Detecting gross alignment errors in the Spoken British National Corpus

Ladan Baghai-Ravary, Sergio Grau, Greg Kochanski
Phonetics Laboratory, University of Oxford, United Kingdom

We present methods for evaluating the alignment accuracy of the transcriptions with 46 recordings taken from the Spoken British National Corpus, which is an extensive and varied corpus of natural unscripted speech. Automatic checking of such alignments is crucial when analysing any very large corpus, since even the best current speech alignment systems will occasionally make serious errors. The methods described here use a hybrid approach based on statistics of the speech signal itself, statistics of the labels being evaluated, and statistics linking the two. This also provides an indication of the location of likely alignment problems; this should allow efficient manual examination of large corpora.

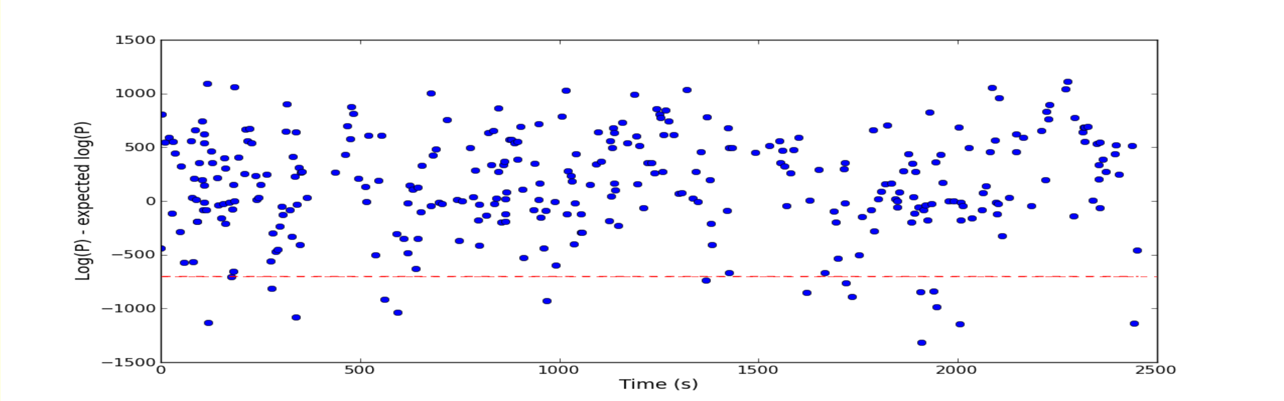
Methods

Word probability ("improbable")

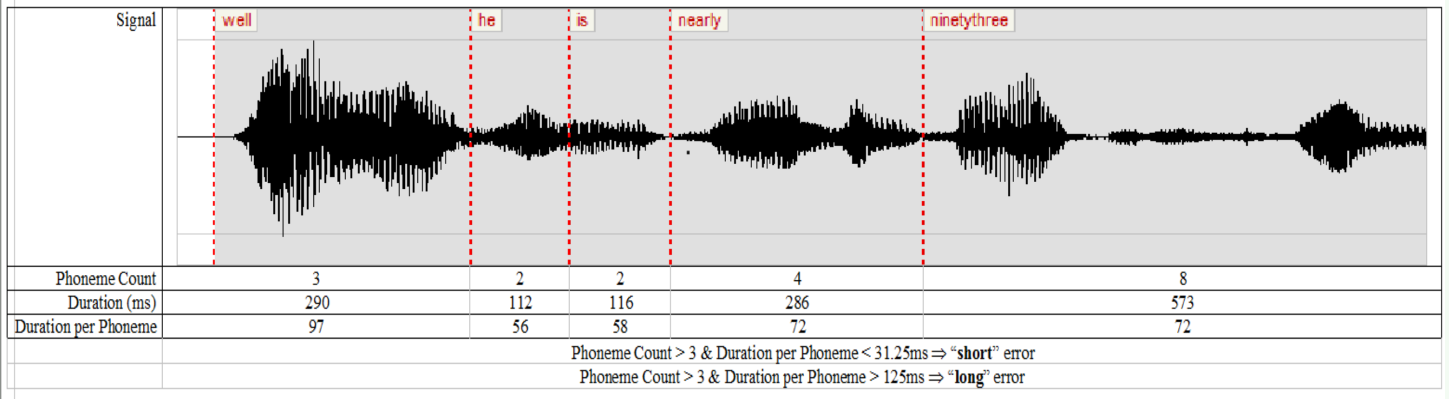
We used words with 4 or more phonemes & normalised the total probability of it with respect to its duration. This log probability per unit time provides a stable indication of the goodness of fit of the observed data to the HMM.

Unexpected probability ("unexpected")

We built a prediction of the aligner's log-probability score per unit length from the corpus as a whole, excluding the data files under analysis.

