Title:

A dynamic programming algorithm for time-aligning and averaging physiological signals related to speech

Running title:

non-linear time-alignment and averaging

Helmer Strik and Louis Boves

University of Nijmegen, Dept. of Language and Speech,

P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

Abstract

In analyzing physiological signals related to speech, it is necessary to average several repetitions in order to improve the Signal to Noise Ratio. However, in a recent experiment, considerable differences were found in the articulation rate of repeated realizations of a medium length utterance, especially for untrained subjects. This makes averaging of related physiological signals a non-trivial problem. A new method of time-alignment and averaging of the physiological signals is described. In this method a dynamic programming algorithm is used, which succesfully corrects for the timing differences between the repetitions.

1. Introduction

A quantitative study of the physiological basis of speech production requires the simultaneous measurement of acoustic signals and a number of physiological signals. The usual procedure to overcome the limitations of low Signal to Noise Ratios in physiological signals, and to avoid misinterpretations caused by idiosyncrasies of single tokens, is to average multiple repetitions of the 'same' utterance (Atkinson, 1978; Baer, Gay, and Niimi, 1976; Collier, 1975; Maeda, 1976). To allow averaging, the utterances must be lined up in time. To that end line-up points must be defined in every repetition. Typical choices are distinctive events like the release of a plosive or the onset of voicing, preferably close to the middle of the utterance.

Time alignment and averaging cannot be applied to all speech signals in the same way. For instance, usually no averaging is applied to fundamental frequency (F_0) signals, probably because of its discontinuous nature which makes straightforward averaging questionable. Instead, the F_0 contour of one of the repetitions is chosen to represent the 'average' F_0 contour (Atkinson, 1978; Collier, 1975; Maeda, 1976).

The applicability of the method of linear time-alignment for averaging, as described above, is limited by the inherent variability of speech production. Two types of variation must be distinguished, viz. variation in speaking rate and variation in articulation. Both kinds of variation are not independent, as a pronounced change in speaking rate is likely to affect articulation as well. But for the experiments we are concerned with, the amount of change in speaking rate is such, that rate induced articulatory variations are unlikely to be a first-order effect. This paper mainly deals with techniques to overcome the effects of temporal variation.

If trained subjects are asked to utter words or short phrases, the variation in articulation speed usually remains within reasonable bounds. But even for a trained subject considerable differences were found in the speaking rate for repetitions of a medium length utterance (Strik and Boves, 1988). If the variation in the speaking

rate is large, averaging after linear time-alignment would result in signals corresponding to different articulatory events being averaged.

In this paper we propose a novel processing technique in which a Dynamic Programmig (DP) algorithm is used to time-align the tokens in a non-linear way. The aim of this method, which is referred to as the method of non-linear time-alignment and averaging, is to obtain such a degree of time-alignment that meaningful averaging remains possible. The method was tested with the data of an experiment with (quasi-)spontaneous speech of a non-trained subject. Results of analysis with linear and non-linear time-alignment are compared.

The proposed method corrects for the variation in speaking rate, but then there is still the problem of variation in articulation. It is safe to assume that repeated realizations of the same utterance are fairly similar. However, if the variation in articulation is too large, meaningful averaging after time-alignment is never possible because then again physiological signals related to different articulatory events are averaged. The data of our experiment were also used to check a posteriori whether the amount of articulatory variation was within reasonable bounds.

2. Method of non-linear time-alignment and averaging

In the method presented in this paper DP is used for non-linear time-alignment of the tokens. DP is successfully used in speech recognition where it is often referred to as Dynamic Time Warping (DTW).

First a brief description of the DP algorithm is given. For explanation of the details of DP the user is referred to the relevant literature (e.g. Sakoe and Chiba, 1978). Next, an overview is given of the six stages of the procedure for non-linear time-alignment and averaging of physiological signals related to speech, followed by a more detailed description of the separate stages.

2.1. The DP algorithm

The algorithm, described here, is based on the flowchart given in the article of Sakoe and Chiba (1978). The DP algorithm finds the optimal time registration between two patterns, a reference pattern R

of length J and a test pattern T of length I. Both patterns are sequences of feature vectors, that are derived from the speech signals by appropriate feature extraction. The frames of the two patterns define a grid of IxJ points (Fig. 1a).

