AN ITERATIVE BILINEAR FREQUENCY WARPING APPROACH TO ROBUST SPEAKER-INDEPENDENT TIME SYNCHRONIZATION

Pieter Soens and Werner Verhelst

Vrije Universiteit Brussel, Department ETRO-DSSP, Pleinlaan 2, B-1050 Brussels, Belgium
Interdisciplinary Institute for Broadband Technology, G. Crommenlaan 8, B-9050 Ghent, Belgium
E-mail: psoens@etro.vub.ac.be (P. Soens); wverhels@etro.vub.ac.be (W. Verhelst)

ABSTRACT

Vocal Tract Length Normalization is a widely deployed speaker normalization technique, which compensates for vocal tract length differences among speakers by appropriately warping the frequency axis of the speech signal. In this work, we study the use of this technique on the time synchronization paradigm. An efficient bilinear frequency warping procedure is proposed, in which the amount of warping is iteratively optimized in accordance with a criterion that is directly related to the output of the standard Dynamic Time Warping algorithm. Subjective listening tests performed on mixed-gender time-aligned results obtained with a subset of data from the English EUROM1 Many Talker Set have shown that the proposed procedure significantly improves the overall speech quality and the time synchronization accuracy with 85% and 91%, respectively.

Index Terms— Time Synchronization, Vocal Tract Length Normalization, Dynamic Time Warping

1. INTRODUCTION

One prominent problem common to many areas of speech research and applied speech processing concerns the degradation in overall system performance due to the across-speaker variability of the acoustic speech signal. While the source of this variability stems from a complex combination of many factors, such as differences in speaking styles and pronunciation, it is commonly agreed that a major part of the variability is due to physiological differences between speakers, in particular due to differences in their vocal tract length (VTL) and shape. In one of the simplest physiological models, the human vocal tract is treated as a uniform tube resonator [1]. According to this model, the resonant or formant frequencies are inversely proportional to the length of the tube. As a result, early Vocal Tract Length Normalization (VTLN) schemes tried to neutralize speaker-specific aspects by linearly warping the frequency axis of the speech signal. In reality however, the relationship between the VTL and the formant positions is highly context-dependent and in consequence not purely linear, which explains the use of more sophisticated warping functions in later approaches. All these functions have in common that they depend on only one (or a few) parameter(s) and that they conserve bandwidth and information in the original spectrum. Broadly speaking, VTLN approaches can be distinguished by the shape of the frequency warping function used and by the method by which the parameter(s) of this function is (are) estimated.

This research was funded in part by the "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 "studie ter Exploitatie van Onderzoeksresultaten op het vlak van Spraakmodificatie".

In recent work [2], we proposed a system for the automatic time synchronization of two renditions of the same speech utterance, and we investigated the viability of such a system for the application of Automatic Dialogue Replacement (ADR), a well-known audio post-production technique, used to synchronize a revoiced studio recording with the corresponding recording made on the film set. Although we demonstrated that the system performs well when the two recordings have been produced by the same speaker, i.e. the actor, further experiments using speakers of opposite gender have shown that the system performance degrades rapidly when moving from speaker-dependent to speaker-independent time synchronization.
The remainder of this paper is organized as follows: in the next section, we shortly review the time synchronization framework. The main contribution of our work is then outlined in section 3, in which we study the effectiveness of a VTLN scheme based on a bilinear frequency warping approach with the aim of improving the across-speaker robustness of time synchronization. The main novelty of the proposed method lies in the way the amount of frequency warping is estimated: while this parameter is typically optimized in the Maximum Likelihood (ML) sense within the context of Hidden Markov Model (HMM) based speech *recognition* [3], we now devise the optimization criterion in the context of a speech *synthesis* application. In section 4, we evaluate the proposed method in terms of overall speech quality and time synchronization accuracy. Finally, in section 5, we draw the conclusions from the results.