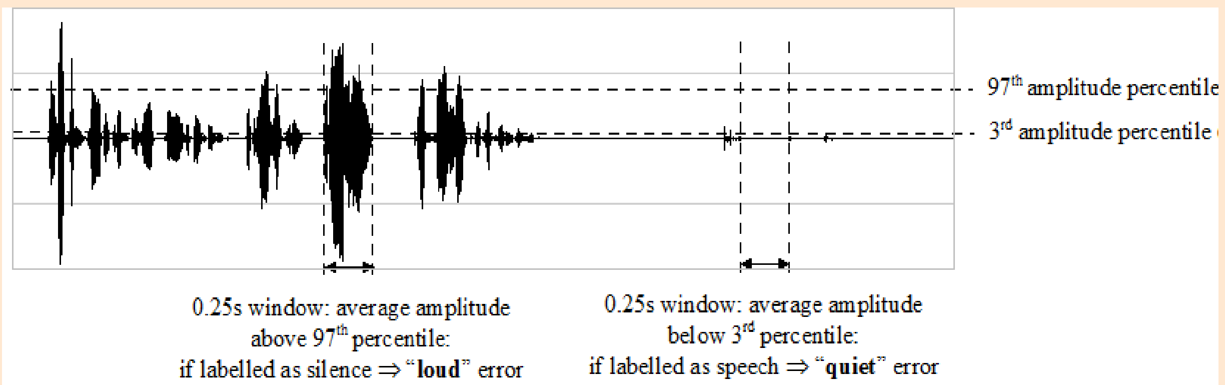


We measure the difference over a second long region. When this smoothed difference of log(P) is more negative than a threshold, the algorithm has identified a suspicious region.



Phoneme Count	3	2	2	4	8
Duration (ms)	290	112	116	286	573
Duration per Phoneme	97	56	58	72	72

Phoneme Count > 3 & Duration per Phoneme < 31.25ms ⇒ "**short**" error
Phoneme Count > 3 & Duration per Phoneme > 125ms ⇒ "**long**" error



0.25s window: average amplitude above 97th percentile: if labelled as silence ⇒ "**loud**" error

0.25s window: average amplitude below 3rd percentile: if labelled as speech ⇒ "**quiet**" error

Extremes of Amplitude ("loud" / "quiet")

The "quiet" and "loud" thresholds were set as the 3rd and the 97th percentile of the amplitudes observed over the whole of the recording. The nominal length of a short word was set to ¼ sec, assuming four phonemes with an average duration of 1/16 sec each. These parameters were estimated by experiment, and chosen to give a relatively small number of false positives. This algorithm produces two outputs ("loud", and "quiet") as it reports the extremes separately.

Word Duration ("short" / "long")

For words with four or more phonemes, we calculate the duration of the region labelled as the word, normalise it by dividing by the number of phonemes & compare it with the largest and the smallest average phoneme duration. Results outside the range: 1/32 s < mean duration < 1/8 s are flagged.

Duration Mismatch ("badlength")

We built a duration model for phonemes and then measured how far each phoneme deviates from the model. Regions are identified as suspicious if the smoothed absolute value of the deviation is large enough to exceed a threshold.

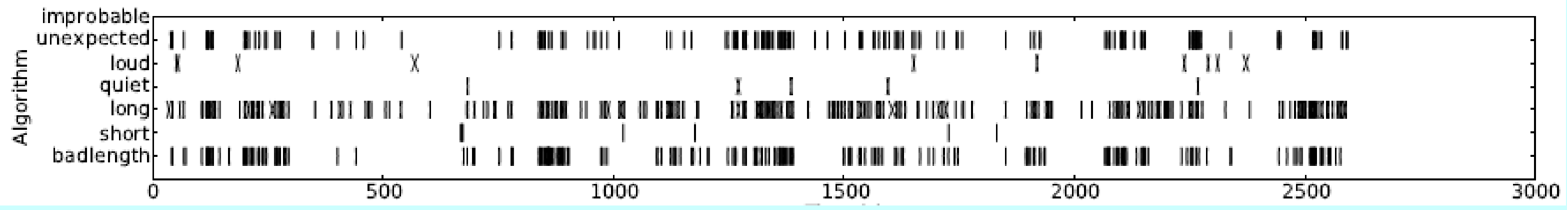
Human Evaluations

Examined 5-second long regions, approximately once every 60 seconds throughout the files, giving a score on a scale between 0 (very poor) and 10 (very good) for each file.

Results

- The evaluation methods were subjected to a number of correlation analyses, with and without two predefined non-linearities intended to make the distributions more normal. **short** was most commonly significant, followed by **badlength** and **loud**. At least one of those factors was significant at the $P < 0.01$ level in each regression.
- Pearson's R^2 averaged 0.66, reaching as high as 0.87 for some regressions, indicates that a combination of the algorithms was effective at matching the human judgements of overall alignment accuracy.
- There is a clear correlation between some of the detectors (**badlength, unexpected**, and **long**). These coincided 3.6 to 5.1 times more often than chance, with statistical significances well beyond $P < 0:001$.

The diagram below shows flagged regions for a typical file: *Each row shows where the respective algorithm flagged a potential alignment error. The audio file was manually rated 8/10 for alignment success.*



- Several of the pairs of algorithms were anticorrelated, notably **loud** vs. **short**, **badlength**, vs. **long**. This is due to the designs of the algorithms: **loud** triggers only on silences, while the others trigger only on speech sounds. As a result, they never pick the same phoneme, and only occasionally pick phonemes within 5 s of each other.
- The remaining pairs were either nearly independent (**loud** & **unexpected**, **badlength** & **short**, **long** & **short**, **short** & **unexpected**) or did not have enough occurrences to draw any reliable conclusion.
- Duration-based measurements (**short** and **badlength**) seem to perform best. One of the most useful indicators of a gross alignment error was the **short** algorithm which detected a sequence of phonemes whose durations were at the minimum allowed by their state topology (here, 3 states or 30 milliseconds). The relative lack of success of the **improbable** and **unexpected** algorithms was unexpected: both good and bad alignments have similar distributions of Log(P) scores.

Conclusion

Automated techniques can usefully identify regions of bad alignment. This research will allow for semi-automatic evaluation of the alignment of large speech corpora, which will be important for their future use in speech research.