# ROBUST TEMPORAL ALIGNMENT OF SPONTANEOUS AND DUBBED SPEECH AND ITS APPLICATION FOR AUTOMATIC DIALOGUE REPLACEMENT

*Pieter Soens and Werner Verhelst*

Interdisciplinary Institute for Broadband Technology - IBBT, Vrije Universiteit Brussel
Dept. ETRO-DSSP, Pleinlaan 2, B-1050 Brussels, Belgium
phone: +(32)(0)2/629.29.30 - fax: +(32)(0)2/629.28.83
email: {psoens,wverhels}@etro.vub.ac.be
web: http://www.etro.vub.ac.be/research/DSSP

**ABSTRACT**

In this paper, we present a robust system for the temporal alignment of 2 renditions of the same speech utterance. The system operates in 2 steps: during analysis, the timing relationships between the speech segments of the utterance that serves as a timing reference and the corresponding speech segments in the replacement utterance are measured by means of a dedicated dynamic time warping algorithm. The obtained warping paths are then processed and used to synthesize a high-quality speech utterance that is time-aligned with the reference. Subjective audio-visual listening tests performed within the context of a difficult Automatic Dialogue Replacement task demonstrated that the proposed system achieves a significant improvement compared to the industry-standard benchmark, both in terms of achieved lip-synchronization accuracy as well as in overall sound quality of the synthesized utterances.

## 1. INTRODUCTION

A system for the temporal alignment of speech utterances modifies the timing structure of a first utterance (replacement, dub) in such a way as to synchronize it with a second utterance (reference, guide), which has the same textual content and has been produced by the same or by a different speaker. In general, such a system achieves the synchronization in 2 steps. First, the time correspondence is measured between the matching phonemes in both utterances. The resulting timing relationship describes the varying amounts of time stretching and compression necessary to bring the time axis of the replacement into optimal alignment with that of the reference. In a second step, the relative timing differences between the utterances are cancelled out by warping the time axis of the replacement in accordance with the measured timing relationship.

Although we can enumerate many possible uses for time alignment systems, our special attention in this paper goes to Automatic Dialogue Replacement (ADR), a well-known post-production technique in the audio-for-video industries. During the production of film soundtracks, dialogues are frequently re-recorded in a studio and used to replace the original ones recorded on the set. Very often, this is necessary because of the poor quality of the original recordings that might for example be corrupted by some kind of background noise that is difficult to control. As another example it is sometimes argued that an actor can produce a markedly improved spoken performance in a studio in comparison to the one produced on the set, which is usually very chaotic and makes it difficult to capture the true mood of a scene. In either case, straightforward replacement of the original recordings by the studio dialogues introduces a lot of mismatches between the lip and mouth movements in the picture and the actual timing and duration of the individual phonemes in the replacement speech. ADR is the most widespread technique used for the (indirect) compensation of such audio-visual "lip-synch" errors. It operates as follows: each actor involved in a particular scene attends a special dubbing session, during which the appropriate pictures are projected onto a screen in front of him, while replaying the original recordings over headphones. The actor then revoices the original dialogues, ensuring not only that his replacement speech precisely synchronizes with the on-screen lip movements, but also that the nuances of his performance match the original. Post-synchronizing dialogue is generally considered very difficult because most actors have a lot of difficulties to maintain synchrony while speaking. In addition, its repetitive nature makes it also very dull and time-consuming as the actor often needs to redeliver his lines until director and dialogue editor have compromised between the desired level of performance and timing. In the past, a few systems have been developed that allow automatically time-aligning the studio dialogues with the original recordings. These systems not only save time and money, they also release the actors from their technical preoccupation of speaking in synchrony with a picture soundtrack and thus allow them to fully concentrate on their primary task of acting and producing great performances.

This paper is organized as follows: section 2 reviews the previous work on automatic temporal alignment and the observed shortcomings in the approaches followed. In section 3 we motivate and discuss the proposed "split time warping" technique, which we evaluate in section 4 by comparing its performance against the industry-standard benchmark. Also are discussed the employed evaluation methodology and database. Finally, in section 5, we discuss the results and conclude the paper.

