On the Automatic Validation of Speech Alignment

Georgios Athanasopoulos, Benoît Macq
ICTEAM-ELEN - Université catholique de Louvain, Belgium

***Abstract*—The alignment of two utterances is the basis of many speech processing applications. The acoustic user interface of such applications should be capable of detecting insufficient alignment results and identifying the responsible input utterances. In this paper, we discuss the automatic validation of speech alignment and propose two new validation algorithms. The first method relies on locating and matching the syllable nuclei of the aligned utterances. The second method performs syllable-level comparison of the speech signal envelopes in accordance to the alignment time-warping path. Experimental results show that the proposed algorithms perform consistently well and can be effectively applied for the validation of different speech alignment methods.**

***Index Terms*—speech alignment, HMM-based forced alignment, dynamic time warping, alignment assessment**

I. Introduction

Speech alignment is fundamental in various speech processing applications such as automatic dialog replacement [1], voice conversion and transplantation [2], pronunciation evaluation in second language learning [3], assessment and processing of speech disorders [4], [5]. The alignment of the user's speech input (source utterance) with a speech model (reference utterance) concerns the identification of the relative timing differences between the corresponding speech signals using a timing analysis technique. In real-world systems, various factors such as the user's pronunciation or the presence of background noise can influence the alignment result.

From an acoustic user interface point of view, it is desirable that only speech inputs of adequate quality are processed. Depending on the application, criteria such as the SNR, distortion analysis (e.g., due to signal clipping), or intelligibility measures [6] can be applied for the assessment of the input speech signal. In applications where a specific user utterance is expected, an invalid input would result in unpredictable processing outcome. It is, therefore, important that the application is capable of detecting and discarding wrong inputs. This task is not always straightforward. For example, the interpretation of the user's input using speech recognition technology is not suitable for all types of applications.

In speech alignment, a good correspondence between the source and the reference utterance is essential to allow for reliable results. The validation of the alignment output is a critical task which could feedback the user input verification process. To the best of our knowledge, no prior research attention has been drawn in the area of the automatic validation of speech alignment. In this paper, we address this requirement by investigating complementary verification strategies. Our goal is the overall evaluation of the alignment for determining whether it has been successfully completed and if its outcome can be used by subsequent speech processing components. To this purpose, two computationally simple methods are introduced. Both methods are designed to be independent of the alignment algorithm. In the first method, we propose locating and matching the syllable nuclei in the aligned source and reference utterances. The second method relies on the comparison of the speech signal envelopes in each syllable in accordance to the alignment time-warping path.

The paper is organized as follows. In Section II, we provide an overview of speech alignment methodologies and further motivate the importance of validating their outcomes. The proposed strategies for the automatic validation of speech alignment are detailed in Section III. In Section IV, we present and discuss our experimental results. Finally, Section V concludes this paper and provides an outlook.

II. Background

Two major approaches exist for estimating the timing relationship (time-warping path) between two speech signals: the Hidden Markov Model (HMM)-based phonetic alignment and the Dynamic Time Warping (DTW). Figure 1 shows an example of a time-warping path. In the HMM-based alignment, the timing relationship is estimated by making use of speech recognition paradigms. An acoustic model is used for providing an overall statistical representation of the distinct sounds of a language that are not specific to one speaker or speaking style. Essentially, the alignment between two speech signals is achieved via their individual phonetic segmentation. Using a pre-trained acoustic model and the phonetic transcription of the utterance, the alignment task reduces to determining the phoneme boundaries in both source and reference acoustic recordings, as shown in Figure 2. The complete time-warping path can be inferred from the phoneme boundaries through interpolation. An overview of HMM-based alignment methods can be found in [7]. HMM-based alignment is known to perform well in various applications. An inherit drawback is that it is language specific and its performance degrades for different speaking styles (e.g., pronunciation variations, expressive speech, dialects). Implementing an HMM-based alignment method typically results in relatively more complex systems and its performance depends on the quality of the acoustic model.

