A Method for Segmentation of Voiced Speech Signals into Pitch Period Segments*

Vlasta Radová and Ondřej Šilhán

University of West Bohemia, Department of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic
radova@kky.zcu.cz
http://artin.zcu.cz/people/radova

Abstract. An algorithm for segmentation of voiced parts of speech waveforms into the segments, each of them corresponding to one pitch period, is described in the paper. The algorithm is based on the similarity of adjacent pitch period segments, and the dynamic time warping procedure is used for the similarity evaluation. Attention is also paid to the proper starting point selection, which will assure that the segmentation will be synchronized across all segmented waveforms.

1 Introduction

Several years ago we presented a speaker identification method that used parts of speech waveforms as the features [1]. The parts had the length of one pitch period and proved itself to be successful in the discrimination between speakers. The problem was, however, that the pitch period segments were selected manually from the speech waveform, which was a very time-consuming process.

In the TSD'98 workshop Vintsiuk [2] presented a method for pitch period discrimination that was based on the similarity of the adjacent segments. The segments were compared linearly in such a way that the longer segment was shortened to the length of the shorter one by cutting off the superfluous end part of the longer segment. Because we knew from our experiments with the speaker identification that the pitch period segments can be compared very well by the dynamic time warping procedure, we tried to develop a segmentation method based on the dynamic time warping. The principle of the method is explained in Sect. 2, and the problems that appeared during its development are described in Sect. 3.

2 Segmentation Algorithm

Let $s(k)$, $k = 0, \ldots, K$ be the samples of a speech waveform of a voiced sound, and let T_0 be the value of the pitch period (in samples) of that sound. The task

* The work was supported by the Grant Agency of the Czech Republic, project no. 102/02/0124, and by the Ministry of Education of the Czech Republic, project no. MSM 235200004

of the described algorithm is to segment the speech waveform into parts S_n, $n = 1, \ldots, N$, (from now on we will call these parts "pitch period segments" or only "segments"), each of them corresponding to one pitch period. Further, let each segment S_n, $n = 1, \ldots, N$, be determined by two values: B_n, which is the beginning of the n-th segment, and L_n, which is the length of the n-th segment. If we accept that the beginning of the $(n + 1)$-th segment can be determined according to the formula

$$B_{n+1} = B_n + L_n \ , \qquad n = 1, \ldots, N - 1 \ , \tag{1}$$

the task of the segmentation algorithm is

1. to determine the beginning of the first segment, i.e. to determine B_1, and
2. to find the length of each segment S_n, $n = 1, \ldots, N$, i.e. to find L_n, $n = 1, \ldots, N$.

2.1 Determination of the Beginning of the First Segment

In some cases, when we would need to segment a waveform into pitch period segments irrespective of where the segmentation starts, the beginning of the first segment B_1 might be determined randomly. However, as it was already mentioned in Sect. 1, our intention is to use this algorithm in a speaker identification/verification task. The waveforms obtained from voiced sounds of both the reference speakers and the unknown speaker will be segmented into pitch period segments, the segments of the unknown speaker will then be compared with the segments of the reference speakers, and according to the result of the comparison a decision about the identity of the unknown speaker will be made [1]. In such a case it is reasonable to synchronize the starting point of the segmentation across all waveforms in some way.

The first idea of how to reach the synchronization across all waveforms was to start segmentation at the greatest local minimum of the waveform in the interval $\langle 0, T_0 \rangle$. However, during the first experiments it was found out that the beginning of the waveform can sometimes be slightly corrupted by the preceding sound, which might cause problems with the synchronization required above (for more detailed description of this phenomenon see Sect. 3). Therefore we decided to omit the initial part of the waveform of the length of T_0, and to determine the beginning of the first segment according to the formula

$$B_1 = \underset{k \in \langle T_0, 3T_0 \rangle}{\operatorname{argmin}} \ (s(k)) \ . \tag{2}$$

2.2 Determination of the Length of Segments

The algorithm for the determination of the length of the segments should satisfy the following requirements:

1. The length of each segment must not differ too much from the pitch period T_0.
2. The adjacent segments have to be similar.
3. Each segment should start in a local minimum of the waveform.