A suitable distance metric is used to calculate the distance at point p_k, which is the distance between frame i of test pattern T and frame j of reference pattern R: $d[p_k] = d[T_i,R_j]$. A path P is a sequence of K grid points (Fig. 1a): $P = p_1, p_2, p_3, \ldots, p_k, \ldots, p_K$; and $p_k = (i,j)$. The total distance between T and R for a given path P is the weighted sum of the local distances:

$$D_P[T,R] = \sum_{k=1}^{K} w_k * d[p_k].$$

By definition, the optimal path P_o is the path that minimizes $D_P[T,R]$. The path P_o represents a function F, which realizes a mapping from the time axis of T onto that of R, called the warping function. The warping function F, or the optimal path P_o, can be used to normalize the time axis of T with respect to the time axis of R. When there are no timing differences between T and R, the path P_o coincides with the line i=j.

The path P usually is constrained. The path has to start in $p_1 = (1,1)$, end in $p_K = (I,J)$, and it must remain within an adjustment window (Fig. 1a). In the method proposed in this paper a slope constraint condition of 1/2 (see Sakoe and Chiba, 1978) is used, which means that a diagonal step can be followed, or preceded, by at most 2 off-diagonal (i.e. horizontal or vertical) steps. The consequence is that only the five step sequences given in Fig. 1b are allowed. The symmetric form DP-matching is used because Sakoe and Chiba found that it gave better results in speech recognition than the asymmetric form.

2.2. General overview of the method

The method of non-linear time-alignment of physiological signals, proposed here, can be split into six successive stages:

1. specification of line-up points,
2. selection of a reference pattern,
3. calculation of cepstrum coefficients of the acoustic signals,
4. calculation of a warping function for each token (DP),
5. mapping of the physiological signals, using the warping function, and
6. calculation of median values and variation of time-normalized signals.

A necessary requirement for this method is that all (physiological) signals be sampled at the same sampling frequency (F_S). For the experiment used for evaluation of the method, F_S is 200 Hz, so the sampling time ($T_S = 1/F_S$) is 5 ms. The individual stages are described below.

2.2.1. Specification of line-up points

Even though DP has proved to be useful in speech recognition, for the purpose at hand some modifications seemed necessary. First of all, in basic speech research one is often interested in the (average) physiological signals before and after an utterance. However, it is difficult to obtain a useful time registration path by comparing silence with silence. Also, it is often desirable to have an exact time-alignment of a particular event in an utterance to study the (average) physiological signals in the neighbourhood of this event. Therefore, our method allows one to define several line-up points in an utterance, that are time-aligned exactly; the DP algorithm is only applied between those line-up points (Fig. 2). The first line-up point is interpreted as the beginning of the utterance, and the last one as the end of the utterance. Before the first line-up point, and after the last line-up point, the time registration path runs diagonally (Fig. 2).

2.2.2. Selection of a reference pattern

One of the tokens is chosen as a reference for time-normalization of the remaining tokens. The best choice for this reference pattern or template seems to be the token with median length, because it requires the least adaptation in the other tokens.

2.2.3. Calculation of feature vectors

The recording conditions during experiments in which several fysiological signals are measured often are such, that the Signal to Noise Ratio (SNR) of the audio signals is not high. The current method should also be applicable to audio signals with mediocre SNR. Cepstrum coefficients are known to give good results in speech recognition (Davis and Mermelstein, 1980; Paliwal and Rao, 1982). Therefore, the first 12 cepstrum coefficients are used as feature vectors. The speech signals were digitized with a sampling frequency of 10 kHz and submitted to a 12th order LPC analysis using a 250 point Hamming window and a window shift of T_S = 5 ms. The vectors of LPC coefficients were subsequently transformed to vectors of 12 cepstrum coefficients (Markel and Gray, 1976).

2.2.4. Determination of optimal time registration path

In the fourth stage the warping function has to be found that minimizes the distance between test pattern and reference pattern. The exact choice of the distance metric does not seem critical for our purpose. A simple Euclidian distance measure proved to be sufficient. However, the definition of the adjustment window is critical. Because there can be a substantial difference in the length of patterns under comparison we used the adjustment window shown in Fig. 1a, which is different from the one given by Sakoe and Chiba (1978).