2. TIME SYNCHRONIZATION FRAMEWORK

Fig. 1 reviews the functional block diagram of our automatic time synchronization system. The system uses a synthesis-after-analysis approach to modify the timing structure of a replacement speech utterance (U_y) such that the result (U_z) is precisely synchronized with the speech utterance that serves as the timing reference (U_x). In the first step of the analysis, which aims to solve the difficult problem of precisely inserting new, and deleting or resizing existing non-speech segments, such as breathing pauses, a dedicated Dynamic Time Warping (DTW) algorithm is used to identify the corresponding speech segments in both the replacement and reference speech waveforms. Thereafter, the timing relationship for each pair of matching speech segments is computed and then processed in such a way that the time-scale modification of the replacement speech segments is performed more gradually (smoothing step) while at the same time the speech rate of the time-scaled result is systematically controlled in relation to that of the timing reference (post-processing step). It was found that this approach produces results that are both natural-sounding and well-synchronized with the timing reference. Details of the system can be found in [2].

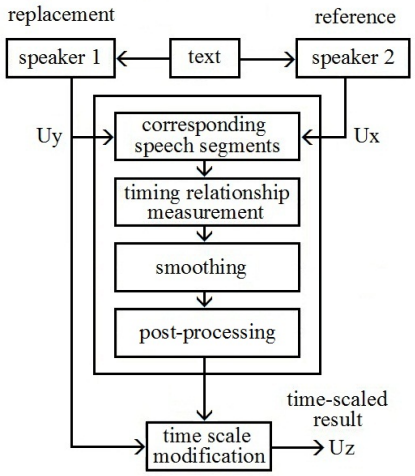

Fig. 1. Time synchronization framework [2].

3. METHOD

This section presents a detailed description of the procedure that is used to achieve a more robust speaker-independent time synchronization. The proposed procedure attempts to reduce the variability in spectral formant peak positions for corresponding speech sounds produced by different speakers. In order to compensate for this variability, the acoustic feature vectors of one of the speakers are transformed in order to improve their match with the corresponding feature vectors of the other speaker[1]. This is achieved by means of a spectral frequency warping operation that is controlled by only one parameter, which we will call the warping factor from now on. In the work described here, an iterative procedure is used to estimate the best warping factor for each pair of speech utterances in accordance with an optimization criterion that is directly related to the output of the DTW algorithm. The following subsection describes exactly how this is accomplished. Further, in subsection 3.2, we discuss how the frequency warping is implemented in the DTW feature extraction front-end. Finally, in subsection 3.3, we shortly address the convergence properties of the proposed procedure. Throughout this work, we assume that the reader is acquainted with the concepts of DTW (A short review can for example be found in [2]).

3.1. Iterative Warping Factor Estimation Procedure

The warping factor estimation procedure can be described mathematically as follows. In the DTW analysis of speech utterance pair u, the samples of the r-th (reference) speech frame of U_x and the s-th (replacement) speech frame of U_y, obtained by applying an N-point tapered (Hamming) window to the sampled and pre-emphasized speech waveform, are denoted by $x_{u,r}(n)$ and $y_{u,s}(n), n=0\ldots N-1$, and the corresponding discrete-time spectra by $X_{u,r}(e^{i\omega})$, $r=1\ldots R$ and $Y_{u,s}(e^{i\omega})$, $s=1\ldots S$, respectively. Furthermore, let $\mathbf{x}_{u,r}$ and $\mathbf{y}_{u,s}$ be the corresponding feature vectors obtained from these spectra, then the utterance pair can be represented by the pair of vector sequences $\mathbf{X}_u=\{\mathbf{x}_{u,1},\mathbf{x}_{u,2},\ldots,\mathbf{x}_{u,R}\}$ and $\mathbf{Y}_u=\{\mathbf{y}_{u,1},\mathbf{y}_{u,2},\ldots,\mathbf{y}_{u,S}\}$.

[1]Without loss of generality, we assume in this work that the transformation is performed on the acoustic features of the reference speaker.