Whereas the first two requirements result from the characteristics of a voiced speech signal and are quite natural, the third requirement can be regarded as a consequence of the requirement 2 and the choice of the first segment beginning. It means that the requirement 3 strongly depends on the choice made in Sect. 2.1, and if that choice changes, the requirement 3 has to change as well.

Let us suppose now that the segment S_n is given, i.e. its beginning B_n and length L_n are already known. In the conformity with the requirement 1 the length L_{n+1} of the segment S_{n+1} can be expressed as

$$L_{n+1} = T_0 + c^*_{n+1} + l^*_{n+1} \ , \qquad n = 2, \ldots, N-1 \ , \tag{3}$$

where c^*_{n+1} and l^*_{n+1} are the best corrections of T_0 with respect to the requirements 2 and 3, respectively. From now on the c will be called the "similarity correction", and the l will be called the "tuning correction".

In order to determine the best similarity correction c^*_{n+1} let us define the $D(S_n, S_{n+1})$ as a distance between the segments S_n and S_{n+1}. Because the segments can differ in their length, it is a good idea to align the endpoints of the segments before the distance is computed. It was shown in [1] that the nonlinear alignment of pitch period segments using the dynamic time warping (DTW) procedure gives better results than the linear alignment. For that reason we decided to align the segments using the DTW procedure. The distance $D(S_n, S_{n+1})$ is then determined as a by-product of the DTW. The best similarity correction c^*_{n+1} can now be determined as such a correction c_{n+1} of the period T_0, for which the distance $D(S_n, S_{n+1})$ is minimal. We have found out experimentally that the best correction c^*_{n+1} should have a value from the interval $\langle -\mathrm{round}(F_\mathrm{S}/1600) \, , +\mathrm{round}(F_\mathrm{S}/1600) \rangle$, where F_S is the sampling frequency of the speech signal, and the function $\mathrm{round}(x)$ returns the value of x rounded to the nearest whole number. Thus, c^*_{n+1} can be obtained from the formula

$$c^*_{n+1} = \underset{c_{n+1} \in \langle -b_c, +b_c \rangle}{\operatorname{argmin}} D(S_n, S_{n+1}) \ , \tag{4}$$

where $b_c = \mathrm{round}(F_\mathrm{S}/1600)$, $D(S_n, S_{n+1})$ is the distance between the segments S_n and S_{n+1} computed as the by-product of the DTW procedure, the segment S_n starts in the point B_n and has the length L_n, and the segment S_{n+1} starts in the point $B_{n+1} = B_n + L_n$ and has the length $T_0 + c_{n+1}$. The type of allowed transitions of the DTW function employed in the alignment process is depicted in Fig. 1.

The best tuning correction l^*_{n+1} is computed when the similarity correction c^*_{n+1} is already known. As it follows from (3), the correction l^*_{n+1} should assure that the next segment (i.e. the segment S_{n+2}) will start in a local minimum of the speech waveform. Thus, in order to determine the best correction l^*_{n+1} we will

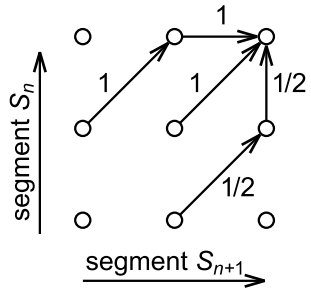

Fig. 1. Employed type of the allowed transitions of the DTW function

look for a local minimum in a vicinity of the speech sample $s\big(B_{n+1} + T_o + c^*_{n+1}\big)$. We have found out experimentally that the best tuning correction l^*_{n+1} should have again a value from the interval $\langle -\text{round}(F_\text{S}/1600)\,, +\text{round}(F_\text{S}/1600)\rangle$, where F_S is the sampling frequency of the speech signal, and the function $\text{round}(x)$ returns the value of x rounded to the nearest whole number. The value of l^*_{n+1} can be then determined according to the formula