2.2.5. Transformation of the physiological signals

The warping functions computed in the previous stage describe the differences in the temporal structure of all tokens relative to the reference token, i.e. they allow normalization of the time axes of the tokens by mapping them onto that of the reference token. Since the physiological signals are measured on the same time axis as the speech signal, their time axes can be normalized using the warp functions derived from the speech signals.

The time-normalized or warped signal W is computed from the original signal S by using a non-linear function F_n: $W(j) = F_n[S(i)]$. The calculation starts at grid point $p_K = (I,J)$, and backtracks to grid point $p_1 = (1,1)$. Because only the five step sequences given in Fig. 1b are allowed, the function F_n only has to be defined for these five partial paths. For time compression, step sequences D and E in Fig. 1b, W(j) is obtained by averaging over two and three samples respectively. For time stretching, step sequences A and B, W(j) and preceding samples are obtained by linear interpolation (Fig. 3). And for a single diagonal step, step sequence C, no local transformation of the time-axes is made.

The result is a function F_n that is defined in the following way:

Step sequence A. $W(j) = [S(i+1) + S(i)]/2;$
$W(j-1) = S(i);$
$W(j-2) = [S(i) + S(i-1)]/2$

Step sequence B. $W(j) = [S(i+1) + 2*S(i)]/3;$
$W(j-1) = [2*S(i) + S(i-1)]/3$

Step sequence C. $W(j) = S(i)$

Step sequence D. $W(j) = [S(i) + S(i-1)]/2$

Step sequence E. $W(j) = [S(i) + S(i-1) + S(i-2)]/3$

As it is impossible to determine a meaningful warping function for the silent intervals before and after the utterances, the best thing one can do is to leave the time structure unchanged. This is achieved by letting the path run diagonally (Fig. 2).

2.2.6. Averaging

For every physiological process the expected value of the time-normalized signals must be computed. We prefer the median over the arithmetic mean value, since it reduces the effect of outliers. The median signals are then smoothed. In addition to the median value, a measure of the variation around the median (the phonatory or articulatory variation) can also be important. We found that the range spanned by all but the n largest and n smallest values, where n is of course (much) less than half the number of available tokens, is a useful measure of variation.

The method of averaging, described above, is appropriate for continuous signals. But F_0, one of the signals that has received much

attention in speech research, is a discontinuous signal. For unvoiced frames F_0 was set to zero. We found that taking the median value of F_0 gives the appropriate voiced-unvoiced decision and the desired average F_0 value.

3. Experimental evaluation

To compare the methods of linear and non-linear time-alignment, data of an experiment were used in which simultaneous recordings were made of the acoustic signal, electroglottogram (EGG), lung volume (V_l), subglottal pressure (P_{sb}), supraglottal or oral pressure (P_{or}), and electromyographic (EMG) activity of the sternohyoid (SH) and vocalis (VOC) muscles. A male non-trained subject was asked to produce an utterance spontaneously. His answer was: "*Ik heb het idee dat mijn keel wordt afgeknepen door die band*" (I have the feeling that my throat is being pinched off by that band). He was then asked to repeat that sentence 29 times. All physiological signals were then pre-processed to obtain signals with a sampling rate of 200 Hz. This experiment is described in more detail elsewhere (Strik and Boves, 1991).

The original, spontaneous sentence deviated from the 29 repetitions because in the original there was a pause of almost half a second, due to a swallowing gesture of the subject. Thus, in order to minimize the risk that utterances containing different articulatory gestures were averaged, only the last 29 sentences were used for analysis.

3.1. Variation in speaking rate

The oscillograms of three audio signals are shown in Fig. 4. It is obvious that there are large differences in the durations of the utterances. The mean length of the 29 utterances was 2310 ms (sd = 130 ms), while the maximum and the minimum length were 2615 ms and 2165 ms, respectively.

