DIVIDE-AND-WARP TEMPORAL ALIGNMENT OF SPEECH SIGNALS BETWEEN SPEAKERS: VALIDATION USING ARTICULATORY DATA

Sai Muralidhar Jayanthi[1,2] *Lucie Ménard*[3] *Catherine Laporte*[2]

[1] Department of Electronics and Electrical Engineering, Indian Institute of Technology, Guwahati, India
[2] Department of Electrical Engineering, École de technologie supérieure, Canada
[3] Department of Linguistics, University of Quebec in Montreal, Canada

ABSTRACT

Meaningful comparisons between sets of speech-induced, dynamically evolving articulatory measurements require that the data be temporally aligned in a manner invariant to speech rate discrepancies. The best known approach to this problem is to apply dynamic time warping (DTW) to the corresponding audio signals. While the usefulness of DTW methods is well established in automatic speech recognition, they were never directly and quantitatively validated as a way of aligning and comparing signals from different speakers in a way that is useful for the study of speech as a biological process. This paper provides the first direct quantitative validation of such an audio-based temporal alignment algorithm, itself based on a new divide-and-warp strategy, using speaker invariant temporal landmarks from articulatory data. Results demonstrate that the proposed temporal alignment algorithm accurately brings these landmarks into correspondence between speakers (mean absolute delay of $\sim 35ms$).

Index Terms— Temporal alignment, dynamic time warping, between-speaker comparisons, validation, articulatory measurements.

1. INTRODUCTION

Observing and characterizing articulation is of great importance to the study of healthy and impaired speech as biological processes. Measurements from electromagnetic articulography (EMA), ultrasound (US) imaging, and magnetic resonance (MR) imaging are commonly used for such purposes. Many methods exist to characterize complex static configurations of articulators such as the shape of the lips or the tongue, extracted at key instants (e.g., the middle of vowels) in EMA, US or MR recordings, in relation to different sounds and/or speaker populations [1, 2, 3]. Other studies have analyzed *changes* in articulator configuration over time, and there is much to be gained by doing so (see, e.g., Woo et al.'s work on elucidating the functional units of speech induced tongue motion [4], or Mitra et al.'s work exploiting articulatory information for speech recognition [5]). However, analyzing articulatory time series in a meaningful way is more difficult than analyzing static data [6]. These difficulties stem from (at least) two fundamental problems: (1) non-uniform variations in speech rate across similar utterances within and between speakers must be considered, and (2) statistical methods must be adapted to the context of multivariate, time-dependent and rate-dependent measurements. This paper considers the first of these problems.

Speech is subject to rate variations both between and within individual speakers, across different utterances of a given sentence, expression, word or phoneme. Meaningful comparisons between sets of speech-induced dynamic articulatory measurements requires that the articulatory data be temporally aligned in a manner invariant to these rate variations, which are typically non-uniform over time [7]. While feasible and sometimes necessary, temporal alignment estimated directly from the articulatory data [8, 9], is influenced by individual or sub-population articulatory idiosyncrasies. This poses a problem as it is precisely such idiosyncrasies that might be of interest when studying speech as a biological process. Thus, the best known way to estimate the required alignment, as independently as possible from the articulatory data of interest, is to apply dynamic time warping (DTW) algorithms to the audio speech signal [10, 11]. DTW and its variants [12, 13, 14, 15] compute the one-to-one monotonic transformation of the time axis (a *time warping function*) that minimizes a dissimilarity measure between a sample signal and a reference signal. This optimal time warping function can then, in principle, be applied to articulatory data acquired simultaneously with the sample audio speech signal to align them with the reference articulatory data for comparison.

The discriminative value of the *dissimilarity measure* minimized by DTW is well understood in automatic speech recognition [16] and indexing applications [17]. However, the literature lacks direct evidence that the *time warping functions* produced by DTW or any of its variants align speech signals produced by different speakers in a way that is useful for the study of speech as a biological process. Also, computation time for DTW scales as the square of recording length, making it impractical for aligning full sentences. This work addresses the latter issue by applying DTW indepen-

dently to the voiced and unvoiced segments of recordings from different speakers (see Section 2). More importantly, it addresses the former, fundamental issue of validation by demonstrating the accuracy of alignment between speaker-invariant temporal landmarks from independently acquired EMA measurements based on the time warping functions estimated from the audio data (Section 3).

2. DIVIDE-AND-WARP TEMPORAL ALIGNMENT

The objective of temporal alignment is to find a non-linear time warping function F_n mapping a test pattern T to a reference pattern R. In automatic speech recognition, the patterns are typically feature vectors extracted from short (e.g. single word) audio speech data sequences. Here, the patterns are raw audio sequences as long as a full sentence.