$$l^*_{n+1} = \underset{l_{n+1}\in\langle -b_l, +b_l\rangle}{\text{argmin}} \big(s\big(B_{n+1} + T_0 + c^*_{n+1} + l_{n+1}\big)\big) \quad , \tag{5}$$

where $b_l = \text{round}(F_\text{S}/1600)$ and $s(k)$ is the k-th sample of the speech waveform.

The formulae (3), (4) and (5) allow to compute the length of a segment recursively from the previous segment. It means, that the length of the first segment has to be determined in a different way. In the algorithm described here we computed the length L_1 of the first segment together with the length L_2 of the second segment (for that reason the n in the formula (3) starts from 2). The lengths L_1 and L_2 were then determined according to the formulae

$$\begin{aligned} L_1 &= T_0 + c^*_1 + l^*_1 \quad , \\ L_2 &= T_0 + c^*_2 + l^*_2 \quad , \end{aligned} \tag{6}$$

where

$$\{c^*_1, c^*_2\} = \underset{\substack{c_1\in\langle -b_c, +b_c\rangle \\ c_2\in\langle -b_c, +b_c\rangle}}{\text{argmin}} D(S_1, S_2) \quad , \tag{7}$$

$b_c = \text{round}(F_\text{S}/1600)$, and $D(S_1, S_2)$ is the distance between the segments S_1 and S_2 determined again as the by-product of the DTW procedure, the segment S_1 starts in the point B_1 given by the formula (2) and has the length $T_0 + c_1$, and the segment S_2 starts in the point B_2 determined by the formula

$$B_2 = B_1 + T_0 + c_1 \tag{8}$$

and has the length $T_0 + c_2$.

When the values of the c_1^* and c_2^* are known, the best tuning corrections l_1^* and l_2^* are computed according to the formulae

$$\begin{aligned} l_1^* &= \underset{l_1 \in \langle -b_l, +b_l \rangle}{\operatorname{argmin}} \; (s(B_1 + T_0 + c_1^* + l_1)) \quad , \\ l_2^* &= \underset{l_2 \in \langle -b_l, +b_l \rangle}{\operatorname{argmin}} \; (s(B_2 + T_0 + c_2^* + l_2)) \quad , \end{aligned} \tag{9}$$

where $b_l = \mathrm{round}(F_\mathrm{S}/1600)$, $s(k)$ is the k-th sample of the speech waveform, B_1 is given by the formula (2), and B_2 is given by the formula

$$B_2 = B_1 + T_0 + c_1^* + l_1^* \quad . \tag{10}$$

3 Experimental Results

The algorithm described in the previous section was used for the segmentation of about 2,000 waveforms. Each waveform corresponded to a Czech vowel extracted from one of several words that were pronounced by 100 speakers.

The segmentation algorithm without the tuning corrections l_{n+1}^*, and l_1^* and l_2^* in (3) and (6), respectively, was used in the first experiments. The following problem was identified during the inspection of several randomly selected waveforms that were segmented with the algorithm: The initial part of the waveform was segmented correctly, however as the time went on the boundaries of the segments moved and the beginnings of the segments at the end of the waveform were not synchronized with the beginnings of the segments in the initial part at all (see Fig. 2a). During a detailed examination of the experimental results we have found out that the values of the $D(S_n, S_{n+1})$ in (4) differ very little or do not differ at all for different similarity corrections $c_{n+1} \in \langle -b_c, +b_c \rangle$, and therefore it is very difficult to find the exact minimum. The error caused by this phenomenon is very small in the initial part of the waveform, so that it seems that this part of the waveform is segmented correctly. However, as the algorithm goes on the error grows, and the segmentation becomes wrong. On the contrary, the values of $D(S_n, S_{n+1})$ for the corrections $c_{n+1} \in \langle -b_c, +b_c \rangle$ are very small in the comparison with the values of $D(S_n, S_{n+1})$ for $c_{n+1} \notin \langle -b_c, +b_c \rangle$. It means that the similarity corrections c_{n+1}^*, $n = 0, \ldots, N-1$, are useful, however an additional correction is necessary in order to "tune" the end of the segment into a correct minimum. Therefore we introduced the tuning corrections l_{n+1}^*, $n = 0, \ldots, N-1$, as described in Sect. 2.2, and after that the segmentation became much better (see Fig. 2b).

