experiments to be described in Section III, the test sentence was the independent variable and was mapped to the abscissa of the warping plane, and the reference sentence was the dependent variable and was mapped to the ordinate of the warping plane. The local distance of (4) was used in all warps.

One point is worth noting about the UELM solution. It should be clear from the above discussion that the UELM optimal path need not end at the grid point (N, M). The optimal path can terminate at an interior point (n_0, M), i.e., the reference pattern ends before the test pattern is used up, or at an interior point (N, m_0), i.e., the test pattern ends before the reference pattern is used up. So long as n_0 is close to N (or m_0 is close to M), there is no problem with the premature termination. However, in cases where $n_0 \ll N$ or $m_0 \ll M$, it is unclear what meaning, if any, to attach to the alignment. We shall discuss this key point again later.

III. Evaluation of DTW-Based Sentence Alignment

In order to evaluate the performance of the DTW-based sentence alignment system of Section II, two different test sentences were chosen for investigation. These were the sentences.

*S*1—This is a test of automatic labeling by dynamic time warping to synthetic speech.

*S*2—Red orange buttons frequently bother all bottled up actors that must wear them.

For sentence *S*1, 27 distinct acoustic events were identified as candidates for labeling. These were the sounds (using ARPABET notation).

/DHIHS/IHZ/AX/TEHST/AXV/AODX/AX/
/MAE/DXIHK/LEYB/BEL/LIHNX/BAY/
/DAY/NAE/MIHK/TAY/MW/AORP/IHNX/
/TUW/SIHN/THEH/DXIHK/SP/IY/CH/

Similarly for sentence *S*2, 20 distinct acoustic events were identified, namely,

/REHD/AOR/AXNJH/BAH/TAXNZ/FRIY/
/KWEHNT/LIY/BAO/DHER/AOL/BAA/
/DXL/DAXP/AEK/TERZ/DHAET/MAXST/
/WEHR/DHEHM/

Clearly some of the events can (and often will) be missing in some pronunciations of the sentences. Such missing sounds must be accounted for by the time alignment procedure.

A total of 24 versions of each sentence were recorded. 20 of the versions were produced by ten talkers (seven men, three women, all native speakers of general American English) who each spoke the two sentences twice. The remaining four versions of each sentence were generated synthetically by rule from a text-based system [11], and a demisyllable-based system [12]. Two sets of duration rules were used in each synthesis giving two versions of each sentence.

The four synthetic versions of sentence $S1$ had an average duration of 6.02 s with a standard deviation of 0.897 s; the 20 natural versions of $S1$ had an average duration of 5.12 s with a standard deviation of 0.472 s. For sentence $S2$, the synthetic versions were 5.55 s average duration with 0.210 s standard deviation, and the natural sentences were, on average, 5.03 s in duration with a standard deviation of 0.687 s. Hence, we see that, in general, the synthetic sentences were longer in duration than the naturally spoken sentences.

The variability in the rate or pronunciation of the 24 repetitions of each sentence $S1$ and $S2$ is reflected in the fact that for sentence $S1$ (with 22 syllables) each syllable took an average of 222 ms ± 27 ms, whereas for sentence $S2$ (with 20 syllables) each syllable took an average of 256 ms ± 33 ms.

A. Labeling the Positions of Acoustic Events Within the Sentences

For each of the 24 versions of each sentence, an interactive display and playback system was used to identify the positions of each of the acoustic events within the sentences. Fig. 3 shows one example of the labeling of sentence $S1$. Shown in this figure is the energy contour of the sentence, along with vertical markers of the 27 acoustic events in the sentence. The marker positions were generally chosen at places where the energy curve of the sentence had a distinct dip. This occurred primarily at syllable boundaries. After tentatively choosing a marker position, the sentence was played from the beginning to the current marker. If by listening it was found that the expected syllable was either chopped too short, or extended into the next sound, the marker position was changed and the playback repeated. This iterative procedure continued until all marker positions were found. On average it took about 30 min to locate all marker positions for a sentence.[1]

Since the goal of our study was to find stable highly reproducible positions for all the markers of a sentence in order to evaluate the capability of aligning the sentences by DTW (as opposed to aligning a transcription with phonetic events) no criterion other than good reproducibility of marker positions was used for defining the segments. Formal tests on relabeling by the same experimenter indicated very low error in market locations (on the order of one frame), thereby indicating a stable criterion on the part of the experimenter for locating events.