The release of the /k/ of "*keel*" was used as the line-up point for the method of linear time-alignment. This line-up point was chosen because it is expected to be clearly distinguishable, and it is

situated near the middle of the sentence. The mean duration of the first part (from beginning to the line-up point) was 880 ms (sd=80 ms), with a maximum of 1075 ms and a minimum of 780 ms. The mean duration of the last part (from line-up point to the end) was 1430 ms (sd=70 ms); the maximum and minimum values were 1590 ms and 1320 ms. Therefore, one can hardly maintain that there is little variation in the temporal structure of the signals. Also, the subject increased his articulation rate as he repeated the utterances more often. But even for the last six sentences the ranges for the first and last parts were 120 ms and 90 ms, respectively. So even after numerous repetitions the variation is still so large that straightforward averaging of the tokens could result in combining physiological signals of different articulatory movements.

3.2. Method of linear time-alignment

Although we did not expect linear time-alignment to produce meaningful results, we still wanted to test its viability. In Fig. 5 the time-aligned transglottal pressure (P_{tr}) signals, corresponding to the audio signals of Fig. 4, are shown in the upper three windows. The timing differences are very large, and the time-alignment is only reasonable just before and after the line-up point. This is reflected in the average signal (Fig. 5, bottom trace) that becomes increasingly meaningless towards both beginning and end of the utterance.

In Fig. 6 the average signals are plotted for F_0, Intensity Level (IL), P_{tr}, P_{or}, P_{sb}, V_l, SH and VOC. Especially for F_0, IL and the pressure signals it is apparent that the averages are only meaningful in the direct neighbourhood of the line-up point.

3.3. Method of nonlinear time-alignment

For the method of non-linear time-alignment and averaging warping functions were calculated for all tokens using the token with median length (2295 ms) as the template. These warping functions were then used to map the physiological signals. Before averaging the signals, we checked whether the degree of time-alignment, obtained by warping the signals, was sufficient.

To that end nine labels were placed manually in all 29 tokens at marked acoustic events. Chosen were releases of unvoiced plosives, one of them being the /k/ that was used as line-up point. The line-up points were used to shift the signals, so after linear time-alignment the line-up points are perfectly time-aligned. This is shown in Fig. 7, where the fifth label is the /k/ that is used as line-up point. Away from the line-up point the degree of time-alignment diminishes. Already for the two neighbouring labels, label 4 and 6, the timing differences are fairly large. The largest timing differences were found at the beginning of the utterances. The warping functions were then used to time-align the labels, and the result is shown in Fig. 8. Apart from some inaccuracies, all labels (i.e. the corresponding acoustic events) seem to be aligned very well. Because the acoustic events of the whole sentence are time-aligned by non-linear time-alignment, meaningful averaging at this stage seems possible.

Median signals are plotted in Fig. 9. It can be seen that the median signals are not only meaningful near the line-up point, but also towards beginning and end of the utterance.

3.4. Variation in pronunciation

Non-linear time-alignment seems succesful in time-aligning the acoustical events of all utterances to a reasonable degree. However, for meaningful averaging another requirement must be fulfilled, viz. that the different realizations of the utterances are produced with essentially the same articulatory gestures. After all, averaging the physiological signals belonging to utterances that were produced very differently, is not a meaningful procedure. We cannot test whether the movements of the articulators were very much alike in the different utterances, but we can check the amount of variation of some relevant physiological signals of the speech production system between the utterances.

The dotted lines in Fig. 9 give an idea of the range of the middle 20 values at each time instant (see method). From these traces we can infer that, apart from V_l, the amount of variation of the physiological signals between the different realizations of an utterance is within reasonable bounds.

4. Conclusions and discussion

Both for untrained and trained (see Strik and Boves, 1988) subjects a substantial degree of time variation between repetitions of a medium length utterance was found. Even after numerous repetitions these timing differences did not disappear. With such differences in temporal structure, linear time-alignment and averaging no longer seems a useful procedure with which to extract meaningful relations.

A possible solution seems to be the following. Define several line-up points in each repetition, time-align these line-up points, and do linear time-alignment in between. However, the timing differences are not distributed uniformly, and therefore the number of line-up points needed to obtain a reasonable overall time-alignment would be very large.

We have shown that the method of non-linear time-alignment, presented here, works satisfactorily, despite the mediocre signal-to-noise ratio of the speech signals and the highly non-stationary character of the noise. Thus, the technique of DP, developed in the framework of automatic speech recognition, can also be a very useful tool in fundamental research for processing physiological (or comparable) signals related to speech. After time normalization, median values are obtained for all measured physiological quantities. These median values can be used for further analysis.