## 2. RELATED WORK

Over the last 4 decades, a considerable amount of research has been carried out on the development of techniques for the automatic time registration of corresponding events in 2 renditions of a same utterance (see for example [1] and related references therein). On the contrary and to the best of the authors' knowledge, very little efforts have been made when it comes down to applying the registered timing relationships for speech synthesis purposes.

The first attempts were made in the eighties, primarily in the early original work of Bloom [2, 3], who developed a digital audio signal processor named WordFit for the automatic post-synchronization of revoiced studio recordings with the corresponding recordings made on the film set. Although this system was designed to work with a variety of audio signals and not only with speech, it was reported that no single set of parameters could be found for which the system or its successor, VocALign PRO, would work under all circumstances [4, 5]. One very important and practical problem that arises in time-aligning sentence long speech utterances is the presence of long inter-word gaps, possibly between different words and/or of different durations, in one or both utterances. During the development of WordFit, Bloom implemented a modified version of the ZIP algorithm [6]. Although this algorithm could solve part of the problems as explained in [2], it is generally not capable of correctly inserting or rejecting pauses into or from the replacement track. Another disadvantage of ZIP is that it

This research was sponsored in part by the IWOIB (Instituut ter bevordering van het Wetenschappelijk Onderzoek en de Innovatie van Brussel) with a grant in the Spin-Off In Brussels program for the project EOS - studie ter Exploitatie van Onderzoeksresultaten op het vlak van Spraakmodificatie.

 ISSN 2076-1465 

belongs just like UELM [7] and MATCH [8] to a class of dynamic time warping (DTW) algorithms that estimate the globally optimum path by tracking a locally optimum path using a local search window. Although the window steering and partial trace back procedures ensure the amount of computation and storage can be kept to modest levels, such algorithms are susceptible to tracking failures (see for example [9]), especially when large timing differences occur in the sentence pairs.

Later on, Verhelst and Borger studied the alignment of speech utterances in the context of prosody transplantation [10]. Such systems can be used to interchange prosodic features, such as timing, pitch and timbre among different renditions of a same utterance. It was concluded that in order to make prosodic transplantations widely applicable, further work had to be done to improve their robustness: informal experiments revealed that, with utterance pairs that are not acoustically and phonetically sufficiently close to each other, local distortions could be regularly perceived, even when only timing is transplanted [11]. Very often, these distortions could be traced to some event in the timing relationship, but could not always be considered to be due to an error in this relationship, nor to the system that was used to perform the time scaling. In general, it was concluded that the perceived distortions could be attributed to 3 different types of acoustic-phonetic differences, which are described in detail in [12]. In [11], Verhelst built a basic system for the automatic post-synchronization of speech utterances based on standard DTW and WSOLA. The system proved to be quite robust to significant timing differences such as those that can for example be observed between speech utterances in which silent pauses occur between different words, but it was also noted that the time-scaled results very often suffered from many audible distortions. The major part of these distortions could be readily identified with the short abrupt transitions in the time warping path and it was shown that they could be straightforwardly smoothed out with the help of a graphical warping path editor that was developed for that purpose [12]. Although it was concluded that such an editor could form an effective tool for the semi-automatic correction of lip-synch errors, no objective criterion was formulated that enabled the consistent and automatic production of high-quality natural sounding results.

Finally, in [13], Resch and Kleijn adopt the approach of Verhelst, but they classify the reference and replacement tracks into speech and silence segments, the information of which is used to bias the warping path towards preferred directions in different situations. The major problem with this approach is that it applies the DP principle to a situation that does not justify its use [14]. Therefore, as is the case with ZIP, the alignment can get stuck in local minima, causing significant misalignments. Such misalignments were also verified from extensive experiments using our implementation of the algorithm, and in many cases the results were found to be inferior to those of VocALign PRO.

## 3. SPLIT DYNAMIC TIME WARPING

Although the several algorithms proposed in literature have been thoughtfully motivated, they all have to contend with specific drawbacks in different situations, which stem from a somewhat contradictory requirement that is imposed on the warping function curvature: at some parts this function should allow very steep or flat gradients to account for the possible different location and/or duration of pauses, while at other positions it should be smooth enough to avoid unnatural sounding artefacts in the time-aligned results. In order to meet this dual requirement in a convenient manner, the proposed system uses a DTW-based timing analysis approach, which splits up the calculation of the final warping path in 2 steps (the details of the WSOLA-based synthesis are identical to those described in [11] and will therefore not be discussed in this paper).