In the context of frequency warping, we now define the warped spectrum $X^{\alpha}_{u,r}(e^{i\omega}) \triangleq X_{u,r}(e^{ig_\alpha(\omega)})$, in which $g_\alpha(\omega)$ denotes the applied frequency warping function with warping factor α. The feature vectors computed from these spectra can then be represented by $\mathbf{x}^{\alpha}_{u,r}$ and the warped representation of U_x as the sequence of vectors $\mathbf{X}^{\alpha}_u=\{\mathbf{x}^{\alpha}_{u,1},\mathbf{x}^{\alpha}_{u,2},\ldots,\mathbf{x}^{\alpha}_{u,R}\}$. Then, the optimal warping factor α^{opt}_u, in the DTW sense, for the u-th pair of speech utterances is obtained from

$$\alpha^{opt}_u = \begin{cases} \arg\min\limits_{\alpha} D_{\boldsymbol{\tau}}(\mathbf{X}^{\alpha}_u,\mathbf{Y}_u) \\ \arg\min\limits_{\alpha} \sum\limits_{k=1}^{K} w_k\, d(\mathbf{x}^{\alpha}_{u,k},\mathbf{y}_{u,k}) \end{cases} \tag{1}$$

in which $d(\mathbf{x}^{\alpha}_{u,k},\mathbf{y}_{u,k})$ denotes a local spectral distance, e.g. the squared Euclidean Distance between Mel Frequency Cepstral Coefficient (MFCC) feature vectors $\mathbf{x}^{\alpha}_{u,k}$ and $\mathbf{y}_{u,k}$, extracted from U_x and U_y at time instants x_k and y_k, respectively. Further, $D_{\boldsymbol{\tau}}(\mathbf{X}^{\alpha}_u,\mathbf{Y}_u)$ represents the global accumulated DTW distance, which is computed as the weighted sum of local spectral distances along the time warping path $\boldsymbol{\tau}_u=\langle(x_k,y_k)\rangle, k=1\ldots K$. This warping path is obtained with the standard DTW algorithm, which assumes that the unbiased Sakoe-Chiba weighting function w_k is used [2], and the vector sequences $\mathbf{X}^{\alpha}_u$ and $\mathbf{Y}_u$. Since a closed form solution for α^{opt}_u from Eq. (1) is difficult to obtain, we propose to compute it by means of an iterative search procedure. Let, for this purpose, $\hat{\alpha}^{(i)}_u$ represent the i-th estimate of α^{opt}_u, and $\boldsymbol{\tau}^{(i)}_u$ the corresponding time warping path (of length K_i) obtained with the standard DTW algorithm and the vector sequences $\mathbf{X}^{\hat{\alpha}^{(i)}_u}_u$ and $\mathbf{Y}_u$, then the $i+1$-th estimate of α^{opt}_u is found by minimizing the global accumulated DTW distance along $\boldsymbol{\tau}^{(i)}_u$ for all possible α within a discrete set of values R_i

$$\hat{\alpha}^{(i+1)}_u = \arg\min_{\alpha\in R_i} \sum_{k=1}^{K_i} w_k\, d(\mathbf{x}^{\alpha}_{u,k},\mathbf{y}_{u,k}) \tag{2}$$

As starting condition, we set $\hat{\alpha}^{(0)}_u=0 \leftrightarrow \boldsymbol{\tau}^{(0)}_u$ (no warping, see subsection 3.2), for which it is assumed that the iteration will converge to $\hat{\alpha}^{(I)}_u \approx \alpha^{opt}_u \leftrightarrow \boldsymbol{\tau}^{(I)}_u \approx \boldsymbol{\tau}^{opt}_u$ after I steps. We will address the validity of this assumption in subsection 3.3. Further, the first estimate of the warping factor, $\hat{\alpha}^{(1)}_u$, is searched over a grid of 51 values spaced evenly between -0.25 and 0.25. This search grid (R_0) was designed to account for a frequency variation of approximately 30% at $\omega=\pi/2$. In the subsequent stages of the iteration, the range of values within which the warping factor was optimized was systematically reduced to $\hat{\alpha}^{(i)}-0.25+0.05\,i \le \hat{\alpha}^{(i+1)} \le \hat{\alpha}^{(i)}+0.25-0.05\,i$ for $0\le i\le 4$, and to $\hat{\alpha}^{(i)}-0.05 \le \hat{\alpha}^{(i+1)} \le \hat{\alpha}^{(i)}+0.05$ for $i\ge 5$.

