Technique for Automatic Sentence Level Alignment of Long Speech and Transcripts

Imran Ahmed, Sunil Kumar Kopparapu

TCS Innovation Labs - Mumbai, Yantra Park, Thane (West), Maharashtra, INDIA

ahmed.imran@tcs.com, sunilkumar.kopparapu@tcs.com

Abstract

A frugal approach to construct speech corpora, specially for resource deficient languages, is to exploit collections of speech and corresponding text data available in audio books, news, lectures. However, using these resources for building speech corpora require an alignment of the long duration speech data with the accompanying text data. Existing techniques for automatic speech-text alignment of long audio files assume availability of a basic speech recognition engine and hence cannot be directly used for resource deficient languages. In this paper, we propose a novel technique for sentence level alignment of long speech-text data by exploiting the syllable information in speech and text data. The proposed technique does not depend on the availability of any speech recognition models and hence can be used for resource deficient languages.

Index Terms: speech-text alignment, long audio, resource deficient, speech corpus, syllable

1. Introduction

A typical speech corpus is a set of audio files and its associated transcriptions. The process of creating a speech corpus in any language is a laborious, expensive and time consuming process, which means several languages do not have a speech corpus available; especially when the language has no viable commercial speech recognition engine. Of late researchers have been working on creating speech corpora for resource deficient languages using online resources (for example, [1, 2]) and a smaller amount of additionally collected speech data [3]. The approach is frugal in all the three senses, namely, it is less expensive, less laborious and less time consuming to construct the speech corpus. Specifically for Indian languages, All India Radio [4] provides online access to archives of news in various Indian languages. Recently [1] shows that frugal speech corpus constructed using online resources is a promising approach for training acoustic models for a resource deficient language, for use in ASR.

An essential requirement for development of a speech corpus is the need to align speech and the corresponding text transcript to make it useful. Typically, speech-text alignment of utterances can be achieved through the forced alignment technique using the Viterbi algorithm [5]. However, online resources are available as long audio and text data. Viterbi algorithm fails for forced alignment of utterances longer than a few minutes, due to memory requirements. Long audio-text alignment techniques proposed in literature are generally based on recursive forced alignment [6, 5] or ASR based techniques [7, 8, 9, 10]. [11] suggests the use of silence and prosody cues to segment long audio into smaller segments, followed by forced alignment; while [12] proposes a modified Viterbi forced alignment algorithm to handle long audio. Both forced alignment and ASR based techniques assume the availability of acoustic models for that language. Thus, the existing techniques for automatic speech-text alignment of long audio files cannot be directly applied for alignment of speech data for resource deficient languages.

In this paper, we propose a novel technique to enable sentence level long speech-text alignment even when there are no acoustic models available for that speech-text language pair. The main idea is to use the knowledge of detected syllables in speech and its transcript [1] to perform automatic sentence level alignment. Our proposed technique to identify the paragraph and sentence boundaries uses measures like syllable count and inter-syllable duration. Since the technique does not require any speech recognition models. It can be directly applied to different (resource deficient) languages. The rest of the paper is organized as follows. In Section 2 we formulate the problem and in Section 3 we discuss the speech syllable detection algorithm and the algorithm to mark paragraph and sentence boundaries. Section 4 discusses experimental results for sentence level alignment of Indian English and Hindi news, followed by the conclusion in Section 5.

2. Sentence level alignment of Long Audio

Let $a(t)$ for $0 < t < T$ be the audio resource (T is typically in order of minutes) and let Θ be the corresponding transcription. Let $a(t) \Leftrightarrow \Theta$ represent equivalence in terms of information content in some loose sense. Further assume that Θ is formatted in the sense of punctuations say paragraph, sentences, etc. The transcription Θ contains K paragraphs and S sentences. If P_k represent the k^{th} paragraph, then

$$\Theta = \bigcup_{1}^{K} P_k \tag{1}$$

Similarly, if P_k contains $|P_k|$ sentences and S_j^k represents the j^{th} sentence of the k^{th} paragraph of Θ, then

$$P_k = \bigcup_{1}^{|P_k|} S_j^k \tag{2}$$

Paragraph $\prec P \succ$and sentence $\prec s \succ$ boundaries are clearly identifiable in Θ as newlines and punctuations annotated into sentence boundaries. We expect to identify these sentences and paragraphs in the speech data; so as to enable sentence level speech-text alignment and segmentation of long audio and transcripts.