### 3.1 Identification of corresponding speech segments

The first step in the timing analysis is motivated by the conclusion that the main concern for the greater part of time alignment applications, and for ADR in particular, is to know the temporal variations

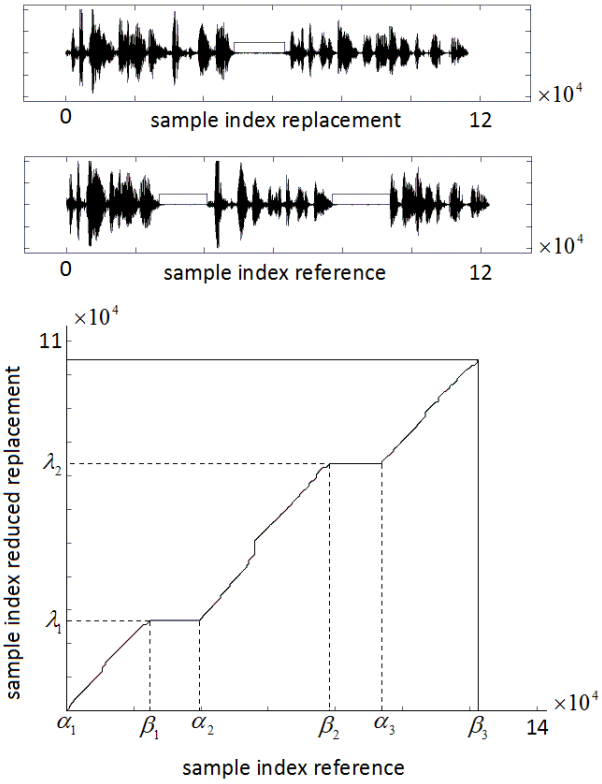


Figure 1: Illustration of split time warping ($f_s$ = 16kHz throughout the figures).

that occur between the corresponding speech segments[1] in the 2 waveforms. Furthermore, it is well-known that the details of a DTW path can be quite arbitrary during the alignment of non-speech segments and can therefore give rise to tracking errors [6]. Hence, it seems reasonable to first segment the 2 waveforms into intervals containing speech and intervals containing non-speech before applying a specific DTW algorithm. Assuming both waveforms have been precisely segmented (more details on the segmentation are given in section 4.3 and [15]), the main idea behind the proposed method is that for each of the $R$ reference speech segments delimited by time markers $(\alpha_r, \beta_r)$ with $1 \leq r \leq R$, there must correspond a replacement speech segment $(\lambda_{r-1}, \lambda_r)$. Since in general the number and/or location of the non-speech segments in the 2 waveforms is different, automatic identification of the matching speech segments is not straightforward. Experiments in [5] demonstrated that the corresponding pairs can be identified by splitting the replacement speech waveform in which all non-speech segments were removed in a preprocessing step ("reduced replacement") at time instants[2]

$$\lambda_r = \frac{\int_{\beta_r}^{\alpha_{r+1}} g(x)\tau(x)dx}{\int_{\beta_r}^{\alpha_{r+1}} g(x)dx} \qquad r = 1 \ldots R-1 \tag{1}$$

where $\lambda_0 = 0$ and $\lambda_R$ equals the duration of the reduced replacement. In this expression, $\tau(x)$ represents the linearly interpolated DTW path between the reference (along the x-axis) and the reduced replacement using the symmetric Sakoe-Chiba local constraint (with zero slope constraint condition) [16]. Furthermore, $g(x)$ is a Gaussian weighting function symmetrically positioned over $[\beta_r, \alpha_{r+1}]$ that is used to bias the split towards the speech segment boundaries.

[1] Without loss of generality, we defined a "speech segment" within the context of ADR as each sequence of phonemes that is not interrupted by a breathing pause, silence or background noise ("non-speech segment").

[2] Preferably, also the non-speech segments in the beginning and at the end of the reference are removed in a preprocessing step.