[1] When a label was missing (i.e., the talker did not use one of the sounds postulated by the linguistic analysis above) a pseudolabel was interpolated to keep label consistency among all versions of the sentences.

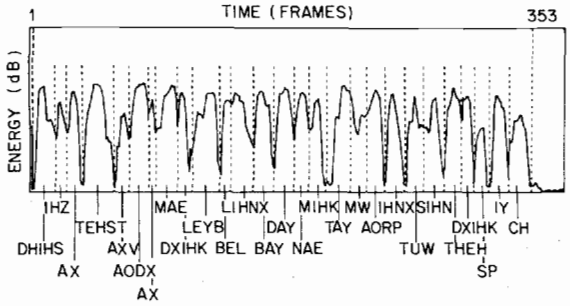

Fig. 3. Example of hand generated labels for a test sentence.

B. Evaluation of Hand Labeling

For a given utterance T we denote the set of Q hand labels (i.e., the frames of each marked event) as n_i, $i = 1, 2, \cdots, Q$. After time aligning a reference utterance R with T, we have the warping path $m = w(n)$, leading to an aligned set of labels

$$m_i = w(n_i), \quad i = 1, 2, \cdots, Q. \tag{10}$$

By hand labeling, for reference R, we have identified the marked events $\hat{m}_i$, $i = 1, 2, \cdots, Q$. Hence, we can define a nominal error p_α as

$$p_\alpha = \left[\frac{1}{Q}\sum_{i=1}^{Q} |m_i - \hat{m}_i|^\alpha\right]^{1/\alpha} \tag{11}$$

where $\alpha = 1$ is the average absolute error, $\alpha = 2$ is the root-mean-square error, and $\alpha = \infty$ is the peak error (the Chebyshev norm).

We can also apply a normalization to the error (to account for speed of speech) of the form

$$\hat{p}_\alpha = \frac{p_\alpha \cdot \bar{N}}{N} \tag{12}$$

where $\bar{N}$ is the average duration (over all replications) of the test sentence, and N is the duration of the actual test sentence used in the DTW warp.

By only considering the 20 pairs of sentences consisting of two tokens of each sentence by each talker, and by calculating the $\hat{p}_\alpha$'s for only these pairs of sentences, a gross measure of the accuracy of hand labeling can be derived. The results of such a test are given in Table I which shows (for $\hat{p}_1$, $\hat{p}_2$, and $\hat{p}_\infty$) the minimum, maximum, and average (over the ten talkers) of each statistic. We make the assumption here that the path found by the DTW of two versions of the same sentence by the same talker gives the "correct" alignment path; hence, deviations of the expert labels from the warping path give a good measure of the error of the hand labeling procedure. As seen in Table I, the average absolute frame error is 0.93 frames (≈14 ms), and the average Chebyshev error is 3.83 frames (≈57 ms). Thus, in general, the average frame error for hand labeling (or equivalently for DTW alignment) is quite small; however, infrequent large frame errors (somewhat longer than transient sounds) can and do occur.

An analysis of the gross labeling errors in this experiment indicated the following problems.

TABLE I
STATISTICS OF ERRORS FOR PARTS OF SENTENCES BY THE SAME TALKER

Statistic	Minimum	Maximum	Average
$\hat{p}_1$	0.57	1.37	0.93
$\hat{p}_2$	0.17	0.40	0.26
$\hat{p}_\infty$	2.09	7.84	3.83

1) In some cases pairs of syllables were heavily coarticulated in one repetition by a talker but not in the second repetition.

2) In some cases the talker inserted mouth clicks or pops at points within the sentence.

3) In some cases a syllable in one replication was more than two times larger than the same syllable in the second replication. Thus, the local constraints of the DTW were insufficient to account for this gross local deviation of the warping curve.

Although all of the above problems led to individual gross errors in label positions, they did not greatly affect the overall accuracy of the alignment procedure. Alternative approaches to DTW, such as the normalize-and-warp procedure [13] would not have corrected any of the above problems.

C. Examples of Warping Outputs

Each of the 24 versions of the two sentences was time aligned with every other version of the same sentences. The DTW warping parameters ϵ and δ were varied to optimize the match (on average) between test and reference. We discuss these parameters in Section III-D. When the warp parameters were properly selected, most of the warps were successful.

Fig. 4 illustrates one such case. Shown in this figure are the accumulated distance function, the warping path, and the hand-marked labels for warping two tokens of sentence *S*1. For this example the test duration was 450 frames (6.7 s) and the reference duration was 399 frames (5.95 s). The test utterance was one of the synthetically generated sentences whereas the reference utterance was a naturally spoken one. Fig. 4 shows that the accumulated distance grows almost linearly throughout the sentence with an average frame distance of about 0.55. The warping path goes through (to within ± 1 frame) most of the labeled points of the sentence.

