Automatic Long Audio Alignment and Confidence Scoring for Conversational Arabic Speech

Mohamed Elmahdy[⋆], **Mark Hasegawa-Johnson**[†], **Eiman Mustafawi**[⋆]

[⋆]Qatar University, Doha, Qatar
[†]University of Illinois at Urbana-Champaign, USA
melmahdy@ieee.org, jhasegawg@illinois.edu, eimanmust@qu.edu.qa

Abstract

In this paper, a framework for long audio alignment for conversational Arabic speech is proposed. Accurate alignments help in many speech processing tasks such as audio indexing, speech recognizer acoustic model (AM) training, audio summarizing and retrieving, etc. We have collected more than 1,400 hours of conversational Arabic besides the corresponding human generated non-aligned transcriptions. Automatic audio segmentation is performed using a split and merge approach. A biased language model (LM) is trained using the corresponding text after a pre-processing stage. Because of the dominance of non-standard Arabic in conversational speech, a graphemic pronunciation model (PM) is utilized. The proposed alignment approach is performed in two passes. Firstly, a generic standard Arabic AM is used along with the biased LM and the graphemic PM in a fast speech recognition pass. In a second pass, a more restricted LM is generated for each audio segment, and unsupervised acoustic model adaptation is applied. The recognizer output is aligned with the processed transcriptions using Levenshtein algorithm. The proposed approach resulted in an initial alignment accuracy of 97.8-99.0% depending on the amount of disfluencies. A confidence scoring metric is proposed to accept/reject aligner output. Using confidence scores, it was possible to reject the majority of mis-aligned segments resulting in alignment accuracy of 99.0-99.8% depending on the speech domain and the amount of disfluencies.

Keywords: conversational Arabic, audio alignment, speech processing

1. Introduction

In many cases, audio speech data is available with the corresponding human generated transcriptions; however they are not aligned (or synchronized) as in the case of meetings, lectures, podcasts, audio books, etc. Long Audio Alignment is a known problem in speech processing in which the goal is to automatically align a long audio input with the corresponding transcriptions. The problem usually deals with very long audio that can exceed one hour length. Accurate long audio alignments can help in many speech processing tasks such as audio indexing, speech recognizer acoustic model training, audio summarizing and retrieving, etc. Manual alignment for large amounts of speech data could be very costly and inefficient.

Our goal, in this research, is to develop an approach that is capable of automatically align long speech audio files, in particular for conversational Arabic. Most of prior work in long audio alignment has focused on English language as in (Hazen, 2006; Liu et al., 2003; Moreno et al., 1998) and to the best of our knowledge there is no prior work done for conversational Arabic long audio alignment. Conversational Arabic speech is mostly spontaneous with the dominance of dialectal Arabic that differs significantly from standard Arabic.

In (Moreno et al., 1998), a recursive long audio alignment approach was proposed . The approach is based on Automatic Speech Recognition (ASR) and evaluated on English speech. A biased language model (LM) is prepared using the corresponding text to the audio file. ASR is applied on the entire audio file. Speech recognition results are aligned with the original text. ASR is then reapplied on smaller segments with a more restricted LM between Anchors. Anchors are common strings between ASR results and the original transcriptions. In (Hazen, 2006), some preliminary analysis of manual transcriptions is provided which show that there is a significant difference between human generated transcripts and what is actually being uttered in the audio file. Thus, a long alignment approach is designed in such a way to detect and correct errors in the initial manually generated transcription.

Since disfluencies occur frequently in spontaneous speech, in (Liu et al., 2003), a number of knowledge sources for disfluency identification were investigated. An automatic alignment system was proposed that was based on acoustic-prosodic features, word-based, and repetition pattern language models.

In this work, an ASR-based long audio aligner for conversational Arabic speech is proposed. Unlike prior work that applies ASR on the whole long audio file, our alignment approach starts first with automatic audio segmentation to split the audio file into small segments. This, in turn, speeds up ASR decoding in addition to improving alignment accuracy. Unlike the work of (Moreno et al., 1998), language model restriction is not only applied between anchors, language model restriction is applied on all segments regardless of anchor rate. Since we are dealing with conversational Arabic with a significant dominance of dialectal Arabic, we propose grapheme-based acoustic modeling in which all short vowels and geminations are implicitly modeled in the AM. Graphemic modeling was introduced for Modern Standard Arabic in (Billa et al., 2002), and applied on dialectal Arabic in (Elmahdy et al., 2011; Elmahdy et al., 2010). Furthermore, a segment-based confidence scoring metric is proposed to accept/reject alignment results.