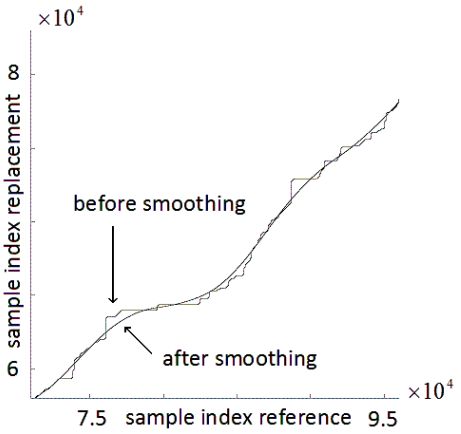


Figure 2: Smoothing process in detail (2$L$=500ms).

Figure 1 illustrates the split time warping concept on a replacement and reference waveform with 1 and 2 non-speech segments, respectively. Whereas the lengths of the horizontal stretches approximately cover the durations of the non-speech segments in the reference, their position along the y-axis indicates where to split the reduced replacement such that the corresponding speech segments can be identified.

### 3.2 Smoothing and postprocessing of the sub warping paths

In a second step, we recalculate the timing relationship for each pair of matching speech segments using the same DTW algorithm as in the previous step. Since the resulting "sub warping paths" $\tau_r(x)$ can be expected to be reasonably close to the diagonal linear path, recomputation was sped up using a global constraint in the form of a small Sakoe-Chiba band [16]. In addition, an Itakura parallelogram [17], defined by lines of slope $1/2$ and 2 was applied to ensure the correct alignment of the speech/non-speech segment boundaries.

***Smoothing:*** as mentioned in section 2, straightforward use of the sub warping paths for overlap-and-add time-scale modification of the replacement speech would inevitably lead to distorted results, mainly due to the short abrupt portions that correspond to the unrealistic time-scaling factors of 0 or ∞. As we suggested in [11], we smoothed the sub warping paths using different techniques such as piece-wise linear smoothing, DTW variants with especially crafted local constraints (as in [18]), and more generative non-linear smoothing techniques such as LOWESS [19]. From the techniques studied, LOWESS smoothing using a zero order degree polynomial proved both very effective as well as computationally efficient. In that case, the r-th smoothed sub warping path $\widetilde{\tau}_r(x)$ is obtained from the centrally weighted moving average represented by expression 2.

$$\widetilde{\tau}_r(x) = \frac{\int_{x-L}^{x+L} w(u-x)\tau_r(u)du}{\int_{-L}^{L} w(u)du} \qquad r = 1\ldots R \tag{2}$$

Figure 2 illustrates the smoothing process in detail: for the calculation of the smoothed sub warping paths, we followed the traditional LOWESS approach in using a tricube window

$$w(x) = \left[1 - \left(\frac{|x|}{L}\right)^3\right]^3 \qquad |x| \le L \tag{3}$$

in which the application-dependent window length 2$L$ largely defines the trade-off between achieved timing accuracy (or lip-synch accuracy in the case of ADR) and perceived voice quality.

***Postprocessing:*** although the smoothing process constrains the first and higher-order derivatives to more realistic values and leads to more smoothly sounding results, occasional peaks in these functions can still be responsible for unnatural sounding speech rates, accelerations and/or decelerations, and should therefore be further

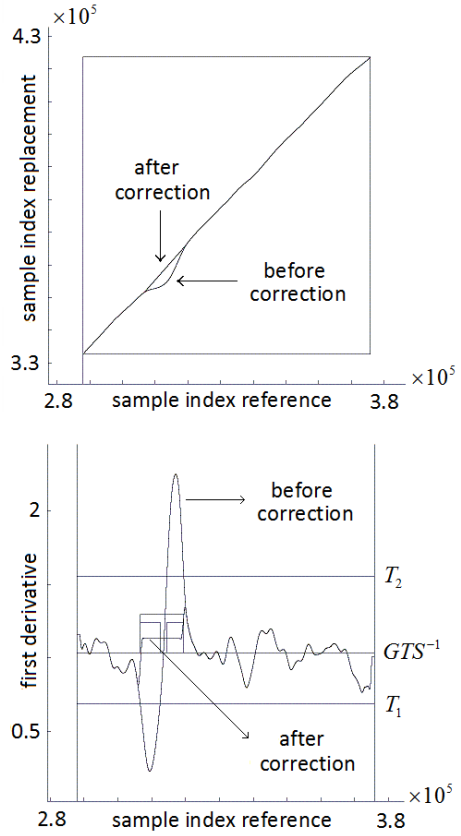


Figure 3: Illustration of the postprocessing stage ($\alpha$=1.5).