The proposed method exploits the fact that typical speech is a juxtaposition of voiced (e.g. vowels and voiced consonants) and unvoiced (e.g. unvoiced consonants) segments. It is assumed that given two utterances R and T of a given expression by two speakers, the correct time warping function should not map samples of one type in T to samples of the other type in R. Such unacceptable mappings are prevented by first detecting the voiced and unvoiced segments in both R and T and independently aligning corresponding segments of the recordings using conventional DTW or one of its variants. This divide-and-warp strategy has the added benefit of drastically reducing DTW computation time, making time alignment tractable and practical over entire sentences.

The remainder of this section first explains how audio data are broken into voiced and unvoiced segments (Section 2.1). It then describes the specific DTW variant used to perform temporal alignment of the resulting segments (Section 2.2), and explains how to apply the resulting time warping functions to articulatory recordings acquired simultaneously with the audio data (Section 2.3).

2.1. Segmentation of audio speech data

In voiced segments, the input excitation is nearly periodic in nature, whereas the excitation in unvoiced segments is similar to random noise. Thus, in this work, detection and labeling of voiced and unvoiced segments is performed based on short term frequency content.

Specifically, local spectral centroids are measured over 25 ms time windows in each audio recording, with consecutive time windows occurring at 5 ms intervals. The voiced segments are taken to correspond to spectral centroids below a certain threshold, and the remaining segments are taken to be unvoiced. A different threshold is required for each given recording; thus, an adaptive, yet straightforward approach is used for threshold selection. The mean spectral centroid over the entire sequence of time windows in the recording was found to be a suitable threshold most of the time. Thus, this is used as an initial threshold value. The threshold is then iteratively altered by small multiplicative factors (no less than 0.75 and no more than 1.25) until the test and reference signals can be divided into equal numbers of (corresponding) segments or until failure is declared. An illustrative example showing the results of this method is shown in Fig. 1.

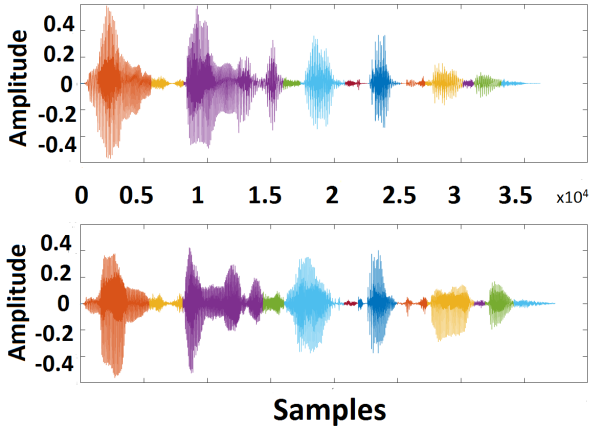


Fig. 1. Divide-and-warp strategy: the reference (top) and test (bottom) recordings of the sentence "Mum strongly dislikes appetizers" are first broken into an equal number of segments corresponding to voiced content, unvoiced content and silence. Corresponding segments (represented by different colors) are then independently aligned to each other.

2.2. Temporal alignment of corresponding segments

Following the segmentation of the audio signal, temporal alignment of corresponding reference segment r of length I and test segment t of length J is performed using a simple variant of classical DTW. Simply put, DTW finds the optimal time alignment between r and t by minimizing the cumulative dissimilarity measure between them along a path within the grid of $I \times J$ points defined by the Cartesian product of $r : \{r_1, r_2, r_3, ..., r_i, .., r_I\}$ and $t : \{t_1, t_2, t_3, ..., t_j, .., t_J\}$.

At each grid point $(i, j) \in \{I, J\}$, a dissimilarity measure $d(i, j)$ between r_i and t_j is computed. In conventional DTW [12, 10], this dissimilarity measure is simply the Euclidean distance $d(i, j) = ||r_i - t_j||$. In this work, we instead use the Derivative DTW dissimilarity measure proposed by Keogh and Pazzani [13], $d(i, j) = ||r'_i - t'_j||$, where

$$r'_i = \frac{(r_i - r_{i-1} + \frac{(r_{i+1} - r_{i-1})}{2})}{2},$$

$$t'_j = \frac{(t_j - t_{j-1} + \frac{(t_{j+1} - t_{j-1})}{2})}{2}$$

are finite difference approximations to the time derivatives of r and t, respectively. This emphasizes matching between subsequences of similar shape rather than similar amplitude, and

prevents undesirable mappings between a single point from one segment to a large subsection of the other.