3.2. Frequency Warping Implementation

In the literature, several techniques have been proposed to implement VTLN. The most intuitive method is to resample the speech waveform in the time domain before front-end processing [4]. Other approaches have implemented VTLN in the frequency domain, either by compressing or expanding the speech signal in the Fourier domain using spectrum interpolation, or by modifying the width and spacing of the component filters of the filterbank, used in the MFCC computation [3]. Yet other approaches are based on the work of Pitz et al., who proved that VTLN can equally be implemented by means of a linear transformation in the cepstral domain, provided that the applied frequency warping function is invertible [5].

One particular example of such a function is the Bilinear Transform (BLT), which for the present purpose will be expressed as

$$Q(z) = \frac{z-\alpha}{1-\alpha z} \tag{3}$$

for some real-valued warping factor α. In [6], McDonough has shown that $Q(z)$ belongs to a class of rational all-pass transforms (RAPT), which, under suitable analyticity constraints, provide the means to transform a causal discrete-time sequence $x(n)$ into a new sequence $x^{\alpha}(n)$, such that their Z-transforms are related by $X^{\alpha}(z) = X(Q(z))$. The transformed sequence is given by

$$x^{\alpha}(n) = \sum_{m=0}^{+\infty} q^{(m)}(n)\, x(m) \tag{4}$$

in which

$$q^{(m)}(n) = Z^{-1}\left\{Q^m(z)\right\} = \frac{1}{2\pi i}\oint_{C^+} Q^m(z)\, z^{n-1} dz \tag{5}$$

for all $m \geq 0$. The integration must be performed in the region of convergence (ROC) of $Q^m(z)$ along a closed contour C in counter-clockwise direction. From Eq. (5), it is apparent that $q^{(0)}(n) = \delta(n)$, and the sequences $q^{(m)}(n)$ can be computed recurrently from the Cauchy product $q^{(m)}(n) = q^{(m-1)}(n) * q^{(1)}(n)$ for all $m \geq 2$, once $q^{(1)}(n)$ is known. Provided that $|\,\alpha\,| < 1$, the latter is available from inspection of the geometric series expansion of $Q(z)$ on the unit circle $z = e^{i\omega}$. In sum, the sequences $q^{(m)}(n)$ can be found from

$$q^{(0)}(n) = \delta(n) \tag{6a}$$

$$q^{(1)}(n) = \begin{cases} -\alpha & n = 0 \\ (1-\alpha^2)\,\alpha^{n-1} & n > 0 \end{cases} \tag{6b}$$

$$q^{(m)}(n) = \sum_{k=0}^{n} q^{(m-1)}(k)\, q^{(1)}(n-k) \qquad m \geq 2, n \geq 0 \tag{6c}$$

and Eqs. (4),(6) now set forth a procedure to transform a sequence $x(n)$ into a new sequence $x^{\alpha}(n)$, whose Fourier transform equals that of the original sequence evaluated on a warped frequency axis $\hat{\omega} = \angle Q(e^{i\omega}) \triangleq g_{\alpha}(\omega)$, such that $X^{\alpha}(e^{i\omega}) = X(e^{i g_{\alpha}(\omega)})$, in which

$$g_{\alpha}(\omega) = \omega + 2\arctan\frac{\alpha\sin\omega}{1-\alpha\cos\omega} \tag{7}$$