The DTW is an earlier, well-known technique for computing an alignment time-warping path between two utterances without requiring a phonetic transcription. In speech processing, it

Parts of the research in this paper were performed in the context of the FEDER project UserMEDIA (#501907-379156) co-funded by the European Union and the Région Wallonne.

ISBN 978-90-827970-1-5 © EURASIP 2018

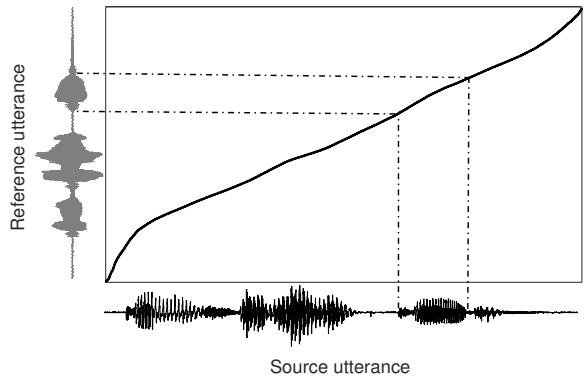

Fig. 1. Time-warping path between the source and reference utterances of speakers with different voice characteristics.

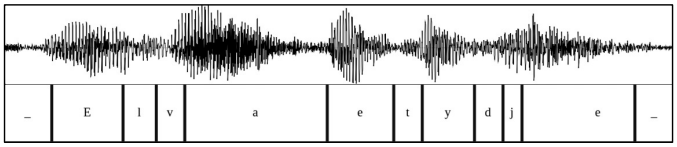

Fig. 2. HMM-based phonetic segmentation of the French utterance "Elle va étudier".

was initially applied for the comparison of speech patterns in the context of speech recognition [8]. As the DTW operates on speech signal level, it is language independent and does not require training an acoustic model. Essentially, DTW algorithms aim at measuring the similarity between two temporal sequences which may vary in speed. The two sequences are aligned by warping the time axis of their feature vectors (e.g., the Mel-frequency cepstral coefficients (MFCC)) iteratively until an optimal match between the two sequences is reached in terms of similarity. The main concept of the algorithm, recent improvements and related alignment techniques can be found in [9].

An important challenge for both methods is how to evaluate the alignment outcome. The output likelihood computed during the HMM-based alignment of a phoneme contain information about how close this uttered phoneme was to the corresponding model. Intuitively, the likelihood score could be used for assessing the alignment's quality. In practice, the likelihoods depend on the quality of the acoustic model (e.g., size of training corpus, coverage of different speaking styles and pronunciations). According to our observations, successfully aligned utterances may still result in relative low likelihoods due to a particular speaking style or accent. More importantly, because the phoneme boundaries are selected so that a maximum likelihood criterion is satisfied, misaligned utterances very often preserve high likelihood scores. Hence, the likelihood scores, either in phoneme level (local) or for the entire utterance (global), do not always provide a reliable measure of the alignment's overall quality. In DTW, a measure of goodness of the alignment can be obtained via the global accumulated DTW distance, which is computed as the weighted sum of feature vector distances along the time-warping path. In our experience, the accumulated distance is pronunciation and voice type dependent, e.g., due to spectral differences in male/female voices. Hence, its use as a single criterion of the alignment's quality can be ambiguous. In the following section, we examine how these uncertainties can be resolved by complementary methods specifically designed for validating the alignment results.

III. PROPOSED METHODS

A. Syllable Nuclei Matching

The intention of the SNM method is to assess the alignment outcome by matching the position of syllable nuclei in the source and reference utterances. The syllable nucleus is typically a vowel or, in some languages, a syllabic consonant. This work considers the French language where the syllable nucleus is always a vowel and its presence is obligatory. Hence, we can detect the syllable nuclei via a vowel detection algorithm, such as the Low Frequency Modulated Energy (LFME) proposed in [10]. This method is based on the rationale that vowels possess a considerable amount of energy in the low frequency region.

