A time synchronization system is a helpful tool for different applications, such as language education and speech therapy. We present a system that performs temporal alignment of two utterances of the same phrase. The system consists of two parts. In the first part the time warping function is determined with Dynamic Time Warping (DTW). In the second part the time scale of one utterance is modified according to the time warping function. To obtain good performance, the dynamic time warping algorithm required significant modifications. Our listening test confirms that our time synchronization system has high precision and the resulting speech utterances are of natural quality.

Keywords: Time Synchronization, Time Scale Modification, DTW, WSOLA

I. INTRODUCTION

A time synchronization system is a helpful tool for different applications, such as language education and speech therapy. We present a system that performs temporal alignment of two utterances of the same phrase. The system consists of two parts. In the first part the time warping function is determined with Dynamic Time Warping (DTW). In the second part the time scale of one utterance is modified according to the time warping function. To obtain good performance, the dynamic time warping algorithm required significant modifications. Our listening test confirms that our time synchronization system has high precision and the resulting speech utterances are of natural quality.

II. METHODOLOGY

The first part is realized with Dynamic Time Warping (DTW) [1]. DTW is mainly known from speech recognition, where it was featured in most systems in the 80’s. Later on it was largely replaced by Hidden Markov Models (HMM’s), as they proved to be advantageous for several reasons [2]. Although displaced from speech recognition, DTW has been in use in the 90’s and later on for different applications such as speaker identification systems [3], signature verification systems [4], and in recent work for gesture recognition [5].

For the second part of the system Waveform Similarity Overlap and Add (WSOLA) synthesis [6] was selected. It falls in the class of time domain based Overlap and Add (OLA) methods. The idea behind the OLA synthesis methods is to create synthesized speech by concatenating small segments of speech. In doing so, the periodic structure of the speech signal has to be preserved. The different OLA methods such as the Synchronous Overlap and Add method (PSOLA) [7] or Pitch Synchronous Overlap and Add method (PSOLA) [8] offer various related solutions to this problem.

In [9] Verhelst presents an earlier system for time synchronization based on DTW and WSOLA. In his system DTW is applied without constraints or modifications. This leads to problems with the sound quality of the modified utterance if the reference and source signal are not sufficiently similar [9]. With our approach we account for acoustic and phonetic differences by introducing an accumulative local penalty constraint and a smoothing stage to the Dynamic Time Warping (sections 3.1 and 3.2).

2.1. DTW algorithm

Dynamic Time Warping is a pattern matching algorithm with a non-linear time normalization effect. It is based on Bellman’s prin-
The dynamic time warping algorithm [1] creates an alignment between two sequences of feature vectors, \((t_1, t_2, \ldots, t_N)\) and \((s_1, s_2, \ldots, s_M)\). A distance \(d(i, j)\) can be evaluated between any two feature vectors \(t_i\) and \(s_j\). This distance is referred to as the local distance. In DTW the global distance \(D(i, j)\) of any two feature vectors \(t_i\) and \(s_j\) is computed recursively by adding its local distance \(d(i, j)\) to the evaluated global distance for the best predecessor. The best predecessor is the one that gives the minimum global distance \(D(i, j)\) at row \(i\) and column \(j\):

\[
D(i, j) = \min_{m \leq i, k \leq j} [D(m, k)] + d(i, j). \quad (1)
\]

The computational complexity can be reduced by imposing constraints that prevent the selection of sequences that can not be optimal [1]. Global constraints affect the maximal overall stretching or compression. Local constraints affect the set of predecessors from which the best predecessor is chosen.

### 2.2. WSOLA algorithm

Waveform similarity overlap and add (WSOLA) is a time domain based algorithm for time scale modifications of speech [6] [11]. It gives high quality speech and allows scaling factors that may be specified in a time-varying fashion. One major advantage of the WSOLA method is that, in contrast to PSOLA [8], no pitch estimation is needed.

InOLA [12] (overlap and add) synthesis the modified signal is obtained by excising segments from the input signal, repositioning them along the time axis and performing a weighted overlap addition to construct the synthesized signal.

The basic idea of the WSOLA algorithm can be best explained graphically (see Fig. 2). The time warping function \(\tau^{-1}(L_k)\) assigns one segment of the source signal to each synthesis instant \(L_k\) in the target signal. A timing offset \(\Delta_t\) within a range of \(\pm \Delta_{max}\) around the time warping function \(\tau^{-1}(L_k)\) is needed to avoid pitch period discontinuities and phase jumps. In this way a proper segment synchronization in the synthesized signal is achieved. The timing offset \(\Delta_t\) is determined such that the synthesized segment maintains maximal local similarity to the natural continuity existing in the original signal. Assume segment (1) in Fig. 2 was the last segment excised from the source signal and added to the target signal. Next, WSOLA tries to find a segment (2) lying in the region \([\tau^{-1}(L_k) - \Delta_{max}, \tau^{-1}(L_k) + \Delta_{max}]\) (shaded region), that is maximally similar to the natural continuation (segment (N1)).