represents the bilinear frequency warping function. A plot of this function for 3 different α-values is shown in Fig. 2. From this figure, we can see that the frequency mapping is inherently non-linear. For example, for positive values of α, all frequencies of the original spectrum, except $\omega = 0$ and $\omega = \pi$, are downshifted in absolute terms; as a matter of fact, the frequency range $0 < \omega < \arccos\alpha$ is compressed and the frequency range $\arccos\alpha < \omega < \pi$ is expanded. Finally, we remark that, although Eq. (4) involves infinite series, the sequences are typically of finite length for $x(m)$, and can be truncated for $x^{\alpha}(n)$ when $|\,\alpha\,| < 1$. In our implementation for example, the reference speech segments $x_{u,r}(n)$ were subject to the transformation, represented by Eq. (4), before traditional MFCC computation, using N terms for both the original and transformed sequences, corresponding to a frame length of $25ms$.

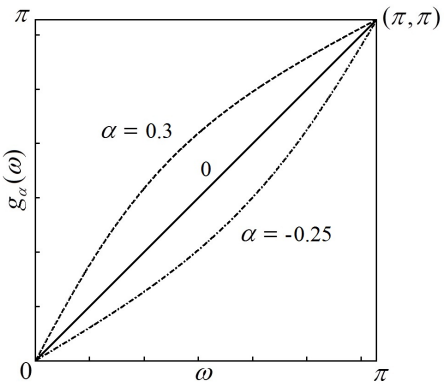

Fig. 2. Illustration of the bilinear frequency warping function $g_{\alpha}(\omega)$ for 3 different values of the warping factor α.

3.3. Analysis of Proposed Iterative Procedure

In this section, we describe an experiment that we performed to better understand the properties of the proposed iterative procedure. For this purpose, we asked one adult male and one adult female speaker to produce each 2 read versions of 8 text fragments (one excerpt from a novel, 2 from a journal paper, and 5 from the EUROM1 corpus [7]). We manually segmented each of these 4 recordings into 37 segments of continuous speech, and then applied the iterative procedure to the 148 pairs of matching speech segments, each time using the time pattern of the male speech samples as the timing reference. Thereafter, we counted the obtained warping factors in each of the bins of the R_0 range. This is illustrated in the resulting histogram of Fig. 3 (a): values of estimated warping factors are displayed along the horizontal axis, and the number of utterance pairs that were assigned to each given warping factor is plotted on the vertical axis. From the figure, we can see that all warping factors vary from -0.23 to -0.12 with 66.89% and 89.19% having a value to within one and two percent of the median ($m = -0.18$), respectively. These negative warping factor values are consistent with the fact that, on average, male speakers have a longer VTL and exhibit lower central formant frequencies than women. As a result, it is reasonable that the proposed procedure chooses to shift the formants of the male spectra upwards ($\alpha < 0$) in order to improve the match with the formants of the corresponding female spectra.

Fig. 3 (b) shows the distribution of the required number of iteration steps that are necessary to reach convergence to within one percent accuracy. From the figure, we can see that for the major part of the iterations (97.30%) convergence was reached in 2, 3 or 4 steps, and in precisely half of the cases this occurred after the median value of 3 steps. At last, we have computed, concurrently with each estimated warping factor value $\hat{\alpha}_u^{(i)}$, $i = 1 \ldots I$, a discrepancy score $D_u^{(i)}$, which reflects the normalized area between the time warping paths $\boldsymbol{\tau}_u^{(i)}$ and $\boldsymbol{\tau}_u^{(i-1)}$. We then pooled individual scores computed for different speech utterance pairs per iteration step index, and represented the average of each cluster as a vertical bar in Fig. 3 (c). From this figure, we can see that the resulting warping path discrepancy function decreases most rapidly for the lower order iteration step indices. Assuming that large discrepancy scores correspond to synthesized results in which large differences in both speech quality and time synchronization accuracy can be perceived, this suggests that the lower order iteration steps contribute most to the overall system performance improvement.