The remainder of this paper is organized as follows: Section 2 presents the collection of data sets used in this research. Section 3 describes the automatic segmentation approach. The proposed long audio alignment approach is described in Section 4. Experimental results are discussed in Section 5. Section 6 concludes this study.

2. Data Collection

Around 2,150 conversational episodes (or podcasts) have been colleced from Al-Jazeerah channel with an overall length of more than 1,400 hours. Episodes vary in length from 20-50 minutes. All episodes have been downloaded from YouTube with the highest available audio quality. The recordings are spanning from year 2008 to 2012. Audio tracks have been extracted, converted to monaural audio, and resampled to 16 kHz. All corresponding raw text has been downloaded from Al-Jazeerah website. A rule-based pre-processing stage is applied on the raw transcription files to remove titles, headings, images, punctuation marks, etc.

In order to evaluate alignment accuracy, two evaluation sets were prepared. The first set, *conv-eval-set*, is for media conversational Arabic that is characterized with significant speech disfluencies. The second set, *read-eval-set*, is for clean read speech that is characterized with very little speech disfluencies. The *conv-eval-set* was prepared by manually aligning five hours from Al-Jazeerah conversational podcasts, which is the same set used previously in (Elmahdy et al., 2013). The *read-eval-set* consists of five hours from the LDC Arabic Broadcast News corpus (Maamouri et al., 2006a; Maamouri et al., 2006b) after the elimination of all synchronization data, transcribed disfluencies and filler tags.

3. Automatic Segmentation

Unlike prior work that applies ASR on the whole long audio file, our proposed alignment approach starts with automatic segmentation to segment the audio file into small segments. Applying ASR on small segments can speed up decoding as well as improving alignment accuracy.

In this section, an automatic split and merge segmentation approach is proposed. Segmentation is applied to segment audio files into small segments of a customized average length of 5-10 sec. For each episode, the energy is computed at each sample with a frame window of 512 samples. Then the mean energy for the whole episode is estimated. Silence threshold is chosen to be a customizable fraction of the mean energy. Empirically, a fraction of 20% percent was found reasonable. Silence duration is configured to have a minimum of 350 ms. Minimum silence duration is important to avoid segmentation at geminated consonants or low energy sounds e.g. /s/. Thus, consecutive frames of duration more than 350 ms and with energy below the silence threshold are assumed to be a silence period. Segmentation is then applied at the center of each silence period. This results in a large number of short segments. Finally, consecutive short segments are merged together as long as the merged segment does not exceed 5-10 sec.

4. Long Audio Alignment

4.1. ASR System Description

The ASR system is a GMM-HMM architecture based on the CMU Sphinx engine (Huggins-Daines et al., 2006). Acoustic models are all fully continuous density context-dependent tri-phones with 3 states per HMM trained with maximum likelihood estimation (MLE). The feature vector consists of 39-dimensional MFCC coefficients. During acoustic model training, linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT) are applied to reduce the dimensionality to 29 dimensions, which improves accuracy as well as recognition speed. Decoding is performed in multi-pass, a fast forward Viterbi search using a lexical tree, followed by a flat-lexicon search and a best-path search over the resulting word lattice.

4.2. First Alignment Pass

For each episode, a biased bi-gram LM model is trained using the corresponding raw text with Witten Bell smoothing applied. Lexicon was restricted to cover only words in the transcriptions of the current episode. Non-Arabic words are excluded from the lexicon and replaced with a garbage model.

In conversational Arabic, speakers tend to use dialectal Arabic rather than standard Arabic. Moreover, they tend to make more grammatical mistakes as changing case endings (e.g. using /u/ rather than /i/). Because of the diglossic nature of Arabic varieties (Ferguson, 1959), it is hard to estimate all pronunciation variants for dialectal words and all possible mis-pronunciations. Thus, a graphemic pronunciation model is used where the pronunciation is simply the word letters rather than the actual pronunciation. In this case, there is one pronunciation for each given word. A graphemic acoustic model is trained with more than 60 hours of standard Arabic speech data. In graphemic modeling, short vowels and geminations are assumed to be implicitly modeled in the acoustic model. A quick overview about Arabic ASR can be found at (Elmahdy et al., 2009).