The detection of the syllable nuclei can be determined from the peaks in the energy contour of the speech signal. The energy contour $e_i(m)$ of a signal $x(n)$ that corresponds to a frequency bin region $[k_1^i, k_2^i]$, can be computed as

$$e_i(m) = \frac{1}{L_{fft} L_{win}} \sum_{k=k_1^i}^{k_2^i} w_k |X_k(m)|^2, \quad i = 0 \dots N-1 \quad (1)$$

where $X_k(m)$ is the STFT of $x(n)$ with k and m the frequency and frame indexes, respectively. L_{fft} represents the length of the FFT, L_{win} is the length of the frame in samples and N is the number of sub-bands. Here, w_k represents the pre-emphasis weights for compensating the energy declination due to the speech spectral slope (commonly $6dB$/octave). The LFME trajectory is calculated from the above as

$$LFME(m) = e_0^2(m) \sum_{i=1}^{N-1} e_i(m) \quad (2)$$

where e_0 is the energy in the lowest frequency band, typically $[300Hz, 800Hz]$. The syllable nuclei are estimated via peak pruning based on amplitude and temporal criteria as detailed in [10]. The speech signals are normalized beforehand according to their RMS energies, allowing common thresholds to be applied for all utterances. Figure 3 shows the normalized LFME trajectory for the same phrase as in Figure 2, and the correspondence of its peaks to the phonetic transcription of the utterance's syllable nuclei.

The following summarizes the proposed matching approach. Starting from the estimated syllable nuclei positions in the source and reference utterances, we cluster them according to the syllable where they belong. With regards to the reference utterance, this task can be performed and, if necessary, be manually adjusted in advance. The boundaries of syllables for both source and reference are considered known or can be inferred from the speech waveform, as it will be discussed

ISBN 978-90-827970-1-5 © EURASIP 2018

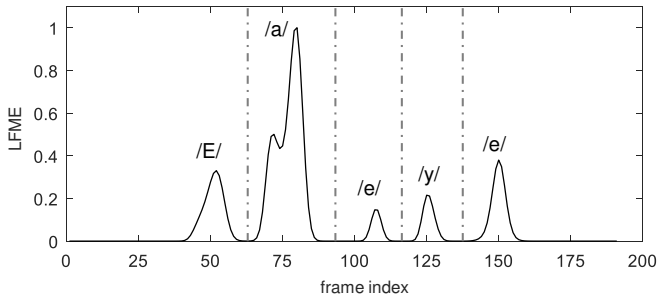

Fig. 3. The LFME and the corresponding syllable nuclei.

latter on. Ideally, we would expect one nucleus per syllable. In reality, additional nuclei candidates or no nucleus at all might be assigned to a syllable due to imperfect estimation, mispronunciations, etc. Supplementary nuclei candidates in a given syllable can also occur due to the presence of pronounced plosives, voiced consonants, or semi-vowels. In a first step, we compare the number of nuclei candidates in each syllable between source and reference utterances. The objective is to identify syllables where either no nucleus is assigned, or the difference in the number of nuclei candidates exceeds a threshold $t_{ns} \in \mathbb{Z}_{\geq 0}$. If the number of syllables identified in the previous step is higher than a threshold t_{nn}, the underlying nuclei mismatch indicates that the two utterances do not sufficiently correspond with each other and the alignment result should be discarded. In order not to favor shorter utterances, we define the threshold t_{nn} proportionally to the number of syllables as $t_{nn} = \max(t_{nn}^{min}, \lfloor \alpha \; n_{syl} \rceil)$, where $t_{nn}^{min} \in \mathbb{Z}_{\geq 0}$ is the minimum threshold value, parameter $\alpha \in [0, 1]$ controls the threshold value, n_{syl} is the number of syllables, and $\lfloor \; \rceil$ denotes the nearest integer function.