constrained. One possible way is to constrain the first derivative of the smoothed sub warping paths in the following manner:

$$T_{1,r} = \frac{GTS_r^{-1}}{\alpha_r} \le \frac{d\widetilde{\tau}_r(x)}{dx} \le \alpha_r \cdot GTS_r^{-1} = T_{2,r} \tag{4}$$

In this expression, $GTS_r$ represents the r-th global time scaling factor, which is defined as the ratio of durations of respectively the r-th reference and replacement speech segments that are being aligned, and $\alpha_r$ is an application-dependent constant in the range $1.1\ldots1.5$. Furthermore, threshold values $T_{1,r}$ and $T_{2,r}$ are defined as the lower and upper bound of inequality 4, respectively. From a physical point of view, expression 4 implies that the instantaneous speech rate of the replacement speech after time-scale modification ($SR_x(t)$) is constrained by that before time-scale modification ($SR_y(t)$) in accordance with

$$T_{1,r} \cdot SR_y(t) \le SR_x(t) \le T_{2,r} \cdot SR_y(t) \tag{5}$$

Figure 3 illustrates the correction procedure that was applied to achieve natural sounding results. The small bend in the smoothed (sub warping) path in the upper panel of figure 3 would generate an unnatural sounding speech deceleration followed by an unnatural sounding speech acceleration. This is correctly reflected in the lower panel by the sharp negative and positive peak in the function that represents the first derivative of the smoothed path. Applying a threshold yields a first estimate of the time intervals where this function should be limited in range (small square wave). Merging of the corresponding adjacent time intervals eventually identifies the portions in the smoothed path that require further processing (large square wave). Allowing each of these portions to be extended forwards and backwards in time, the applied correction procedure replaces the smoothed warping path by the shortest possible straight line, the slope of which satisfies expression (5).

## 4. EVALUATION OF SYSTEM PERFORMANCE

### 4.1 Evaluation methodology

The objective evaluation of the overall timing accuracy of a given time alignment system is a difficult task, which is mainly due to the inherent subjectiveness that exists in identifying the corresponding phoneme boundaries in 2 renditions of a same speech utterance [1]. In automatic phoneme alignment for example, evaluation is most often reported in terms of what percentage of a set of automatically generated phoneme boundaries are within a given time threshold of a known set of manually generated boundaries (there is a general consensus that the latter are the most accurate that can be obtained). At first glance, this strategy could be readily applied to our problem, for example by manually labeling the time-aligned results and the corresponding reference samples and relating the 2 series of time markers to estimate the overall timing error. However, this approach would only be acceptable if the difference between the 2 sets of time markers can be considered greater than the difference between the individual sets of manually generated time markers and the unknown correct time markers, a condition which is usually not met since our time-aligned results are generally well aligned with the reference samples. Furthermore, with regard to the objective evaluation of the overall speech quality of the time-aligned results, "full-reference" evaluation methods such as PESQ are not appropriate in this situation since they apply a temporal alignment procedure for comparison of the "signal under test" (time-scaled result) with the undistorted (dub) signal [20]. Because objective evaluation for time synchronization is difficult, we evaluated the proposed system within the context of ADR by means of a subjective audio-visual listening test. Although the proprietary nature of the industry-standard VocALign PRO (V4.0) hampers insight in the algorithmic details of the alignment process (the output waveforms are the only information available for evaluation), it was selected as a baseline for comparison, since it is world-wide considered the benchmark system for automatic time synchronization and ADR [21].