In the first ASR decoding pass, all segments for the current episode are decoded using the biased LM, restricted PM, and the standard Arabic AM. In the first pass, relatively fast AM is used that consists of a mixture of 8 Gaussians per state. The ASR output is aligned with the processed transcriptions using Levenshtein distance algorithm. This way we can ensure error recovery where mis-alignment of a certain segment does not affect the alignment of later segments. For more illustration, in Figure 1, an example for aligning two consecutive segments is shown. The first row shows the results of the speech recognizer. The second row shows final aligned transcriptions.

Anchor rate is determined by the number of correct words in ASR results, which are matched with the original text, divided by the total number of words in the original text. Anchor rate was found to be directly proportional with final alignment accuracy.

4.3. Second Alignment Pass

4.3.1. Unsupervised AM adaptation

Anchor rate was found to decrease in noisy segments and with dialectal Arabic segments where the acoustic features

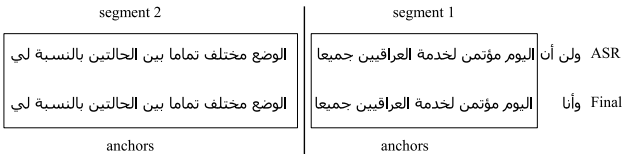

Figure 1: Speech recognizer's output aligned with corresponding transcription text.

differs from the AM training data. It is also affected by the amount of speech disfluencies. In order to increase anchor rate and hence improving alignment accuracy, similarly to what was proposed in (Elmahdy et al., 2010), unsupervised AM adaptation is applied. Results from first alignment pass along with corresponding audio data have been used in unsupervised acoustic model adaptation using Maximum Likelihood Linear Regression (MLLR) (Leggetter and Woodland, 1995). In the second pass, a better AM is used with a mixture of 16 Gaussian densities per state.

4.3.2. Restricted Language Modeling

It was noticed that most of mis-alignmets in the first pass occurs at segment boundaries with only one or two mis-aligned words. For this reason, it is more efficient to use a more restricted LM rather than a LM trained with the whole episode's text. Alignment results from the first pass are used to generate small restricted LMs for each segment. Each restricted model is trained using aligned text of the current segment along with the alignment of the previous and following segments. Similarly to the first pass, the adapted AM along with the restricted LMs are used in batch decoding for all the segments. In Figure 2, a high level block diagram is shown to summarize the whole long audio alignment framework.

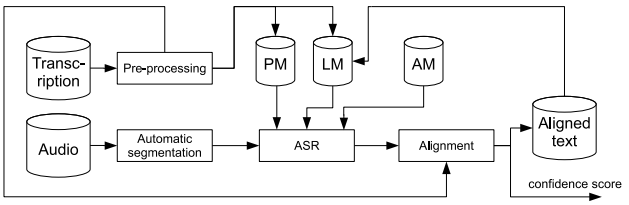

Figure 2: A Block diagram illustrating the proposed long audio alignment framework.

4.4. Alignment Results

We started by evaluating alignment accuracy on the conversational evaluation set *conv-eval-set* which is characterized with the presence of many speech disfluencies like: hesitations, truncated words, ah, etc. First alignment pass has resulted in an anchor rate of 84.6% and word alignment accuracy of 95.4% as shown in Table 1. In the second alignment pass, by using AM adaptation, alignment was significantly improved with anchor rate of 87.1% and word alignment accuracy of 96.2% as shown in Table 1. The significant increase in accuracy in the second pass is mainly interpreted to the mismatch between the acoustic model (standard Arabic) and the speech domain (conversational Arabic). Due to the adaptation applied in the second pass, the impact of this mismatch on ASR is decreased. By using restricted LM for each segment along with the adapted AM, anchor rate was significantly increased to 91.9% achieving word alignment accuracy of 97.7%.

It was noticed that the majority of mis-aligned words tends to be with relatively shorter words. This can interpret the slightly better alignment accuracy when calculating the accuracy in terms of characters rather than words as shown in Table 1.

Alignment pass	Anchor rate	Align. Accuracy	
		words	char.
pass 1	84.6%	95.4%	95.8%
pass 2 adapt.	87.1%	96.2%	96.3%
pass 2 adapt./LM restrict.	91.9%	97.7%	97.8%

Table 1: Alignment accuracy and anchor rate results on the *conv-eval-set*.