When the number of identified syllables is below t_{nn}, we further evaluate the alignment result through the Euclidean distance of candidate nuclei positions between source and reference syllables. In this computation, candidate nuclei positions of the source utterance are projected to the reference through the alignment's time-warping path. Syllables with no assignments are excluded from this computation. For syllables with unequal number of assignments, only those assignments that result in minimum distances are considered. The distance measure is either evaluated in syllable level for the identification of local alignment mismatches, or the mean distance is used for the validation of the overall alignment. If the distance exceeds a threshold $t_{nd} \in \Re_{\geq 0}$, this is a sign of inaccuracies in the alignment and its result can be discarded.

B. Syllable Envelopes Comparison

The key idea of the SEC method is to perform a syllable by syllable comparison of the speech signal envelopes for validating the alignment outcome. The rational behind this choice is that speech signal envelopes convey various temporal cues related to aspects such as articulation, voicing, vowel identity, or syllabification [11]. Different methods for the extraction of a signal's envelope exist in the literature [12]. We compute the speech envelope using an efficient three-step approach proposed in [13]. In this method, the amount of smoothing can be directly influenced through the analysis window length. The obtained smoothed envelope curve is segmented according to the syllable boundaries. Similarly to the SNM method, we assume that the syllable boundaries of both source and reference are known or can be inferred from the speech signal. The envelope curve of each source syllable is time-scaled according to the time-warping path. In order to account for amplitude differences between the signals, both source and reference envelopes are normalized for each syllable according to their RMS amplitude.

The comparison of the envelope curves is performed in terms of Fréchet distance. The Fréchet distance between two curves in a metric space is a measure of the similarity between two curves and it can be computed as described in [14]. This distance measure can be interpreted as the minimum length of a line that connects a point on each curve, and allows one to traverse both curves from start to finish. Syllable pairs whose distance measure exceeds a threshold $t_{es} \in \Re_{\geq 0}$ are considered as potential misalignment candidates. If the total number of misalignment candidates is higher than a threshold t_{en}, we are probably dealing with two utterances that exhibit different temporal cues and the alignment result should be discarded. Similarly to before, the threshold t_{en} is defined as $t_{en} = \max(t_{en}^{min}, \lfloor \beta \; n_{syl} \rceil)$, with $t_{en}^{min} \in \mathbb{Z}_{\geq 0}$ the minimum threshold value and parameter $\beta \in [0, 1]$ adjusting the threshold value proportionally to the number of syllables n_{syl}.

IV. RESULTS & DISCUSSION

Our evaluation is based on a French speech corpus consisting of 127 utterances (up to 10 syllables long) by five different speakers, two native (one male, one female) and three non-native (one male, two female). Every speaker uttered the complete set of phrases. The utterances are properly selected in order to offer a good phonetic coverage of the French language. The sampling rate is $16kHz$ and the recordings took place in clean conditions. A training corpus was defined as a subset of 60 utterances. The remaining utterances form the test corpus. One of the native speaker sets was used as the alignment's reference utterances and the remaining speaker sets as the source utterances. The optimization of the parameters related to LFME algorithm was performed similarly to [10]. The selection of parameters for each proposed method, as well as for the combination of the two, was performed using the training corpus. Pairs of training corpus utterances were given as input to the speech alignment. The parameter values were iteratively optimized so that the following criterion was met: all alignment results from matching source and reference utterances to be successfully validated, while maximizing the number of misalignment detections due to non-matching utterances.