[1] we use *text* and *transcripts* interchangeably in this paper.

This translates to identifying paragraph boundary intervals in $a\,(t)$ such that

$$a\,(I_{P_k}) \Leftrightarrow P_k \tag{3}$$

where I_{P_k} is a time interval $(I^-_{P_k}, I^+_{P_k})$ and $0 < I^-_{P_k} < I^+_{P_k} < T$. Further, to identify sentence boundary in intervals $(I^-_{P_k}, I^+_{P_k})$ such that

$$a\left(I_{S^k_j}\right) \Leftrightarrow S^k_j \tag{4}$$

where $I_{S^k_j}$ is a time interval $(I^-_{S^k_j}, I^+_{S^k_j})$ and $0 < I^-_{P_k} < I^-_{S^k_j} < I^+_{S^k_j} < I^+_{P_k} < T$.

3. Syllable based Approach

We describe a novel syllable based technique to perform sentence level alignment of a long duration speech file to its transcript. We first automatically identify the syllables in spoken audio and also mark the syllables present in the corresponding text transcript. Then we use information about the sequence of syllables in speech and text to mark paragraph and sentence boundaries in speech. We first discuss detection of syllables in speech.

3.1. Syllable Detection in Speech

The algorithm described in [13] is used to detect syllable nuclei in spoken speech. The syllables in spoken speech are detected in two steps:

Step 1. Identify the peaks in intensity (dB) that are preceded and followed by dips in intensity. Only those intensity peaks that are above a certain intensity threshold are marked as potential intensity peaks.

Step 2. Discard the intensity peaks that are unvoiced. The voiced intensity peaks are the syllable nuclei.

In Step 2 it is required to identify the voiced and unvoiced regions in the spoken speech. These regions were identified using the pitch contour calculated as discussed in [14]. Detection of syllable in speech being a crucial component in aligning long speech-text in our approach, the algorithm for syllable detection [13] was tested for its performance accuracy. Table 1 captures the performance in terms of the average percentage error in the number of syllables detected by the algorithm with respect to the actual number of syllables present in the speech. A total of $5(3)$ Indian English (Hindi) news samples, each of, $5(10)$ min duration was used. The actual number of syllables in English news was calculated from the news transcript using [15]. And the actual number of syllables in Hindi news was calculated from the news transcript using method discussed in [16].

Table 1: Performance of Speech Syllable Detection.

	Indian English	**Hindi**
# of News Samples	5	3
Total Duration (min)	25	30
Syllable Count Error (%)	5.3	6.1

3.2. Marking $\prec P \succ$ and $\prec s \succ$ Boundaries in Speech

In an ideal scenario when the number of syllables identified in the text (N) are same as the syllables detected in the speech, then

$$s(m) \Leftrightarrow t(m) \quad \text{for} \quad 1 \leq m \leq N \tag{5}$$

where $s(m)$ is the m^{th} syllable detected in speech and $t(m)$ is the m^{th} syllable in the text data. This results in a perfect speech-text alignment. However, as seen in Table 1 the detection of syllables in speech (N_{ss}) do not match the number of syllables in text (N_{ts}), namely, $N_{ss} \neq N_{ts}$. Even if the overall number of syllables matched, namely, $N_{ss} = N_{ts}$, it is not guaranteed that (5) is valid. In general, the m^{th} syllable detected in speech will in reality actually correspond to the n^{th} syllable in text, namely,

$$s(m) \Leftrightarrow t(n), \tag{6}$$

where, $m = 1, \cdots, N_{ss}$ and $n = 1, \cdots, N_{ts}$. Note that in order to identify the paragraph and sentence boundaries in speech which corresponds to the $\prec P \succ$ and $\prec s \succ$ boundaries in text, we need to identify specific syllables in speech which correspond to text syllables at these boundaries. To identify syllables in speech which correspond to paragraph and sentence boundaries, we use three different seemingly orthogonal criteria.