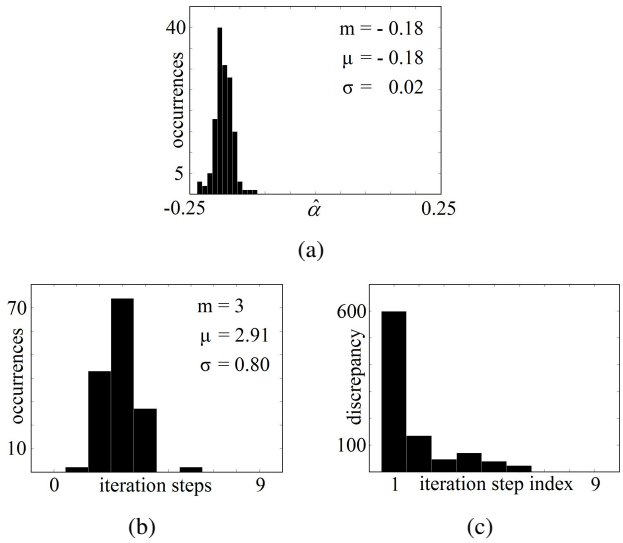

Fig. 3. (a) Histogram of estimated warping factors; (b) Distribution of required number of iteration steps necessary to reach convergence to within one percent; (c) Bar graph of mean normalized warping path discrepancy in function of the iteration step index.

4. EVALUATION

4.1. Database

We have evaluated the proposed procedure in terms of overall speech quality and time synchronization accuracy by means of subjective listening tests using time-aligned results obtained with data from the English sentence recordings of the EUROM1 "Many Talker Set" [7]. These recordings have been structured into 10 blocks (F0 to F9), each one of which is composed of 6 renditions of a sequence of 5 sentences that are isolated by pauses. Each of these 6 five-filler-sentence sequences has been produced by a combination of 6 male and female adults, all native speakers of their own language. We remark that both speakers as well as sentence sequences differ across the 10 blocks, and that there was an equal amount of 30 male and 30 female speakers in total. However, within each block, the number of male and female speakers was not always evenly balanced.

4.2. Selection of Mixed-Gender Speaker Pairs

A subset of the aforesaid data was selected with the aim of keeping a reasonable amount of evaluation data, small enough to be viable for subjective evaluation, but large enough to add sufficient power to the statistical analysis. Moreover, we ensured that those speech sample pairs were included for which it could be reasonably anticipated that VTLN could be most desirable. This was accomplished by means of the following procedure: we first manually endpointed each of the 5 sentences in all of the 60 five-filler-sentence sequences, and then removed non-speech segments (e.g. short breathing pauses), in so far as they occurred, in all of the 300 resulting speech samples. Thereafter, we considered within each block all possible combinations of one male and one female speaker and, for each combination, we applied the proposed procedure to the 5 corresponding speech sample pairs, each time using the time pattern of the male speech sample as the timing reference. For each pair of speech samples u, we then computed (as in subsection 3.3) the discrepancy between the smoothed initial and converged warping paths ($\widetilde{\tau}_u^{(0)}$ and $\widetilde{\tau}_u^{(I)}$ in Table 1, respectively). The 5 discrepancy scores for each pair of speakers were then averaged, and the largest mean score of each block was then added to a list. Eventually, the 5 largest scores of this list defined which speaker pairs (and corresponding speech sample pairs) were kept for the experiment. In order to even the balance of mixed-gender results in the evaluation phase, we applied a similar procedure to select another 25 pairs of speech samples, but this time we used the time pattern of the female speech samples as timing reference in the iterative procedure.

Time-aligned result	Smoothed DTW Path	Post-processing
Baseline (B)	$\widetilde{\tau}_u^{(0)}$	no
Intermediate (I)	$\widetilde{\tau}_u^{(I)}$	no
Proposed (P)	$\widetilde{\tau}_u^{(I)}$	yes

Table 1. Types of time-aligned results. Smoothing window used was a $500ms$ tricube window, and the postprocessing parameter was set equal to 1.1 (see also [2]). Some experimental results can be downloaded from http://www.etro.vub.ac.be/research/DSSP/demo/ADR.