Both HMM-based and DTW alignment methods were considered during the evaluation. Our implementation of the former is similar to that of [15]. Regarding the DTW, it was

ISBN 978-90-827970-1-5 © EURASIP 2018

developed in the scope of [3] and incorporates improvements proposed in [1], [16]. In this evaluation, we assume that the syllable boundaries of all utterances are known. This allows us to decouple the methods' performance from the syllable estimation task. To this purpose, we obtain the corresponding syllables structure along with the phonetic transcriptions required for the HMM-based alignment from the orthographic transcription through eLite-HTS [17]. In the case of DTW, the boundaries are inferred from the reference by applying the alignment time-warping path (see dashed lines in Figure 1 for an example). In the absence of phonetic segmentation for the reference utterance, the closely spaced LFME peaks can be clustered according to a distance criterion. Syllable boundaries can then be roughly approximated as the equidistant locations between clusters, as in the example of Figure 3 (vertical dashed lines). Although informal testing shows that this approach is sufficient in our context, interested readers may refer to elaborated algorithms that attempt to segment a speech waveform into syllables [18].

We performed an objective assessment of SNM and SEC, as well as a serial combination of the two (SNM-SEC). Our intention was to evaluate the capacity of the proposed methods in validating the global alignment outcome. Low level (i.e., phoneme by phoneme) evaluation of the alignment's quality was not in our scope. The performance was assessed in terms of misalignment detection rate, defined as the percentage of accurate detections over the total misaligned results. First, global misaligned results were triggered by attempting to align mismatching input utterances. All pair combinations of non-corresponding source and reference utterances of the test corpus were used. In a second step, misalignment was induced locally by altering the source utterances. Alterations were performed in syllable level. Three types were considered: deletion, insertion and replacement of syllables. All alterations were performed randomly, affecting up to three syllables per utterance.

Consistent evaluation results were observed for both alignment methods, with small variations being due to the different impact the performed alternations have on the alignment result of each method. Table I summarizes the aggregated detection rates of both alignment methods. High detection rates are observed for the global misaligned utterances with SEC outperforming SNM, while SNM-SEC achieves a detection rate up to 89.1%. This underlines the capacity of the proposed methods in identifying invalid input utterances. As far as the local misalignments are concerned, SNM results in somewhat higher detection rates than SEC. However, when the two methods are combined, a better performance is achieved for all investigated misalignment types. Our results confirm that the detection of local misalignments is a more challenging task especially when, as it was chosen in this evaluation, only a small number of syllables is affected. Our analysis indicates that misdetections are mainly due to the preservation of nuclei locations and envelope similarities in the altered utterances. This observation could motivate in the future the choice of alternative validation criteria specific to the assessment of local alignment results.

TABLE I
DETECTION RATES FOR VARIOUS MISALIGNMENT TYPES.

Misalignment Type	SNM	SEC	SNM-SEC
Global Mismatch	70.9%	79.5%	89.1%
Local Insertions	41.8%	37.3%	55.5%
Local Deletions	55.5%	38.2%	64.5%
Local Replacements	50.1%	46.4%	73.6%

Finally, a subjective experience evaluation was conducted in the context of a user experience testing related to [3]. Ten non-native French speakers participated in the test. Analysis of the users post-experience interviews indicates that the application was perceived as more well-performing when it was capable of detecting invalid input utterances. This highlights the importance of the automatic speech alignment validation from an acoustic user interface point of view. Moreover, a post-analysis of video recordings containing each user's interaction shows that the selection of the algorithm's parameters plays a very important role in real-world scenarios. Parameters resulting in more frequent rejections of valid input utterances, e.g., due to pronunciation variations but also due to presence of ambient noise or reverberation, were found more likely to decrease the user experience satisfaction (e.g., preventing some users from completing certain tasks). Overall, the experience evaluation confirmed that the proposed methods can be efficiently applied in real-time applications and verified the capacity of the proposed methods in the automatic validation of speech alignment in real-world settings.

V. CONCLUSION

In this paper, we investigated the importance of automatically validating the results of speech alignment algorithms. After providing an overview of speech alignment methodologies, we introduced two novel validation methods. The SNM method relies on locating and matching the syllable nuclei of the aligned utterances, while the SEC method performs syllable-level comparison of the speech signal envelopes in accordance to the alignment time-warping path.