Fig. 5 illustrates a failure of the DTW alignment procedure. In this case the test was a 296 frame naturally spoken sentence and the reference was the 450 frame synthetic sentence used in Fig. 4. It can be seen in Fig. 5 that the warping path reaches the end of the test long before the end of the reference pattern. The average distance along the path is about 1.1 indicating very poor matching of the utterances. It is also seen that the warping path does not go through any of the hand labels.

For cases such as the one shown in Fig. 5, there is very little one can do to improve the match. Fortunately, these cases do not occur often and when they do occur they signal themselves by having high average distance scores. In the experiment all pairs of the 24 utterances (both for *S*1 and *S*2) were processed. Warp failures almost never occurred when comparing natural to natural or synthetic to synthetic sentences. Failures

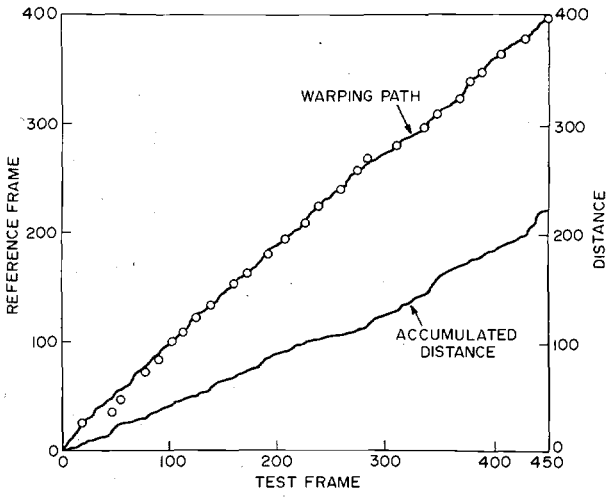

Fig. 4. Time alignment and accumulated distance for a synthetic test and a natural reference utterance. Circles show the location of hand-marked labels.

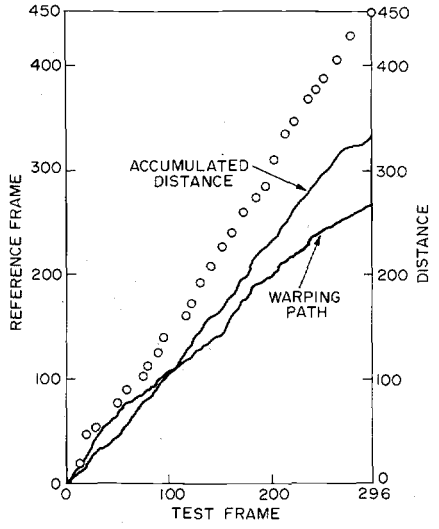

Fig. 5. Time alignment failure for a natural test and a synthetic reference utterance. The alignment failed since the test finished long before the reference finished. Circles show the location of hand-marked labels.

occurred in about half of the cases when the synthetic utterance was the reference, and essentially did not occur at all when the synthetic utterance was the test. This asymmetry between test and reference has been noted [10].

D. Effects of DTW Parameters on Performance

Figs. 6 and 7 show plots of the effects of varying δ and ϵ on the performance measures, $\hat{p}_1$, $\hat{p}_2$ and $\hat{p}_\infty$. For the data of Fig. 6 a value of $\epsilon = 50$ was used, and δ was varied from 1 to 50. The curves here show that for $\delta \geq 5$, no change in performance is obtained. Thus, a starting alignment region of $\delta \approx 5$ frames is adequate for all warps.

Fig. 7 shows the effect of varying ϵ (the range width) on the performance scores. For this case the value of δ was set to 5 and ϵ was varied from 5 to 100. It can be seen that for small values of ϵ, the performance is very poor. For example, for $\epsilon = 5$ the value of average frame error $\hat{p}_1$ is about 9.5 frames.

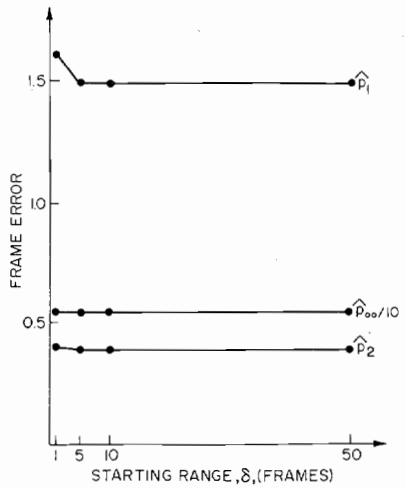

Fig. 6. Average frame error versus starting range parameter, δ, for three performance measures.