4.3. Experiment

For each of the 50 selected pairs of speech samples, we used the Waveform Similarity OverLap-and-Add (WSOLA) algorithm with the same parameter settings as in [2] to time-align the replacement speech waveform with the corresponding speech waveform that served as timing reference. Depending on the warping path that was used to achieve this, we distinguish 3 types of time-aligned results (see Table 1). In addition, we created a second set of evaluation data by mixing each of the 150 time-aligned results with their corresponding timing reference. We will refer to these mixed speech waveforms as BR, IR and PR for the baseline, intermediate and proposed methods, respectively.

4.4. Subjective Listening Tests

In order to assess and compare the results in terms of perceived overall speech quality and time synchronization accuracy, we performed 2 subjective listening tests using a group of 18 test listeners. For this purpose, the results were assembled in 2 groups of 50 triplets (X,Y, Z), in which X, Y and Z represent either a single-ended (B,I,P) or a mixed (BR,IR,PR) result. In both tests, the order of presentation of the samples in each triplet was randomized with the constraint that X, Y, and Z represented a result obtained with the baseline, intermediate or proposed system in at least 32% of the times in order to neutralize observator bias effects such as the primacy and recency effect. For each of the 50 triplets in each test, we asked the test subjects to rate the perceived overall speech quality or respectively time synchronization accuracy of the corresponding X, Y and Z samples by assigning scores to their opinions. Each of these scores is a number in the range from 1 to 5 and has the following meaning: the overall perceived quality (of the speech sample itself or respectively its time synchronization accuracy with the reference) was excellent (5), good (4), fair (3), poor (2), or bad (1).

4.5. Results

Tables 2 and 3 show for both studied aspects the arithmetic means evaluated from the opinion scores for each and across all speaker pairs (MOS), as well as the sample standard deviation (s), standard error of the mean (SEM) and the 95% confidence interval (95%CI).

Overall Speech Quality				
SC_x	SC_y	B	I	P
MP	MJ	1.31	3.49	4.20
NC	NM	2.31	3.25	3.77
NX	MT	1.36	1.95	4.29
NO	NY	1.76	3.47	4.10
NG	MW	1.09	1.87	3.47
MJ	MP	1.60	3.22	4.29
OE	NL	1.91	3.64	4.18
NM	NC	2.41	4.03	4.34
MT	NX	1.41	3.03	4.27
MW	NG	1.34	2.71	4.19
MOS		**1.65**	**3.06**	**4.11**
s		**0.86**	**1.18**	**0.89**
SEM		**0.03**	**0.04**	**0.03**
95%CI		**1.59...1.71**	**2.99...3.14**	**4.05...4.17**
Δ_{xy},Δ_{yz},Δ_{xz}		**1.41**	**1.05**	**2.46**
d_{xy},d_{yz},d_{xz}		**1.36**	**1.00**	**2.79**
δ_{xy},δ_{yz},δ_{xz}(%)		**85.45**	**34.31**	**149.09**
$F(p)$		**1408(<0.0001)**		
p_{xy},p_{yz},p_{xz}		**<0.001**	**<0.001**	**<0.001**

Table 2. Results of statistical analysis with regard to the subjective evaluation of the overall speech quality of the time-aligned results. SC_x and SC_y represent EUROM1 speaker codes. Subscripts x and y indicate which speaker was considered the reference or respectively replacement speaker in the time alignment procedure.

Overall Time Synchronization Accuracy				
SC_x	SC_y	BR	IR	PR
MP	MJ	1.97	4.07	3.53
NC	NM	2.35	4.02	3.34
NX	MT	1.58	2.62	3.98
NO	NY	2.04	4.22	4.15
NG	MW	1.28	2.42	2.89
MJ	MP	1.96	3.74	3.43
OE	NL	2.25	4.16	3.69
NM	NC	2.52	4.19	3.90
MT	NX	1.41	3.19	4.02
MW	NG	1.37	3.04	3.27
MOS		**1.87**	**3.57**	**3.62**
s		**0.96**	**1.17**	**1.00**
SEM		**0.03**	**0.04**	**0.03**
95%CI		**1.81...1.94**	**3.49...3.64**	**3.55...3.69**
Δ_{xy},Δ_{yz},Δ_{xz}		**1.70**	**0.05**	**1.75**
d_{xy},d_{yz},d_{xz}		**1.58**	**0.05**	**1.77**
δ_{xy},δ_{yz},δ_{xz}(%)		**90.91**	**1.40**	**93.58**
$F(p)$		**1067(<0.0001)**		
p_{xy},p_{yz},p_{xz}		**<0.001**	**>0.05**	**<0.001**