Our experimental evaluation and subjective experience testing in real-world settings confirmed the potential of both proposed methods (as well as of their combination) in validating the global alignment outcome, and especially in identifying invalid input utterances. The detection of insufficient local alignment results appears to be a more challenging task, especially when only a small number of syllables is affected. Therefore, the results of this research work could motivate future developments specific to local alignment validation (e.g., due to variations in pronunciation). Finally, future work could also investigate the automatic validation of phonetic alignment where no reference utterance is available and, hence, the validation needs to be performed against the phonetic transcription of the utterance.

ISBN 978-90-827970-1-5

REFERENCES

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

[2] J. Nirmal, M. Zaveri, S. Patnaik, and P. Kachare, "Novel Approach of MFCC Based Alignment and WD-residual Modification for Voice Conversion Using RBF," *Neurocomputing*, vol. 237, no. C, pp. 39–49, 2017.

[3] G. Athanasopoulos, K. Hagihara, A. Cierro, R. Guérit, J. Chatelain, C. Lucas, and B. Macq, "3D Immersive Karaoke for the Learning of Foreign Language Pronunciation," in *Proceedings of the International Conference on 3D Immersion*, 2017.

[4] D. Le and E. M. Provost, "Modeling Pronunciation, Rhythm, and Intonation for Automatic Assessment of Speech Quality in Aphasia Rehabilitation," in *Proceedings of INTERSPEECH*, 2014.

[5] F. Rudzicz, "Adjusting Dysarthric Speech Signals to be more Intelligible," *Computer Speech & Language*, vol. 27, no. 6, pp. 1163–1177, 2013.

[6] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Evaluation of Objective Measures for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *J. Acoust. Soc. Am.*, vol. 130, no. 5, 2011.

[7] S. Brognaux, "Expressive Speech Synthesis: Research and System Design with Hidden Markov Models," Ph.D. dissertation, Université catholique de Louvain (UCL) - Université de Mons (UMONS), 2015.

[8] 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, 1978.

[9] M. Müller, *Information Retrieval for Music and Motion*. Springer, Berlin, Heidelberg, 2007, ch. Dynamic Time Warping, pp. 69–84.

[10] T. Dekens, H. Martens, G. V. Nuffelen, M. D. Bodt, and W. Verhelst, "Speech Rate Determination by Vowel Detection on the Modulated Energy Envelope," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2014.

[11] S. Rosen, "Temporal Information in Speech: Acoustic, Auditory and Linguistic Aspects," *Philosophical Transactions: Biological Sciences*, vol. 336, no. 1278, pp. 367–373, 1992.

[12] W. Biesmans, J. Vanthornhout, J. Wouters, M. Moonen, T. Francart, and A. Bertrand, "Comparison of Speech Envelope Extraction Methods for EEG-based Auditory Attention Detection in a Cocktail Party Scenario," in *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015.

[13] C. Jarne, "Simple Empirical Algorithm to Obtain Signal Envelope in three Steps," *CoRR*, 2017.

[14] P. K. Agarwal, R. B. Avraham, H. Kaplan, and M. Sharir, "Computing the Discrete Fréchet Distance in Subquadratic Time," *CoRR*, 2012.

[15] S. Brognaux and T. Drugman, "HMM-based Speech Segmentation: Improvements of Fully Automatic Approaches," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 1, pp. 5–15, 2016.

[16] P. Soens and W. Verhelst, "An Iterative Bilinear Frequency Warping Approach to Robust Speaker-Independent Time Synchronization," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2012.

[17] S. Roekhaut, S. Brognaux, R. Beaufort, and T. Dutoit, "eLite-HTS: A NLP Tool for French HMM-Based Speech Synthesis," in *Proceedings of INTERSPEECH Show&Tell*, 2014.

[18] N. Obin, F. Lamare, and A. Roebel, "Syll-O-Matic: an Adaptive Time-Frequency Representation for the Automatic Segmentation of Speech into Syllables," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

ISBN 978-90-827970-1-5 © EURASIP 2018