The method of non-linear time-alignment has some further advantages. In contrast with the method of linear time-alignment, this method also yields an average signal for F_0. Furthermore, the technique can be used (semi-) automatically, which makes it very attractive in a research situation that is characterized by the need to handle large amount of signals.

Finally, the method can be used to time-align and average all kinds of signals for which timing differences are apparent.

Acknowledgements

This research was supported by the Foundation for Linguistic Research, which is funded by the Netherlands Organization for Scientific Research, N.W.O. Special thanks are due to Harco de Blaauw who was the subject of the present experiment, to Philip Blok M.D. who inserted the EMG electrodes and the catheter, to Hans Zondag who helped in organizing and running the experiment, and to Jan Strik who assisted in the processing of the data.

References

Atkinson, J.E. (1978) Correlation analysis of the physiological features controlling fundamental voice frequency, Journal of the Acoustical Society of America, 63, 211-222.

Baer, T.; Gay, T. and Niimi, S. (1976) Control of fundamental frequency, intensity and register of phonation, Haskins Laboratory Status Report on Speech Research, SR-45/46, 175-185.

Collier, R. (1975) Physiological correlates of intonation patterns, Journal of the Acoustical Society of America, 58, 249-255.

Davis, S.B. and Mermelstein, P. (1980) Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences, IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-28, 357-366.

Maeda, S. (1976) A characterization of American English intonation. Ph.D. thesis, MIT, Cambridge.

Markel, J.D. and Gray Jr., A.H. (1976) Linear prediction of speech. Berlin: Springer-Verlag.

Paliwal, K.K. and Rao, P.V.S. (1982) Evaluation of various linear prediction parametric representations in vowel recognition, Signal processing, 4, 323-327.

Sakoe, H. and Chiba, S. (1978) Dynamic programming algorithm optimization for spoken word recognition, IEEE Trans. Acoustics, Speech, and Signal Proc., Vol. ASSP-26, 43-49.

Strik, H. and Boves, L. (1988) Averaging physiological signals with the use of a DTW algorithm. In Proceedings SPEECH’88, 7th FASE Symposium, Edinburgh, Book 3, 883-890.

Strik, H. and Boves, L. (1991) Control of fundamental frequency, intensity and voice quality in speech. This issue.

Figure captions

Figure 1. (a) A graphical representation of the DP algorithm, with (b) the five possible step sequences (A-E) in the symmetric DP algorithm when the slope constraint condition is 1/2. Indicated in Italics are the weighting coefficients w_k.

Figure 2. A graphical representation of non-linear time-alignment, when three line-up points are used. B indicates the beginning of the utterance, E the end, and L an acoustic event near the middle of the utterance.

Figure 3. An example of the function F_n for time stretching (step sequence A). In this example a straight line is used as the input signal S.

Figure 4. Oscillograms of the audio signals of three repetitions of the same utterance.The straight vertical line at 1.3 s connects the line-up points of the individual signals.

Figure 5. In the three upper panels the transglottal pressure signals are shown of the three utterances given in Figure 5. The lower panel contains the average transglottal pressure signal for 29 repetitions. The straight vertical line at 1.3 s connects the line-up points of the individual signals.

Figure 6. Average physiological signals for fundamental frequency, intensity level, transglottal pressure, oral pressure, subglottal pressure, lung volume, and electromyographic activity of the sternohyoid and vocalis muscles, obtained by the method of linear time-alignment. The straight vertical line at 1.3 s connects the line-up points of the individual signals.

Figure 7. The labels of the 29 utterances after linear time-alignment. The straight vertical line at 1.3 s connects the line-up points of the individual signals.

Figure 8. The labels of the 29 utterances after non-linear time-alignment. The straight vertical line at 1.3 s connects the line-up points of the individual signals.

Figure 9. Median physiological signals (solid lines) for fundamental frequency, intensity level, transglottal pressure, oral pressure,

subglottal pressure, lung volume, and electromyographic activity of the sternohyoid and vocalis muscles, obtained by the method of non-linear time-alignment and averaging. The dotted lines are a measure for the amount of variation (see text). The straight vertical line at 1.3 s connects the line-up points of the individual signals.