Table 3. Results of statistical analysis with regard to the subjective evaluation of the overall time synchronization accuracy between the time-aligned results and their corresponding timing references.

In addition are given the raw (Δ_{xy},Δ_{yz},Δ_{xz}) and standardized differences (d_{xy},d_{yz},d_{xz}) in overall mean MOS scores, and also the relative improvement of the different systems over each other (δ_{xy},δ_{yz},δ_{xz}). Since the distribution of the MOS scores proved to be far from Gaussian, we performed, for each of the studied aspects, the Friedman test followed by Dunn's post test to verify the statistical significance of the observed differences in mean MOS scores between the 3 systems. Tables 2 and 3 report the Friedman statistic (F) as well as the overall (p) and between-groups (p_{xy},p_{yz},p_{xz}) p-values. With regard to the overall speech quality, we can conclude that there is compelling evidence at the 5% level that the differences in overall mean MOS scores for the 3 systems are statistically significant. A similar conclusion can be drawn with regard to the overall time synchronization accuracy, except for the difference in overall mean MOS scores between the intermediate and proposed system, which proved to be statistically less significant ($p_{yz} > 0.05$).

5. CONCLUSIONS

From the results, we can draw a number of interesting conclusions. At first, with regard to the overall speech quality of the time-aligned results, we can conclude that both the VTLN and DTW path postprocessing procedures significantly improve the results (from in-between bad and poor over fair to good), and that the contribution of the former to the overall improvement is more pronounced than that of the latter. Furthermore, with regard to the overall time synchronization accuracy, the results show that the contribution of the VTLN procedure to the overall improvement is even slightly more pronounced than that to the overall speech quality improvement. However, the post-processing procedure has no additional effect this time: this is not surprising, as it was specifically designed to make the time-scaled results sound more natural [2].

We remark that the large scores for the overall relative improvement in speech quality and time synchronization accuracy (149% and 94%, respectively) should be considered upper limits that are valid for this particular data set: in reality, it can be anticipated that these scores will be smaller when the EUROM1 speech sample pairs would have been selected at random. Nevertheless, the effectiveness of the proposed VTLN procedure to the robustness improvement of speaker-independent time synchronization has been clearly demonstrated.

6. REFERENCES

[1] G. Fant, "Non-uniform vowel normalization," *Speech Transmission Laboratory – Quarterly Progress and Status Reports (STL-QPSR)*, vol. 16, no. 2-3, pp. 1–19, 1975.

[2] P. Soens and W. Verhelst, "On split Dynamic Time Warping for robust Automatic Dialogue Replacement," *Signal Processing*, vol. 92, no. 2, pp. 439–454, February 2012.

[3] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 49–60, January 1998.

[4] A. Andreou, T. Kamm, and J. Cohen, "Experiments in vocal tract normalization," in *Proceedings CAIP Workshop: Frontiers in Speech recognition II (CAIP-1994)*, 1994.

[5] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 930–944, September 2005.

[6] J.W. McDonough, *Speaker compensation with all-pass transforms*, Ph.D. thesis, Johns Hopkins University, 2000.

[7] D. Chan, A. Fourcin, D. Gibbon, B. Grandstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. in 't Veld, and J. Zeiliger, "EUROM - A spoken language resource for the EU - The SAM projects," in *Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH-1995)*, Madrid, Spain, September 18-21 1995, vol. 1, pp. 867–870.