3.2.1. Inter-Syllable Duration (ISD)

Inter-Syllable duration (ISD) can be calculated from syllables identified in the speech data. The ISD for the m^{th} speech syllable is given as:

$$ISD = (T_m - T_{m-1}) \tag{7}$$

where T_m is time of occurrence of the m^{th} syllable and T_{m-1} is time of occurrence of the $(m-1)^{th}$ syllable in speech. Note that the ISD is a characteristic feature of a paragraph and a sentence boundary because these boundaries are generally preceded by a pause, when spoken, leading to a higher ISD. For example as shown in Fig. 1 the utterance */Now the weather forecast/*, which is usually spoken after the end of general news and before the weather news has a longish pause leading to a higher ISD than the ISD of the syllables within this utterance. We exploit this observation to align $\prec P \succ$ and $\prec s \succ$. Fig. 2 is a histogram of the ISD for a news audio of 5 min duration. This news audio has 1258 detected syllables, distributed into bins, each of size 20 msec. It can be seen from the histogram that the number of syllables having higher ISD is smaller. It is not difficult to guess that these large valued ISD correspond to the paragraph. And as we keep moving to the left along the histogram, we come across syllables with ISD values corresponding to sentence boundaries.

However, note that though it is expected that the higher ISD values (or long pauses) correspond to paragraph and sentence boundaries the top $(K + S)$ syllables with the highest ISD values in speech may not necessarily correspond to the actual K paragraph and S sentence boundaries. For instance, for the histogram in Fig. 2, 2 out of 8 paragraph boundaries did not map to the top 10 ISD values (or pauses) and 6 out of 35 sentence boundaries did not map to the top 35 ISD values (or pauses) Clearly, the ISD measure (and pauses) alone is not sufficient to mark sentence boundaries in speech. Therefore we further define two more constraints for the occurrence of the boundaries in speech. To define the first constraint, we use the correspondence between the syllables in a given segment of speech and its transcription. To define the second constraint we projected

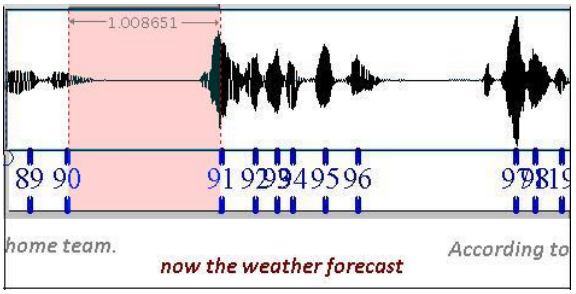

Figure 1: News utterance */Now the weather forecast/* as a new paragraph.

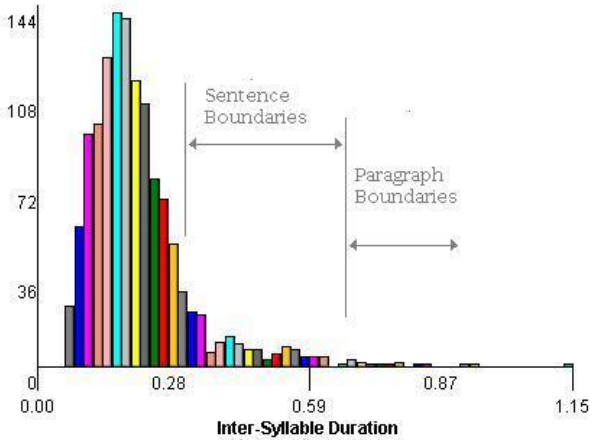

Figure 2: ISD Histogram for a 5 min English news audio. (No. of syllables = 1258, Bin size = 20 msec)

time of occurence of a syllable in text, using an average ISD from speech. These constraints are discussed in detail in the following sections.

3.2.2. Syllable Interval (SI) for Speech Boundary

We discuss a method to map a syllable at end of a paragraph (or sentence) from the text space into the speech space. Let N_{ss} be the number of syllables detected in speech and let N_{ts} be the number of syllables present in the actual text. Define a ratio $r \triangleq \frac{N_{ss}}{N_{ts}}$ and an error $\epsilon \triangleq |N_{ts} - N_{ss}|$. Now the n^{th} syllable in text can be mapped to a syllable in speech as

$$m = n * r \tag{8}$$

where m is m^{th} speech syllable (m is rounded of to the nearest integer). Similarly, if $N_{ts}^{P_k}$ is the number of text syllables in the k^{th} paragraph P_k, the number of speech syllables ($N_{ss}^{P_k}$) in the k^{th} paragraph P_k can be approximated as

$$N_{ss}^{P_k} = N_{ts}^{P_k} * r \tag{9}$$

Using this the syllable (m_{P_k}) at the boundary of the k^{th} paragraph in speech can be calculated as

$$m_{P_k} = m_{P_{k-1}} + N_{ss}^{P_k} \tag{10}$$

where, $m_{P_{k-1}}$ is the syllable at the boundary of the $(k-1)^{th}$ paragraph in speech.