### III. TIME ALIGNMENT

Dynamic Time Warping (DTW) is used to establish a time scale alignment between source and reference signal. It results in a time warping vector \(\Theta\), describing the time alignment of segments of the two signals. \(\Theta\) assigns a certain segment of the source signal to each of a set of regularly spaced synthesis instants in the target signal.

A preprocessing step is taken to remove silence in the beginning and the end of each utterance. This is done by applying a threshold on the energy of the signal evaluated in blocks of length 125 ms and overlap of 15 ms. The feature extraction is performed on the remaining signal. Attributes of speech relevant for differentiating phonemes are measured over short time intervals, within which speech is considered to be quasi-stationary. The feature vectors are extracted from windowed segments of the signal of length 20 ms with 50% overlap. The chosen features are 12 MelCepstrum coefficients [2] and the log energy.

The Euclidean distance \(D(t, s)\) is applied to determine the distance between the feature vectors of the two sequences. As a global constraint, the search space of the DTW is limited to fall in a band of width \(G\). This is illustrated in Fig. 3 a). \(G\) is determined by

\[
G = 20 \cdot \lceil \log_2 \frac{M}{N} \rceil + 40, \quad (2)
\]

where \(N\) is the number of feature vectors for the source signal and \(M\) for the reference signal. Thus, the bandwidth is dependent on \(M/N\), the ratio of reference signal and source signal length. By using the base-2 logarithm an equal sized bandwidth is achieved for a time stretching by factor 2, as for time compression by factor 1/2. Fig. 3 b) shows the global constraint width \(G\) dependent on \(M/N\). For the local distance, a modified version of the Sakoe-Chiba [13] local constraint has been used, as described in more detail in the next section.

### 3.1. Accumulated local penalty constraint

By choosing an appropriate local constraint, the first derivative of the time warping vector \(\Theta\) may be limited in range. In [1] various local constraints are presented. One of the main criteria in the choice of the local constraint for our system is to preserve a certain flexibility in the alignment, needed to cope with local differences in the speaking rate within a wide range. For this reason we apply a symmetric Sakoe-Chiba local constraint without slope constraint.
Three possible predecessors are considered as candidates for the calculation of the global minimum distance:

\[ D(i, j) = \min[D(i-1, j-1), D(i-1, j), D(i, j-1)] + d(i, j) \]

Despite of the desired flexibility, we need to control the warping curvature. Applying a local constraint without a slope constraint tends to result in a path with long horizontal and vertical subpaths. Horizontal subpaths result in a problem in the signal modification part, as they lead to the repetition of one segment several times. This yields a synthetic sound. Vertical subpaths denote the skipping of several segments of the original signal, resulting in an unnatural sound because of syllables or even whole words being omitted. In addition, vertical subpaths can result in pitch discontinuities in the target signal. Changes in pitch mostly occur during vowels in natural speech. In the case of modification from a long to a short vowel, the alignment may contain a vertical subpath, such that the modified vowel is composed by the onset and offset of the long vowel, omitting the middle part. If the long vowel contains a modulation in pitch this leads to a clearly audible overlap.

Vertical and horizontal subpaths are necessary in the alignment of two utterances containing pauses of different length. By a vertical subpath a pause can be cut off, by a horizontal subpath a pause can be extended to a longer pause. Hence it seems reasonable to distinguish between segments containing speech and segments containing silence. A classification for each segment can be done by comparing its signal energy to a threshold obtained by statistics drawn from all segments. Sorting all occurring segment log energy values into 10 equally spaced groups, the threshold is chosen sufficiently large that the position of the natural continuation segment falls in the tolerance region, it leads to a repetition of the same segment even if the time warping function has been modified as described in section 3.2. With a slope limit of 0.5, horizontal subpaths are not suitable for time scaling with WSOLA. To ensure that the time warping vector \( \Theta \) does not have a local slope smaller than a certain value \( \sigma \) for segments classified as speech, we use a smoothing stage. \( \sigma \) is chosen to be 0.5, corresponding to time stretching to the double length. The time warping vector can easily be modified in a left-to-right fashion. Horizontal subpaths are replaced by a subpath with slope \( \sigma \) and extended forwards and backwards in time with a minimum slope of \( \sigma \). The new subpath meets the original subpath. This is illustrated in Fig. 5.

For segments classified as silence, horizontal subpaths remain to allow an extension of speaking pauses in the target signal. Repetition of the same segment inevitably results in synthetic sound, even if the segment contains no speech but background or breathing noise. To alleviate this we attenuate the repeated parts smoothed with a Hann window as illustrated in Fig. 6.

The output from DTW is a time warping vector containing subpaths with slopes 0, 1 and infinity, corresponding to horizontal, diagonal and vertical moves. Longer horizontal subpaths are not suitable for time scaling with WSOLA. To ensure that the time warping vector \( \Theta \) does not have a local slope smaller than a certain value \( \sigma \) for segments classified as speech, we use a smoothing stage. \( \sigma \) is chosen to be 0.5, corresponding to time stretching to the double length. The time warping vector can easily be modified in a left-to-right fashion. Horizontal subpaths are replaced by a subpath with slope \( \sigma \) and extended forwards and backwards in time with a minimum slope of \( \sigma \). The new subpath meets the original subpath. This is illustrated in Fig. 5.