A path P is a sequence of K grid points $\{p_1, p_2, p_3, ..., p_K\}$ where $p_k = (i_k, j_k)$. The cumulative dissimilarity between t and r for a given path P is the weighted sum of the local distances given by $D[r,t] = \sum_{k=1}^{K} w_k d(p_k)$. In this paper, the symmetric form DP-matching is used with horizontal, vertical and diagonal step-weight coefficients w_k equal to 1, 1 and 2 respectively. The optimal path $P^\star$ is the path that minimizes $D[r,t]$, and is estimated using dynamic programming [12, 10]. The path $P^\star$ realizes a mapping from the time axis of t onto that of r, called the *time warping function*.

Additionally, constraints are imposed on the warping path because when it is viewed as a mapping from the time axis of pattern t onto that of pattern r, it must preserve linguistically meaningful structures in pattern t and vice-versa [12]. The path has to start in $p_1 = (1,1)$ and end in $p_K = (I,J)$ and is constrained by an *adjustment window* to limit the search within a reasonable range, as a good warping path is unlikely to wander far away from the diagonal.

In this work, a dynamic adjustment window parallel to the line joining $(1,1)$ and (I,J) is used. The size of the window linearly increases to a maximum starting from 1 and decreases back to 1 and the size is decided by the path position p_k and hence is maximal at the center of the $I \times J$ grid.

In the proposed implementation, if the path p_k moves along the i (or j)-axis consecutively for $q = 3$ steps, the path is not allowed to step further in the same direction before making at least $p = 1$ step in the diagonal direction. A larger p/q ratio may result in more many-to-one correspondences and thereby hitting horizontal or vertical limit (depending on which sequence is shorter) before reaching (I,J).

2.3. Applying time warping functions to articulatory data

The time warping function estimated from the reference audio data can now be imposed on the test audio data and its corresponding articulatory data. Since the sampling rate of the articulatory data is typically much lower than that of the audio data, the former are linearly interpolated to the length of the latter. The time-aligned articulatory data sequence B : $\{b_1, b_2, b_3, ..., b_I\}$is computed from the original articulatory data sequence A using the non-linear time warping function $F_n : B_i = F_n(A_j)$ corresponding to the cumulation of optimal warp paths $P^\star$ estimated using the DTW algorithm on the segments extracted as part of the divide-and-warp strategy described earlier. The optimal time warping function is likely to be composed of one of the following five correspondences and hence the function F_n is defined in the following way:

1. One reference sample mapped to one test sample:

$$b_i = a_j.$$

2. One reference sample mapped to m test samples:

$$b_i = \sum_{j}^{j+m} a_j, m > 1.$$

3. Two reference samples mapped to one test sample:

$$b_i = \frac{a_{j+1} + 2 * a_j}{3},\ b_{i-1} = \frac{2 * a_j + a_{j-1}}{3}.$$

4. Three reference samples mapped to one test sample:

$$b_i = \frac{a_{j+1} + a_j}{2},\ b_{i-1} = a_j,\ b_{i-2} = \frac{a_j + a_{j-1}}{2}.$$

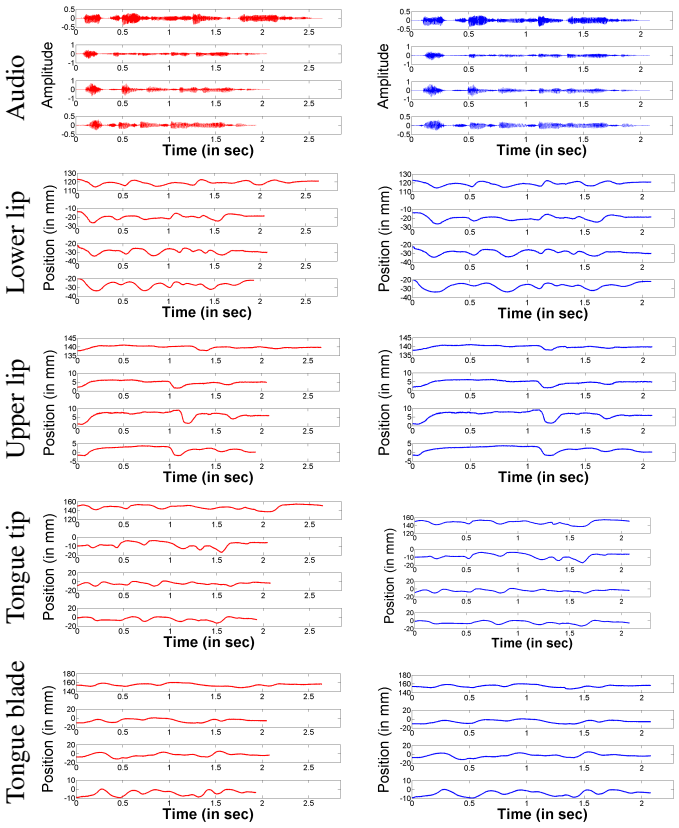
5. $(m > 1)$ reference samples mapped to one test sample: The range between a_{j-1} and a_j, a_j and a_{j+1} is linearly divided into $\lfloor \frac{m}{2} \rfloor$ and $m - \lfloor \frac{m}{2} \rfloor$ parts respectively and assigned to the range $\{b_{i-(m-1)}, b_{i-(m-2)}, ..., b_i\}$.