However, it is very unlikely that the n^{th} syllable in text exactly matches the m^{th} syllable in speech, so instead we define an interval $[a, b]$ for m_{P_k} such that

$$[a, b] = [(m_{P_k} - \delta), (m_{P_k} + \delta)] \tag{11}$$

where δ provides the interval around m_{P_k}, where the paragraph boundary might be present. In all our experiments to mark paragraph boundaries in speech we used

$$\delta = \lceil N_{ts}^{P_k} * \frac{\epsilon}{N_{ts}} \rceil \tag{12}$$

3.2.3. Time Interval (TI) for Occurrence of Speech Boundary

We discuss a method to translate a paragraph (or sentence) boundary from the text space to the time space of speech. If we consider speech of duration T and assume that the speech was spoken at a uniform rate, the average ISD would be

$$\overline{ISD}_t = \frac{T}{N_{ts}} \tag{13}$$

Using (13), we compute

$$T_m = n * \overline{ISD}_t \tag{14}$$

where T_m is time of occurrence of the m^{th} speech syllable, obtained using n^{th} text syllable and $\overline{ISD}_t$. Accordingly, the time ($T_{m_{P_k}}$) of occurrence of the syllable at the boundary of the k^{th} paragraph in speech can be approximated as

$$T_{m_{P_k}} = T_{m_{P_{k-1}}} + (N_{ts}^{P_k} * \overline{ISD}_t) \tag{15}$$

where $T_{m_{P_{k-1}}}$ is the time of occurrence of syllable at the boundary of the $(k-1)^{th}$ paragraph in speech. And we define an interval $[T_a, T_b]$ around $T_{m_{P_k}}$ such that

$$[T_a, T_b] = [(T_{m_{P_k}} - T_\Delta), (T_{m_{P_k}} + T_\Delta)] \tag{16}$$

where $T_\Delta = \delta * \overline{ISD}_t$ and δ is calculated using (12).

3.3. Marking a Boundary in Speech

Given a long speech and the corresponding text data having K paragraphs, determine T, N_{ts} and N_{ss} and calculate r, ϵ and $\overline{ISD}_t$. For all the speech syllables, calculate ISD using (7). Start with the first paragraph ($k = 1$) and iterate following steps to mark the paragraph boundaries in speech:

Step 1 From text syllables, get the number of syllables $N_{ts}^{P_k}$ for the k^{th} paragraph. Calculate the speech syllable interval $[a, b]$ using (9), (10), (11) and (12). This interval is depicted as *Criteria: SI* in Fig. 3.

Step 2 Calculate time interval $[T_a, T_b]$ using (15) and (16). This interval is depicted as *Criteria: TI* in Fig. 3.

Step 3 From the speech syllables that lie in the union of intervals calculated in Step 1 and Step 2, mark the syllable that has the highest ISD value, as final paragraph boundary.

Step 4 Steps 1 to 3 are repeated until all paragraph boundaries are identified.

Once the boundaries of a paragraph are marked in speech, the boundaries of all the sentences in that paragraph can be marked using the same steps described above.