### 4.2 Database recordings

With a view to the experiment in section 4.3, we recorded an audio-visual corpus, comprising 80 different samples, produced by 2 male and 6 female native Dutch speakers. The data from this corpus was extracted from 2 sets of recording sessions. In a first series, we invited each time 2 speakers for a 30 minute table talk. From each of these conversations, 5 samples were extracted for each person. In doing so, care was taken the selected samples were sufficiently long such that they would cover a wide range of speaking rates as well as pauses of different kinds and durations. In a second series, the same speakers were asked to mimic the selected parts of their conversations by revoicing the literally transcribed lines from a large screen at a pace they felt comfortable with. In contrast to the traditional approach in ADR, we did not require the speakers to deliver performances with near-perfect lip-synch accuracy. As a consequence, we can generally observe substantial timing differences between the corresponding sample pairs, which therefore constitute a suitable test database to research the alignment capabilities of the proposed algorithm and in particular its robustness against the acoustic-phonetic differences described in [12]. For both the spontaneous and revoiced speech samples, table 1 shows the average overall duration (**OD**), average speech and non-speech rate (**SR** resp. **NSR**) and average duration of short (**DS**), medium (**DM**) and long (**DL**) non-speech segments. We remark that both classification as well as observed distribution of the short ($<$ 200ms), medium ($>$ 200ms,$<$ 1s) and long ($>$ 1s) non-speech segments were in agreement with [22].

### 4.3 Experiment

A complete set of 80 alignment runs was made by synchronizing all dubbed speech samples with the corresponding spontaneously spoken samples using both the proposed and baseline system (some examples can be downloaded from http://www.etro.vub.ac.be/research/DSSP/demo).

| | **spontaneous speech** | **revoiced speech** |
|---|---|---|
| **OD [s]** | $20.3 \pm 4.4$ | $25.0 \pm 6.2$ |
| **SR [syll/s]** | $5.41 \pm 0.90$ | $4.76 \pm 0.92$ |
| **NSR [1/s]** | $0.238 \pm 0.070$ | $0.374 \pm 0.083$ |
| **DS [ms]** | $155 \pm 26$ | $151 \pm 29$ |
| **DM [ms]** | $499 \pm 189$ | $494 \pm 180$ |
| **DL [ms]** | $1255 \pm 258$ | $1193 \pm 205$ |

Table 1: Major database statistics.

Due to its proprietary nature, users only have limited control in the way the time-aligned results are produced with VocALign PRO. First, one has to select a type of alignment mode ("basic" or "advanced"), each one of which has 5 possible different settings that control the internal parameters of the applied alignment algorithm. In contrast to the basic mode, which only allows time scaling ratios in the range $1/2 \ldots 2$, the advanced mode allows much larger amounts of time compression and expansion to occur. To make a proper choice among the 2 modes and their respective settings, the user can resort to a manual, which describes for each combination the amounts of time compression/expansion that can be provided, the nature of the input waveforms for which it is applicable and the expected output sound quality [21]. For the alignment of our recordings, we found that the "advanced" mode with the "high flexibility" setting produced far better results than the other combinations: it was therefore chosen as final setting for all alignments. In addition, the alignment process can *optionally* be further controlled by targeting the alignment at specified pairs of "synch points" in the 2 waveforms. However, it must be remarked that such points are always interpreted as *suggested* points, which the alignment algorithm will *try to match*, and which can therefore be ignored completely. Although we tried to improve the results for the "difficult" alignment pairs by manually identifying the corresponding speech/non-speech transitions in the 2 waveforms, in most of the cases these pairs of anchor points were ignored or gave rise to the error message "no warping path could be fit through one or more of the waypoints".

For the alignment of our recordings by means of the proposed system, we first segmented the samples into speech and non-speech intervals. In the first instance, this was accomplished automatically. However, because of the importance of an accurate speech/non-speech discrimination, we further manually inspected the speech/non-speech transitions and, where necessary, corrected them in a very efficient interactive way by means of a GUI specifically designed for ADR [15]: at each time this tool allows to zoom into and slide through the waveforms, select and audition specific portions, and make corrections by dragging the speech/non-speech boundaries to the left or the right. After segmentation, the data were processed according to the details described in section 3.