By applying the proposed alignment approach on the *read-eval-set* which is characterized with the presence of rare disfluencies. First alignment pass has resulted in an anchor rate of 91.3% and word alignment accuracy of 98.9% that is significantly better than the case of as the *conv-eval-set* as shown in Table 2. In the second alignment pass, by using AM adaptation, anchor rate was further improved to 93.9% as shown in Table 2, however, word alignment accuracy was not improved as in the case of the *conv-eval-set*. By using restricted LM for each segment along with the adapted AM, anchor rate was significantly increased to 96.4%, however, alignment accuracy was slightly increased to 99.0%.

Alignment pass	Anchor rate	Align. Accuracy	
		words	char.
pass 1	91.3%	98.9%	98.9%
pass 2 adapt.	93.9%	98.9%	98.9%
pass 2 adapt./LM restrict.	96.4%	99.0%	99.0%

Table 2: Alignment accuracy and anchor rate results on the *read-eval-set*.

5. Confidence Scoring

Most of the mis-alignment errors were found to be with segments having significant background noise, like: music, channel noise, cross-talk, etc., or significant speech disfluencies (truncated words, repeated words, hesitations, etc.). It should be noted that, in human generated transcripts which are associated with long audio files, disfluencies are rarely transcribed.

For some speech processing tasks like acoustic model training, it is required to eliminate mis-aligned segments and non-speech segments from the training data. Also, in a semi-automated alignment process, it would be more efficient to identify mis-aligned segments, so that they can be aligned manually.

So, in this section, a confidence scoring metric is proposed to accept/reject aligner output. We have found out that the anchor rate is highly correlated with the final alignment accuracy. The proposed confidence score for an aligned segment is the Levenshtein distance between the recognizer output and the aligned text, which is eventually the word anchor rate calculated for each segment.
Different threshold values were applied to filter out segments with low confidence score to check the improvement in alignment accuracy for the remaining segments. For the *conv-eval-set*, by considering segments with a confidence score greater than 20.0%, it was found that 5.2% of the total aligned segments was filtered out as shown in Table 3. By filtering out segments with confidence score less than 20%, word alignment accuracy is increased to 98.4% as shown in Table 3. By increasing the threshold value to filter out segments with anchor rate less than 40%, word alignment accuracy is further increased to 98.5% with 6.2% of the segments filtered out. By considering only segments with high confidence score greater than 90%, word alignment accuracy was significantly increased to 99.2% with 23.9% of segments filtered out as shown in Table 3.

Confidence score threshold	Segments filtered	Alignment accuracy
baseline (no thresh.)	0.0%	97.7%
>20.0%	5.2%	98.4%
>40.0%	6.2%	98.5%
>60.0%	8.7%	98.7%
>80.0%	14.9%	99.0%
>90.0%	23.9%	99.2%

Table 3: Alignment accuracy and filtered segments amount (%) on the *conv-eval-set* after confidence score thresholding.

For the *read-eval-set*, the initial alignment accuracy is considered high; however, we tried to filter out low confidence segments and to check how this affects alignment accuracy on the remaining data. By considering segments with a confidence score greater than 20.0%, it was found that only 2.5% of the total segments was filtered out as shown in Table 4, and word alignment accuracy was 99.4% for the remaining segments. By increasing the threshold value 40%, word alignment accuracy is slightly increased to 98.5% with 2.7% of the segments filtered out. By considering only segments with high confidence score greater than 90.0%, word alignment accuracy was increased to 99.8% with the rejectio of 10.0% of the audio segments as shown in Table 4.

6. Conclusions and Future Work

In this paper, a framework for conversational Arabic long audio alignment is proposed. Long audio files are automatically segmented using a split and merge approach. A biased language model (LM) is trained on the fly using corresponding human generated transcriptions. Since phonemic pronunciation modeling is not always possible for non-standard Arabic words, a graphemic pronunciation model (PM) is utilized to generate one pronunciation variant for each word.

Confidence score threshold	Segments filtered	Alignment accuracy
baseline (no thresh.)	0.0%	99.0%
>20.0%	2.5%	99.4%
>40.0%	2.7%	99.5%
>60.0%	3.7%	99.7%
>80.0%	5.6%	99.8%
>90.0%	10.0%	99.8%

Table 4: Alignment accuracy and filtered segments amount (%) on the *read-eval-set* after confidence score thresholding.