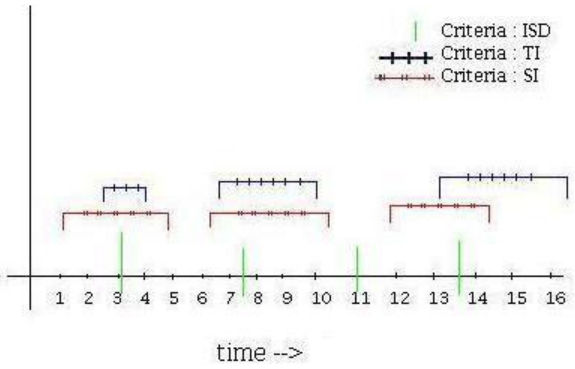

Figure 3: Steps to Mark a Speech Boundary

4. Experimental Results and Discussion

To evaluate the performance of our proposed technique, we used 2 hrs of Indian English news and 2 hrs of Hindi news available at [4]. The news is available as audio files of 10 min duration each; with its text transcription in a separate document. We used our algorithm described in Section 3.3 to mark paragraph and sentence boundaries in the speech files. The Indian English news contains 540 sentences and 103 paragraphs. 81% of the paragraph and sentence boundaries map to the top highest ISD values (or long pauses). However, with the *Criteria: SI* and *Criteria: TI* discussed in Section 3.2, our algorithm correctly identifies 97% of the paragraph and sentence boundaries in speech. Similarly, the Hindi news contains 670 sentences and 144 paragraphs. 74% of the paragraph and sentence boundaries map to the top highest ISD values (or long pauses). However, our algorithm correctly identifies 94% of these paragraph and sentence boundaries.

5. Conclusion

In order to build a speech corpus for resource deficient languages, people are using online speech and text data. However, to make this online resources useful it is required to align the long speech-text data. In this paper we proposed a novel technique based on syllable detection to align long speech-text data at sentence level. The idea is to map the boundaries corresponding to a paragraph and a sentence in the text data to the corresponding boundaries in the speech data. We first identified three different criteria and used the intervals derived from these three criteria to mark the final boundary in speech data corresponding to a paragraph and a sentence. These criteria overcome the errors made by techniques using silence or pauses to segment long audio into sentences. The proposed technique does not require any speech recognition models. For this reason, the approach can be directly applied to any (resource deficient) language.

6. Acknowledgements

The authors would like to thank Mithun and Chitralekha Bhat for their contribution in building the test data sets.

7. References

[1] I. Ahmed and S. Kopparapu, "Speech recognition for resource deficient languages using frugal speech corpus," in *ICSPCC2012*, Hong Kong, China, Aug 2012, (to appear).

[2] M. Davel, C. Van Heerden, N. Kleynhans, and E. Barnard, "Efficient harvesting of internet audio for resource-scarce asr," in *Interspeech 2011*, Florence, Italy, 2011.

[3] I. Ahmed and S. Kopparapu, "A frugal method and system for creating speech corpus," US20130030810, Tata Consultancy Services Limited, 2011.

[4] AIR, "All india radio news archives." [Online]. Available: http://www.newsonair.com/

[5] P. Moreno, C. Joerg, J. Van Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *ICSLP 98*, 1998, pp. 2711–2714.

[6] T. J. Hazen., "Automatic alignment and error correction of human generated transcripts for long speech recordings," in *ICSLP 06*, 2006, pp. 1606–1609.

[7] A. Katsamanis, M. Black, G. Georgiou, L. Goldstein, and S. Narayanan, "Sailalign: Robust long speech-text alignment," in *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, Jan. 2011. [Online]. Available: sail.usc.edu/software/SailAlign/

[8] C. Barras, G. Adda, M. Decker, B. Habert, P. Boula, and P. Paroubek, "Automatic audio and manual transcripts alignment, time-code transfer and selection of exact transcripts," in *Fourth International Conference on Language Resources and Evaluation (LREC)*, Lisbonne, 2004.

[9] C. Huang, W. Hsu, and S. Chang, "Automatic closed caption alignment based on speech recognition transcripts," Columbia University, Tech. Rep., December 2003.

[10] I. Trancoso and et al, "Spoken language technologies applied to digital talking books," in *INTERSPEECH 2006 ICSLP, Ninth International Conference on Spoken Language Processing*, Pittsburgh, PA, USA, September 2006.

[11] Y. Tao, X. Li, and B. Wu, "A dynamic alignment algorithm for imperfect speech and transcript," *Computer Science and Information Systems*, vol. 7, 2010.

[12] K. Prahallad and A. W. Black, "Segmentation of monologues in audio books for building synthetic voices," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, pp. 1444–1449, 2011.

[13] N. de Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior Research Methods*, vol. 41, pp. 385–390, 2009.

[14] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Institute of Phonetic Sciences, University of Amsterdam, Proceedings 17*, 1993.

[15] NU, "English syllable counter from northwestern university." [Online]. Available: http://morphadorner.northwestern.edu/morphadorner/documentation/javadoc/edu/northwestern/at/utils/corpuslinguistics/syllablecounter/EnglishSyllableCounter.html

[16] K. Prahallad and A. W. Black, "Unit size in unit selection speech synthesis," in *Eurospeech 2003*, Geneva, Switerzland, 2003.