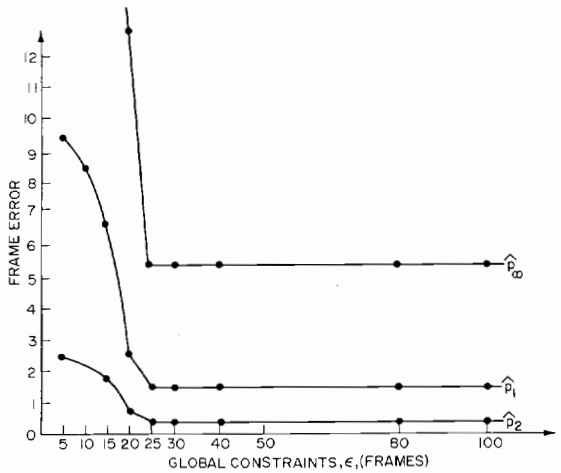

Fig. 7. Average frame error versus range parameter ϵ for three performance measures.

The curves of Fig. 7 show that for values of $\epsilon \geqslant 25$ the performance stabilizes at the best values. The results for $\epsilon = 25$ show an average absolute frame error of 1.7 frames (25 ms) in label position.

IV. Discussion

The purpose of the investigation was to find out how well an automatic time alignment procedure could time align events within long sentences. To study this question two long test sentences were used. Both sentences were designed so that a wide range of pronunciations were possible. Both synthetic and natural versions of each of the test sentences were produced. Each sentence was hand labeled to identify location of acoustic events within the utterance. A simple check on the hand labeling indicated fairly good alignment between replications by the same talker. A complete set of alignment runs were made on 24 versions of each utterance against all other versions of the same utterance. Based on the results given in Section III, the following conclusions are drawn.

1) It is possible to reliably time align two versions of a sentence produced by essentially any talker—in spite of the variability of pronunciations and talkers.

2) In some cases it is possible to reliably align a synthetic sentence with a naturally produced version of the same sentence. For best alignments it is preferable for the synthetic utterance to be the test and the naturally produced utterance to be the reference. This forces every frame in the natural utterance to be aligned with some frame of the synthetic one.

3) For cases when the DTW algorithm is able to time align pairs of sentences, the reliability of the internal alignment points is quite good with an average duration error on the order of 25 ms or less.

4) For purposes of synthesis, if one is careful in carrying out the time alignments, one could gain great insight into the mechanics of durations of sounds within natural sentences. Hence, it should be possible to greatly improve synthesis-by-rule duration rules in an interactive mode based on DTW alignments with naturally produced sentences.

5) For purposes of recognition, it seems clear that one could gather a wide variety of statistics on speech sounds in the context of a sentence by using DTW alignments of a set of natural productions of a sentence against hand labeling of a single naturally produced sentence.

For cases in which the DTW alignment failed, alternative alignment procedures could be considered. However, it is not felt that such efforts are justified in that the rate of such failures is low and the failure mechanism is well understood and easily detected automatically.

In summary we have shown how standard isolated word time alignment procedures can be extended to the case of time aligning productions of sentence length material. The resulting time alignments indicate good accuracy of aligning events within the sentence.

The way in which the proposed sentence alignment procedure could aid speech synthesis is to begin with an automatically labeled synthetic sentence, or a hand labeled natural sentence, and then time align multiple repetitions of the sentence by different talkers. In this manner, highly reliable duration statistics on syllables, words, phrases, etc., could be obtained which could then be used to improve word duration rules for speech synthesis by rule.

For speech recognition the proposed sentence alignment procedure could be used to automatically obtain a training base of syllables, words, phrases, etc., as extracted from sentences. Such a training base could be used for connected and continuous speech recognition applications.

Acknowledgment

The authors acknowledge Dr. J. L. Flanagan for proposing this topic of investigation, and C. P. Browman for providing tokens of demisyllable based synthetic speech. Thanks also go to C. Myers and A. E. Rosenberg for technical discussions during the course of the work.

References

[1] L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, and T. M. Zampini, "A bootstrapping training technique for obtaining demisyllable reference patterns," *J. Acoust. Soc. Amer.*, vol. 71, pp. 1588–1595, June 1982.