### 4.4 Subjective audio-visual listening test

The time-aligned samples obtained with the 2 synchronization systems were re-assembled with the corresponding video fragments from the conversation sessions and subsequently randomly arranged and presented in an equal amount of triplets (A,B,C) and (A,C,B), in which A denotes an original video sample, and where B and C represent this same fragment but with the audio replaced by the time-aligned result obtained with the baseline and proposed system, respectively. For each of the 40 triplets, we first asked 8 listeners to view fragment A and then rate both the perceived audio-visual lip-synch accuracy as well as the overall sound quality (naturalness & intelligibility) of B and C by assigning scores to their opinions, according to the ITU-R 5-point degradation scale [23]. These scores represent a number in the range 1 to 5, which provides a numerical indication of the quality of the considered audio(-visual) feature.

### 4.5 Results

Table 2 shows the arithmetic means evaluated from the opinion scores for each and across all speakers (**DMOS**) (scores were first averaged across all test listeners, and then over the different video samples), as well as the sample standard deviation (**s**), standard error of the mean (**SEM**) and the 95% confidence interval (**95%CI**).

| | lip-synch accuracy | | speech quality | |
|---|---|---|---|---|
| | **Baseline** | **Proposed** | **Baseline** | **Proposed** |
| **S1** | 2.45 | 4.10 | 1.86 | 3.35 |
| **S2** | 2.47 | 3.84 | 1.99 | 3.16 |
| **S3** | 3.14 | 3.61 | 2.54 | 3.66 |
| **S4** | 2.82 | 3.99 | 2.43 | 3.71 |
| **S5** | 3.05 | 3.77 | 2.52 | 3.49 |
| **S6** | 2.29 | 3.60 | 2.12 | 3.29 |
| **S7** | 2.19 | 4.10 | 2.35 | 3.88 |
| **S8** | 3.04 | 4.05 | 2.86 | 3.79 |
| **DMOS** | **2.68** | **3.88** | **2.33** | **3.54** |
| **s** | **0.85** | **0.38** | **0.51** | **0.39** |
| **SEM** | **0.13** | **0.06** | **0.08** | **0.06** |
| **95%CI** | **2.41...2.95** | **3.76...4.00** | **2.17...2.49** | **3.41...3.66** |
| **Δ** | **1.2** | | **1.21** | |
| **d** | **1.81** | | **2.64** | |
| **RI(%)** | **44.8** | | **51.9** | |

Table 2: Major statistical analysis results for the listening test.

In addition are given the raw (**Δ**) and standardized differences (unbiased Cohen's **d** effect size) in overall mean DMOS scores, and also the relative improvement of the proposed system over the baseline system (**RI**). Since the distribution of the assigned scores was far from Gaussian, we used the Wilcoxon matched-pairs signed-ranks test with a threshold significance level $\alpha = 0.05$ to prove the statistical significance of the observed differences between the mean DMOS scores of the baseline and proposed system for both features studied. We remark that the computed p-values in both paired tests were smaller than 0.0001: this is mainly explained from the observation that the listeners preferred the baseline system over the proposed system in only 6.25% (lip-synch accuracy) and 4.37% (speech quality) of the cases. For the sake of completeness, we report that no difference could be perceived in 18.12% and 11.56% of the cases, respectively.

## 5. CONCLUSION

From the results, we can conclude that, both for the samples that were processed with the proposed as with the baseline system, audio-visual lip-synch errors could still be observed at some points. However, while the latter were on average perceived as in-between "disturbing" and just "slightly disturbing", the former were perceived as "not disturbing or annoying". With regard to the quality of processed speech samples, similar conclusions can be drawn, although the overall DMOS scores are somewhat smaller. Furthermore, we can see that the non-zero differences in DMOS scores as well as their variabilities are quite pronounced: this is chiefly explained from the difficulties that were experienced in aligning the database samples by means of the baseline system. For the major part of the sample pairs, the timing structure discrepancies are quite large due to their relative differences in speech rates and number, duration and nature of pauses used. It was observed that for such pairs, the baseline system regularly produced unacceptable results, which could not be further corrected, neither by selecting a different alignment algorithm and/or setting, nor by manually placing corresponding "synch points" in the 2 waveforms.

In summary, we can conclude that the proposed system has demonstrated an overall relative improvement in DMOS score of 44.8% (lip-synch accuracy) and 51.9% (speech quality) over the baseline system when it is used for the temporal alignment of utterances in which large structural timing differences occur, such as those between spontaneous and dubbed speech.