3. RESULTS

The proposed temporal alignment method was validated using the MOCHA-TIMIT data set [18] containing simultaneous audio and EMA recordings of a large phonetically balanced corpus of short English sentences (ranging over a few seconds) from four speakers, two male and two female, with different accents. The EMA measurements include the positions of the upper and lower incisors, upper and lower lips, tongue tip, tongue blade and tongue dorsum.

Validation was performed using the first 50 sentences in the database. Among the four recordings of the same sentence, one was arbitrarily chosen as a reference, as it was found that the choice of reference did not strongly influence the outcome of temporal alignment. The divide-and-warp DTW algorithm (Section 2) was used to estimate time warping functions from the audio data, which were then applied to EMA measurements of the upper and lower lip y-coordinates, and the y-coordinates of the tongue tip and tongue blade. Out of the 50 sentences initially considered, 9 (sentences 5, 8, 9, 11, 25, 29, 30, 37 and 48) were excluded from further evaluation due to conflicts in the number of voiced and unvoiced segments extracted from the different utterances.

Fig. 2 shows temporal alignment results for the sentence "bright sunshine shimmers on the ocean". Note how the EMA measurements are put into correspondence after audio-based temporal alignment. The accuracy of temporal alignment accuracy was measured using the mean absolute delay between corresponding temporal landmarks in the reference and test EMA sentences, before and after temporal alignment, with perfect alignment producing a mean absolute delay of zero. Phonetically speaking, it is safe to assume that the *instants* of maximal and minimal lip aperture (though not the *magnitude* of this aperture) depend primarily on the content of the uttered sentence and are speaker independent. Thus, local maxima and minima in the difference between upper and lower



(a) Before alignment (b) After alignment

Fig. 2. Audio-based temporal alignment of sentence "bright sunshine shimmers on the ocean" between the 4 speakers (the first speaker is the reference) applied to audio and EMA data.

lip y-coordinates, corresponding to instants of maximal and minimal lip aperture, were automatically extracted from the EMA data, before and after temporal alignment, to be used as temporal landmarks, as illustrated in Fig. 3.

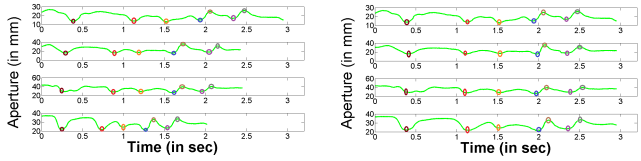


Fig. 3. Instants of minimal and maximal lip aperture (circles) before and after audio-based temporal alignment for the sentence "alimony harms a divorced man's wealth". Temporal alignment places the landmarks into correspondence.

Figure 4 shows a box plot of the mean absolute delays between corresponding temporal landmarks the reference and test EMA sentences, before and after temporal alignment. The mis-alignment of the landmarks in the original EMA data (prior to temporal alignement) was highly variable, with a median mean absolute delay of 121.2 ms (60.6 EMA samples) between landmarks in the reference and test signals. Temporal alignment reduces this median mean absolute delay to 34.8 ms (17.4 EMA samples), with 75% of the sentences registered within an average of 47.4 ms (23.7 EMA samples) of their respective references. Considering that the temporal landmarks were independently detected in the original and time aligned EMA data (which underwent additional interpolation operations), the correspondence between the temporal landmarks may not be exact, which may account for some of the residual inaccuracy. Nonetheless, the results demonstrate that the proposed audio-based temporal alignment procedure successfully accounts for time varying fluctuations in speech rate and speaker differences.