Since it is not possible to check out the segmentation of all 2,000 waveforms, only several waveforms were selected randomly, and those ones were inspected visually. Because such a random inspection did not show any longer serious problem in the segmentation, we decided to use the segmented waveforms in a speaker identification experiment. The reason for such a decision was as follows:

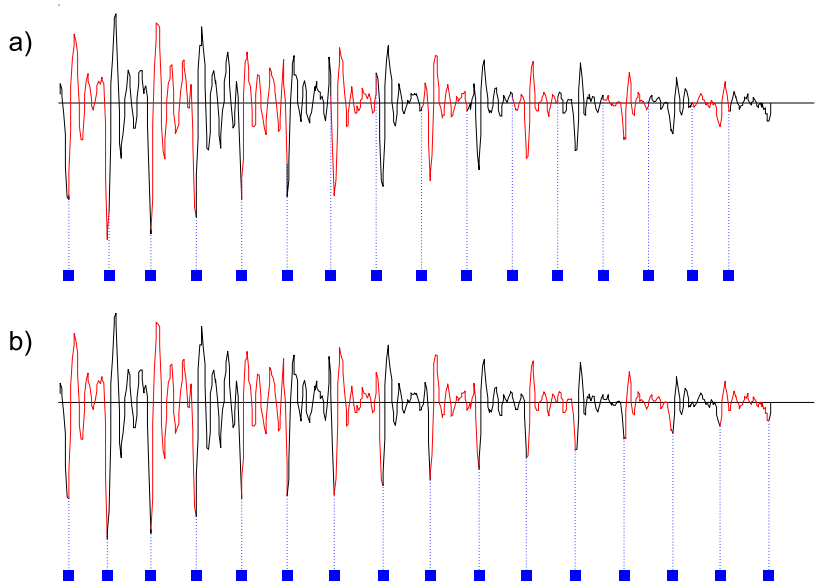

Fig. 2. The results of the segmentation when a) the tuning corrections were not used, and b) the tuning corrections were used.

In [1] we presented a speaker identification method based on the pitch period segments. The method showed itself to be very successful. So if the same experiment is performed now and some problems occur, they will be probably caused by a wrong segmentation.

One more problem appeared during the speaker identification experiment. When the beginning of the first segment B_1 was placed into the greatest local minimum from the interval $\langle 0, T_0 \rangle$ (see Sect. 2.1), a plenty of speakers were identified incorrectly. The reason of it can be seen from Fig. 3. In Fig. 3a and Fig. 3b there are two waveforms from one speaker. Each of the waveforms is segmented quite well, however the segments in Fig. 3a are not synchronized with the segments in Fig. 3b. As a result of it, the pitch period segments which do not correspond to each other were compared during the identification experiment, and it produced a great difference between the segments even if they belonged to the same speaker. When we searched for a cause of the wrong segmentation we noticed that plenty of waveforms that produced problems have an abnormality in its beginning. Such an abnormality can be seen in the waveform in Fig. 3b as well: The first pitch period segment (delimited by two small arrows) is shorter than the others, and therefore the segmentation algorithm found a false end of it. The cause of such an abnormality is probably the co-articulation, which means that the beginning of the waveform is influenced by the preceding sound. In order to prevent such problems we decided to move the beginning of the segmentation from the local minimum in the interval $\langle 0, T_0 \rangle$ into the greatest local minimum

in the interval $\langle T_0, 3T_0 \rangle$ (see the formula (2)). The result of the segmentation was then already satisfactory, as can be seen from the Fig. 3c.