## REFERENCES


[1] J.P. Hosom, *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information*, Ph.D. thesis, Oregon Graduate Institute of Science and Technology, May 2000.

[2] J.P. Bloom, "Use of dynamic programming for automatic synchronization of two similar speech signals," in *Proc. of ICASSP'84*, San Diego, USA, March 19-21 1984, vol. 9, pp. 69–72.

[3] J.P. Bloom and G.D. Marshall, "Method and apparatus for use in processing signals," US Patent #4,591,928, May 27 1986.

[4] J.P. Bloom, G.W. McNally, and N.J. Rose, "A digital signal processing system for automatic dialogue post-synchronisation," presented at the 83rd AES Convention, Preprint 2546 (K-7), New York, USA, October 16-19 1987.

[5] P. Soens and W. Verhelst, "Split time warping of speech for robust Automatic Dialogue Replacement," in *Proc. of XIII-th Convention of Electrical Engineering (CIE)*, Santa Clara, Cuba, June 18-22 2007.

[6] R.M. Chamberlain and J.S. Bridle, "ZIP: a dynamic programming algorithm for time-aligning two indefinitely long utterances," in *Proc. of ICASSP'83*, Boston, USA, April 14-16 1983, vol. 8, pp. 816–819.

[7] L.R. Rabiner, A.E. Rosenberg, and S.E. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," *IEEE Trans. on ASSP*, vol. 26, no. 6, pp. 575–582, December 1978.

[8] S. Dixon and G. Widmer, "MATCH: a music alignment tool chest," in *Proc. of ISMIR'05*, London, UK, September 11-15 2005, pp. 492–497.

[9] M.J. Hunt, "Time alignment of natural speech to synthetic speech," in *Proc. of ICASSP'84*, San Diego, USA, March 19-21 1984, vol. 9, pp. 65–68.

[10] W. Verhelst and M. Borger, "Intra-speaker transplantation of speech characteristics. An application of waveform vocoding techniques and DTW," in *Proc. of EUROSPEECH'91*, Genova, Italy, September 24-26 1991, pp. 1319–1322.

[11] W. Verhelst, "Automatic post-synchronization of speech utterances," in *Proc. of EUROSPEECH'97*, Rhodes, Greece, September 22-25 1997, pp. 899–902.

[12] W. Verhelst and H. Brouckxon, "Rejection phenomena in inter-signal voice transplantations," in *Proc. of IEEE WASPAA'03*, New Paltz, New York, October 19-22 2003, pp. 165–168.

[13] B. Resch and W.B. Kleijn, "Time synchronization of speech," in *Proc. of MAVEBA'03*, Firenze, Italy, December 10-12 2003, pp. 215–218.

[14] R.E. Bellman and S.E. Dreyfus, "Applied dynamic programming," Princeton University Press, Princeton, New Jersey, 1962.

[15] P. Soens and W. Verhelst, "Automatische Nachsynchronisierung von Dialogen - Vorstellung einer ADR-Softwarelösung," *Fachverlag Schiele & Schön GmbH, FKT*, vol. 62, pp. 681–685, December 2008 (in German).

[16] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. on ASSP*, vol. 26, no. 1, pp. 43–49, February 1978.

[17] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. on ASSP*, vol. 23, no. 1, pp. 67–72, February 1975.

[18] C. Myers, L.R. Rabiner, and A.E. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Trans. on ASSP*, vol. 28, no. 6, pp. 623–635, December 1980.

[19] W.S. Cleveland and S.J. Devlin, "Locally weighted regression: an approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 596–610, September 1988.

[20] "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Tech. Rep. ITU-T Rec. P.862, Geneva, Switzerland, February 2001.

[21] Synchro Arts Ltd. (http://www.synchroarts.com/), *VocALign PRO AudioSuite Plug-In for Digidesign® Pro Tools® user manual (version 1.0.5)*, 2005.

[22] E. Campione and J. Véronis, "A large-scale multilingual study of silent pause durations," in *Proc. of Speech Prosody'02*, Aix-en-Provence, France, April 11-13 2002, pp. 199–202.

[23] "Methods for subjective determination of transmission quality," Tech. Rep. ITU-T Rec. P.800, Geneva, Switzerland, August 1996.