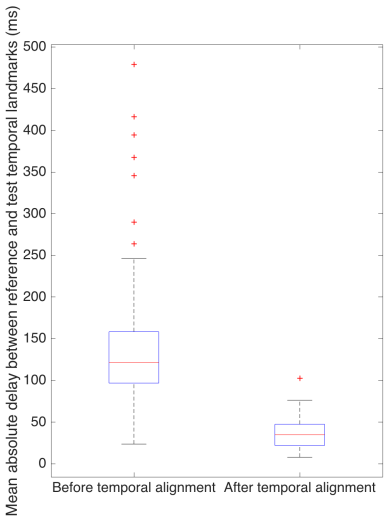


Fig. 4. Box plot showing the distribution of the mean absolute delays between corresponding temporal landmarks (maxima and minima in lip aperture) within a sentence in the reference and test EMA data, before and after temporal alignment.

4. CONCLUSIONS

This paper provides the first direct quantitative validation of audio-based temporal alignment of speech signals for the purpose of comparing speech processes between speakers. It also introduces an efficient divide-and-warp temporal alignment strategy that performs DTW alignment independently on matching voiced and unvoiced segments of audio speech data. Results show that the proposed approach aligns speech signals in a phonetically meaningful way between speakers. However, conflicts in the number of voiced and unvoiced segments between multiple utterances of the same sentence occasionally preclude further processing. Thus, future work will aim at improving the robustness of the audio segmentation process, opening the way towards investigation of audio-based temporal alignment between speakers for the study of speech impairments, a more challenging problem.

5. REFERENCES

[1] C. Qin, M. Á. Carreira-Perpiñán, K. Richmond, A. Wrench, and S. Renals, "Predicting tongue shapes from a few landmark locations," in *INTERSPEECH*, 2008, pp. 2306–2309.

[2] L. Ménard, J. Aubin, M. Thibeault, and G. Richard, "Measuring tongue shapes and positions with ultrasound imaging: a validation experiment using an articulatory model," *Folia Phoniatrica et Logopaedica*, vol. 64, pp. 64–72, 2012.

[3] M. Stone, J. M. Langguth, J. Woo, H. Chen, and J. L. Prince, "Tongue motion patterns in post-glossectomy and typical speakers: A principal components analysis," *Journal of Speech, Language and Hearing Research*, vol. 57, pp. 707–717, 2014.

[4] J. Woo et al., "Determining functional units of tongue motion via graph-regularized sparse non-negative matrix factorization," in *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2014.

[5] Vikramjit Mitra, Hosung Nam, Carol Y. Espy-Wilson, Elliot Saltzman, and Louis Goldstein, "Articulatory information for noise robust speech recognition," *IEEE Transactinos on Audio, Speech and Language Processing*, vol. 19, no. 7, 2011.

[6] L. Lancia and M. Tiede, "A survey of methods for analysis of the temporal evolution of speech articulator trajectories," in *Speech Planning and Dynamics*, S. Fuchs et al., Eds. Peter Lang Publishers, 2012.

[7] J. C. Lucero, K. G. Munhall, V. L. Gracco, and J. O. Ramsay, "On the registration of time and the patterning of speech movements," *Journal of Speech, Language and Hearing Research*, vol. 40, pp. 1111–1117, 1997.

[8] M. Li, C. Kambhamettu, and M. Stone, "Tongue motion averaging from contour sequences," *Clinical Linguistics and Phonetics*, vol. 19, no. 6-7, pp. 515–528, 2005.

[9] J. Kim, A. Lammert, P. Ghosh, and S. S. Narayanan, "Spatial and temporal alignment of multimodal human speech production data: real time imaging, flesh point tracking and audio," in *IEEE International Conference on Acoustics Speech and Signal Processing*, 2013, pp. 3637–3641.

[10] H. Strik and L. Boves, "A dynamic programming algorithm for time-aligning and averaging physiological signals related to speech," *Journal of Phonetics*, vol. 19, pp. 367–378, 1991.

[11] C. Yang and M. Stone, "Dynamic programming method for temporal registration of three-dimensional tongue surface motion from multiple utterances," *Speech Communication*, vol. 38, pp. 201–209, 2002.

[12] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, Feb 1978.

[13] E. J. Keogh and M. J. Pazzani, "Derivative dynamic time warping," in *SIAM Conference on Data Mining*, 2001.

[14] J. Listgarten, R. M. Neal, S. T. Roweis, and A. Emili, "Multiple alignment of continuous time series," in *Advances in Neural Information Processing Systems*, 2005, pp. 817–824.

[15] F. Zhou and F. de la Torre, "Canonical time warping of human behavior," in *Advances in Neural Information Processing Systems*, 2009.

[16] C. Myers, L. Rabiner, and A. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 6, pp. 623–635, 1980.

[17] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and Information Systems*, vol. 7, no. 3, pp. 358–386, 2005.

[18] A. A. Wrench, "A multi-channel/multi-speaker articulatory database for continuous speech recognition research.," in *Phonus 5*, 2000.