The time scale of the target signal is changed using the WSOLA algorithm. The time warping vector \( \Theta \) determines a position of the segment to be excised from the source signal for each synthesis instant \( L_k \) as described in section 2.2.

The segment length is 20 ms, as used in the feature extraction and signal alignment procedure. Hann windowing with a 50% overlap is applied. In WSOLA the segment for synthesis is picked from the tolerance region \( \left[ \tau^{-1}(L_k) - \Delta_{\max}, \tau^{-1}(L_k) + \Delta_{\max} \right] \) around the ‘true’ time instant \( \tau^{-1}(L_k) \) as described in section 2.2. If \( \Delta_{\max} \) is chosen sufficiently large that the position of the natural continuation segment falls in the tolerance region, it leads to a repetition of the same segment even if the time warping function has been modified as described in section 3.2. With a slope limit of 0.5, \( \Delta_{\max} \) needs to be smaller than a quarter of the segment length. Hence the tolerance \( \Delta_{\max} \) is chosen to be 4.9 ms. For good performance, \( \Delta_{\max} \) must be selected to be larger then half a pitch period. Thus, our system functions well for a pitch down to 100 Hz. The cross-AMDF coefficient is used as measure of similarity [6].

Time stretching for unvoiced segments often leads to an audible periodicity using the basic WSOLA method. To reduce these artefacts, segments are classified as voiced or unvoiced, and every third consecutive unvoiced segment gets reversed in time. A similar procedure is known from PSOLA, where every other unvoiced segment is reversed [8]. Our classification is done by means of counting the zero crossings, where a high number of zero crossings indicates unvoiced sounds. With the described method a more natural sound is achieved.

We carried out a listening test on 17 listeners to evaluate the time synchronization system. We selected 10 different utterances from the TIMIT database, five spoken by male, five by female speakers.
To get a measure for synchronicity we recorded test utterances spoken by two male and two female speakers. The speakers were first asked to read a sentence independently, then in synchronization with a TIMIT sentence. They could practise as often as required to achieve satisfactory synchronization. The independently spoken utterances were processed by our time synchronization system. For the listening test, we selected 10 recorded utterances, that account for a wide range of time scaling factors. Four of these sentences were read by male speakers, six by female speakers.

The first part of the listening test was an A-B preference test, presenting the sentences synchronized by the speaker and by our system to the listeners. The 17 listeners were asked to judge the accuracy in synchronization. Fig. 7 a) shows the result of the preference test for the different test utterances. The height of the bars indicate the preference of the listeners. It can be seen that in most of the cases our time synchronization system is clearly preferred over to natural synchronization.

In the comparison one has to consider that the prosody of the independently spoken sentence might differ from the TIMIT reference sentence, whereas the speaker automatically will adjust the prosody speaking simultaneously with the reference. Therefore, the natural synchronization will sometimes be felt as more synchronous, even if it is not. This is the case for the second sentence from the test, which contains a large difference between the independent and reference prosody.

The aim of the second part of the listening test was to evaluate the quality of the processed files. To get a more balanced ratio of increased and decreased speaking rate, we added four additional test sentences where the TIMIT utterances were processed to make them synchronous with our recorded utterances. Thus, we obtained additional examples where the speaking rate is decreased, since the recorded test sentences are on average slower than the TIMIT utterances. The target signal and source signal were presented, and the same 17 listeners asked for a comparative qualitative rating between -5 (much worse) to +5 (much better). In Fig. 7 b) the mean rating and standard deviation for the test sentences over all listeners are depicted. The test results of the second part were inconsistent, showing that the judgement of speech quality for one sentence differs significantly for different persons. A reason for that might be that the judgment are influenced by the prosody, which is automatically changed by changing the time scale. Nevertheless, it can be concluded that the time modified sentences are experienced as being of good quality on average, in many cases rated better than the original.

VI. CONCLUSION

We presented a system that performs time synchronization between two different utterances of the same sentence based on DTW and WSOLA. In contrast to an earlier system (presented in [9]), our system can align utterances that differ severely (caused by a different speaker and speaking style), and makes the resulting time scaled utterances sound natural.

To obtain good time synchronization, major modifications are necessary to make the algorithms suitable for our application. We introduced an accumulated local penalty constraint in DTW to control the curvature of the time warping function. The constraint is made dependent on a classification of segments into speech or silence. A smoothing stage was added to handle the limitations of the WSOLA method in dealing with low slopes in the time warping function for speech segments. By that we achieved flexibility in the time warping function to cope with large local differences, as, for example, longer silence parts between words, while maintaining properties that guarantee good quality. Moreover, the speech quality was improved compared to the basic WSOLA algorithm by time reversing unvoiced segments.

REFERENCES