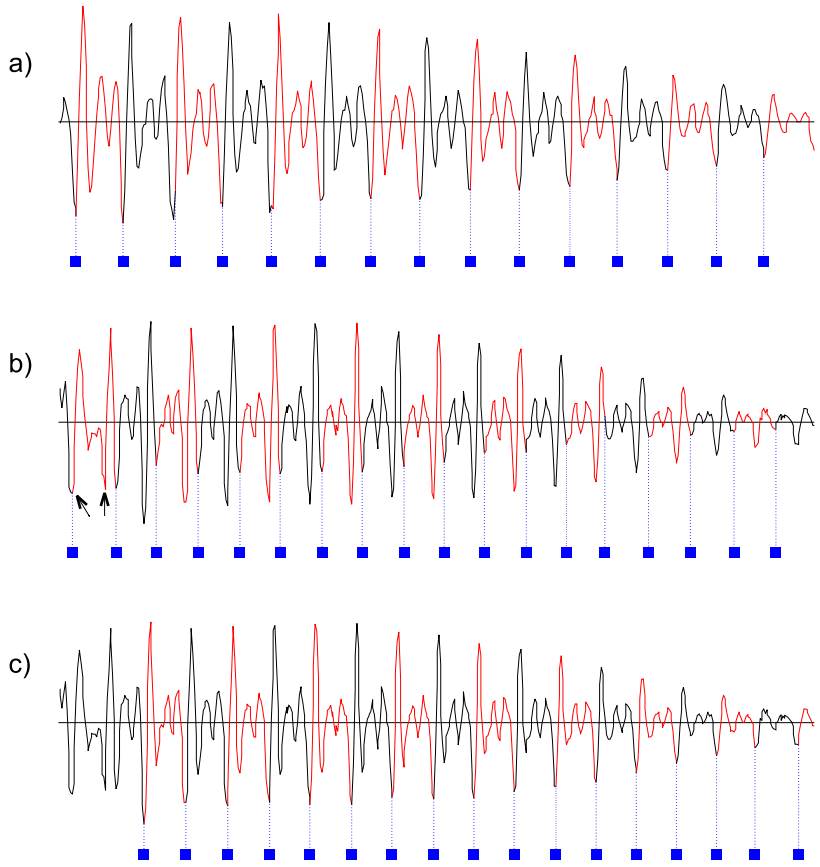

Fig. 3. The results of the segmentation for the vowel "a" of the speaker AFI. a) and b) The segmentation started in the greatest local minimum in the interval $\langle 0, T_0 \rangle$. The small arrows in b) delimit the part that should have been marked as the first pitch period segment. c) The same waveform as in b), however, the segmentation started in the greatest local minimum in the interval $\langle T_0, 3T_0 \rangle$.

4 Conclusion

The procedure described in this paper allows to segment voiced parts of a speech waveform into pitch period segments, i.e into segments, each of them corresponds to one pitch period. In spite of the fact that the procedure was developed with

the intention to use it in our speaker recognition method, it will certainly be useful everywhere where the pitch synchronous speech analysis is necessary.

References

1. Radová, V., Psutka, J.: An Approach to Speaker Identification Using Multiple Classifiers. In: Proc. of the International Conference on Acoustics, Speech, and Signal Processing ICASSP'97, Munich, Germany (1997) 1135–1138.
2. Vintsjuk, T. K.: Optimal Joint Procedure for Current Pitch Period Discrimination and Speech Signal Partition into Quasi-Periodic and Non-Periodic Segments. In: Sojka, P., Matoušek, V., Pala, K., Kopeček, I. (eds.): Text, Speech, Dialogue. Proc. of the First Workshop on Text, Speech, Dialogue. Brno, Czech Republic (1998) 135–140.