[2] A. E. Rosenberg, L. R. Rabiner, J. G. Wilpon, and D. Kahn, "Demisyllable based isolated word recognition system," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, pp. 713–725, June 1983.
[3] H. Sakoe, "Two level DP matching–A dynamic programming based pattern matching algorithm for connected word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 588–595, Dec. 1979.
[4] C. S. Myers and L. R. Rabiner, "Connected digit recognition using a level building DTW algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 351–363, June 1981.
[5] —, "An automated directory listing retrieval system based on recognition of connected letter strings," *J. Acoust. Soc. Amer.*, vol. 71, pp. 716–727, Mar. 1982.
[6] L. R. Rabiner, J. G. Wilpon, and A. Bergh, "An improved training procedure for connected digit recognition," *Bell Syst. Tech. J.*, vol. 61, pp. 981–1001, July–Aug. 1982.
[7] M. Wagner, "Automatic labeling of continuous speech with a given phonetic transcription using dynamic programming algorithms," in *Proc. ICASSP*, Apr. 1981, pp. 1156–1159.
[8] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 43–49, Feb. 1978.
[9] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 67–72, Feb. 1975.
[10] L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson, "Considerations in dynamic time warping for discrete word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 575–582, Dec. 1978.
[11] C. H. Coker, N. Umeda, and C. P. Browman, "Automatic synthesis from ordinary English text," *IEEE Trans. Audio Electroacoust.*, vol. 21, pp. 293–298, Feb. 1973.
[12] C. P. Browman, "Rules for demisyllable synthesis using lingua, a language interpreter," in *Proc. ICASSP*, Apr. 1980, pp. 561–564.
[13] C. S. Myers, L. R. Rabiner, and A. E. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 623–635, Dec. 1980.

Hans D. Höhne was born in Herzberg, Germany on September 28, 1934. He received the Dipl.-Ing. and Dr.-Ing. degrees from the Technische Universität, Berlin, Germany, in 1962 and 1971, respectively.

Since 1963 he has been a Scientist at the Heinrich-Hertz-Institut Berlin. From 1963 through 1969 he worked on human factors in telephone connections with long transmission delay and taught electrical engineering at the Staatliche Techniker Schule, Berlin. From 1970 to 1973, he did research in mathematical statistics. Since 1972 he has given lectures at the Technische Universität, Berlin. His current lecture is on digital signal processing and speech processing. From 1974 through 1982, he pursued problems in speaker and speech recognition. During that period he was a Guest Scientist at the Acoustics Research Department of Bell Laboratories, Murray Hill, NJ for half a year. Presently, he is engaged in research in digital signal transmission.

Dr. Höhne is a member of the German Nachrichtentechnische Gesellschaft NTG and of the IEEE Acoustics, Speech, and Signal Processing Society, the IEEE Computer Society, and the IEEE Information Theory Group.

Cecil Coker received the M.S. degree from Mississippi State University, Mississippi State, in 1956, and the Ph.D. degree in electrical engineering from the University of Wisconsin in 1960.

From 1954 to 1957, he did research on aircraft control systems at Mississippi State. In the 1959–1960 year, he was an Assistant Professor of electrical engineering the University of Wisconsin, Madison. He joined the Acoustics Research Department of Bell Laboratories, Murray Hill, NJ, in 1960. There, his work was centered on analysis, synthesis and feature extraction of speech, and processing of signals from multiple microphones to find the direction of the talker. He is known for his work in articulatory modeling and speech synthesis from text.

Stephen E. Levinson (S'72–M'74–SM'82) was born in New York, NY, on September 27, 1944. He received the B.A. degree in engineering sciences from Harvard University, Cambridge, MA, in 1966, and the M.S. and Ph.D. degrees in electrical engineering from the University of Rhode Island, Kingston, in 1972 and 1974, respectively.

From 1966–1969, he was a Design Engineer at the Electric Boat Division of General Dynamics, Groton, CT. In 1974–1975, he held a J. Willard Gibbs Instructorship in Computer Science at Yale University, New Haven, CT. In 1976, he joined the Technical Staff at Bell Laboratories, Murray Hill, NJ, where he is pursuing research in the area of speech recognition and cybernetics.

Dr. Levinson is a member of the Association for Computing Machinery and the Acoustical Society of America.

Lawrence R. Rabiner (S'62–M'67–SM'75–F'75), for a photograph and biography, see p. 726 of the June 1983 issue of this TRANSACTIONS.