Initial alignment resulted in an accuracy of 95.4% and 98.9%, in terms of words, on the *conv-eval-set* and *read-eval-set* respectively. After applying unsupervised acoustic model adaptation, alignment accuracy is increased to 96.2% on the *conv-eval-set*. On the other hand, on the *read-eval-set*, AM adaptation has significantly improved anchor rate, however, alignment accuracy has not been improved. This is mainly interpreted because of the almost absence of speech disfluencies in the *read-eval-set*. Restricted language models have further increased anchor rate to 91.9% and 96.4% on the *conv-eval-set* and *read-eval-set* respectively. In this case, alignment accuracy was significantly increased to 97.7% on *conv-eval-set*, whilst for *read-eval-set*, negligible accuracy improvement was noticed . Most of mis-alignment errors were found to be with segments having significant background noise or significant speech disfluencies. A confidence scoring metric is proposed to accept/reject aligner output. By using confidence scores, it was possible to reject the majority of mis-aligned segments in the *conv-eval-set* resulting in more than 99.0% alignment accuracy.
For future work, conversational Arabic acoustic modeling can be improved using the proposed long audio alignment and confidence scoring applied on large amounts of Arabic speech data. Speech recognition results can be compared with conventional system trained with manually labeled speech data.

7. Acknowledgments

This publication was made possible by a grant from the Qatar National Research Fund under its National Priorities Research Program (NPRP) award number NPRP 09-410-1-069. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Qatar National Research Fund. We would like also to acknowledge the European Language Resources Association (ELRA) for providing us with MSA speech data resources used in AM training.

8. References

Billa, J., Noamany, M., Srivastava, A., Liu, D., Stone, R., Xu, J., Makhoul, J., and Kubala, F. (2002). Audio indexing of Arabic broadcast news. In *Proceedings of ICASSP*, volume 1, pages 5–8.

Elmahdy, M., Gruhn, R., Minker, W., and Abdennadher, S. (2009). Survey on common Arabic language forms from a speech recognition point of view. In *Proceedings of the International Conference on Acoustics (NAG/DAGA)*, pages 63–66.

Elmahdy, M., Gruhn, R., Minker, W., and Abdennadher, S. (2010). Cross-lingual acoustic modeling for dialectal Arabic speech recognition. In *Proceedings of INTERSPEECH*, pages 873–876.

Elmahdy, M., Gruhn, R., and Minker, W. (2011). *Novel Techniques for Dialectal Arabic Speech Recognition*. Springer, New York, first edition.

Elmahdy, M., Hasegawa-Johnson, M., and Mustafawi, E. (2013). A framework for conversational arabic speech long audio alignment. In *Proceedings of the 6th Language and Technology Conference (LTC)*, pages 290–293.

Ferguson, C. A. (1959). Diglossia. *Word*, 15:325–340.

Hazen, T.J. (2006). Automatic alignment and error correction of human generated transcripts for long speech recordings. In *Proceedings of INTERSPEECH*, pages 1606–1609.

Huggins-Daines, D., Kumar, M., Chan, A., Black, A.W., Ravishankar, M., and Rudnicky, A.I. (2006). Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *Proceedings of ICASSP*, volume 1, pages 185–188.

Leggetter, C.J. and Woodland, P.C. (1995). Maximum likelihood linear regression for speaker adaptation of the parameters of continuous density hidden Markov models. *Computer Speech and Language*, 9:171–185.

Liu, Y., Shriberg, E., and Stolcke, A. Y. (2003). Automatic disfluency identification in conversational speech using multiple knowledge sources. In *Proceedings of Eurospeech*, pages 957–960.

Maamouri, M., Graff, D., and Cieri, C. (2006a). *Arabic Broadcast News Speech, LDC Catalog No.: LDC2006S46*. Linguistic Data Consortium, Philadelphia.

Maamouri, M., Graff, D., and Cieri, C. (2006b). *Arabic Broadcast News Transcripts, LDC Catalog No.: LDC2006T20*. Linguistic Data Consortium, Philadelphia.

Moreno, P.J., Joerg, C., Thong, JM. V., and Glickman, O. (1998). A recursive algorithm for the forced alignment of very long audio segments. In *Proceedings of ICSLP*.