

Word level precision of the NALIGN automatic segmentation algorithm

Kåre Sjölander and Mattias Heldner

Centre for Speech Technology (CTT), KTH, Stockholm

Abstract

This work presents an evaluation of the word level precision of an automatic segmentation algorithm, NALIGN. Measurements of the portion of temporal overlap between automatic and manual word level segmentations show that a precision of 90 to 95% can be achieved, partly depending on the type of speech material. These precision figures are furthermore only marginally lower than the gold standard obtained comparing independent manual segmentations.

Introduction

Automatic alignment of speech and written representations of the spoken material; or put differently automatic segmentation of speech; enables speech researchers to do extensive quantitative studies that would have been very time consuming and expensive otherwise. However, a prerequisite for using automatic segmentations in quantitative research is that the quality of the segmentation is known. The researcher needs to know the agreement to be expected between automatic and manual segmentations, if there are systematic errors affecting the outcome of the investigation, etc.

This work presents an evaluation of the quality of the word-level segmentation of an automatic segmentation algorithm, NALIGN. The evaluation includes the precision of the aligner, measured as the portion of temporal overlap between automatic and manual segmentations, as well as qualitative analyses of systematic errors.

Method

Speech material

Speech material collected within the GROG project (e.g. Carlson et al., 2002; Heldner & Megyesi, 2003) representing two different speaking styles, a read-aloud monologue and a spontaneous dialogue¹ was used in this evaluation. The read-aloud monologue data consists of a rendering of a children's book by a male professional reader at a publisher of talking books.

The recording is 17 minutes long and contains about 2700 words. The spontaneous dialogue material is made up of a radio broadcast, where a well-known female Swedish politician is interviewed by one male and one female interviewer. The interview lasts 25 minutes and contains approximately 4100 words.

Automatic segmentation

The automatic segmentation of the speech material was achieved by means of NALIGN, an alignment algorithm developed at CTT by the first author (e.g. Sjölander, 2003). NALIGN uses a Hidden Markov Model based method to generate word and phoneme level transcriptions from a verbatim orthographic transcription and a sound file. As speech contains various non-word phenomena with temporal extent, such as pauses, breaths and sighs, the aligner also inserts symbols for silences between words where needed. Other important features are that NALIGN can handle long sound files (> 1 h); and that processing time is proportional to the length of the sound file.

Manual segmentation

To establish a reference for the automatic segmentation, as well as a gold standard for the alignment precision (see below), the authors independently segmented the entire speech material. The manual segmentation was guided primarily by auditory perceptual criteria. Each word (as segmented by NALIGN) was presented in isolation, and the word boundaries were adjusted so that the audible traces of preceding or following words were minimized.

Alignment precision

The alignment precision was measured as the portion of temporal overlap between the automatic segmentation and the manual segmentations. The precision was estimated on a frame-by-frame basis (every 10th ms) including the silences inserted by the aligner, as well as on a word-by-word basis excluding those silences. The measurement of the portion of temporal overlap is illustrated in Figure 1.

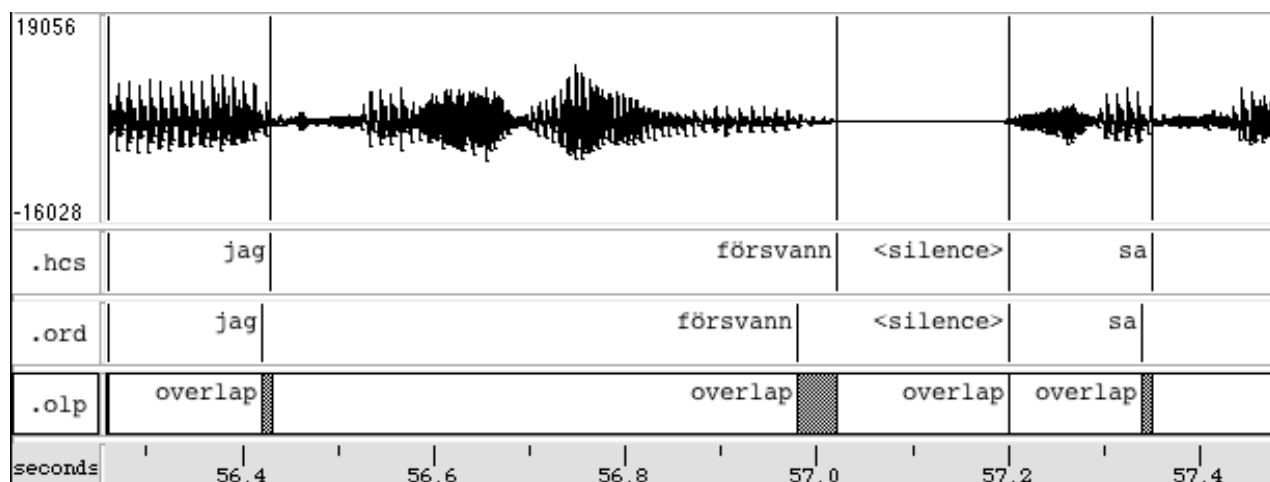


Figure 1. Illustration of the measurements of temporal overlap. The three label tiers beneath the waveform contain the manual segmentation (top); the output of the aligner (middle); and overlapping and non-overlapping portions (bottom). Non-overlapping portions are marked in dark grey.

The frame-by-frame overlap was measured as the duration of overlapping frames divided by the duration of all frames, including overlaps in silent portions (such as the overlap in <silence> in Figure 1).

The second measure, the word-by-word overlap, was used for more detailed analyses of words with different properties. The word-by-word overlap was measured as the duration of overlap in the words, divided by the duration of all words in the manual segmentation. That is, overlaps in silent portions (e.g. the overlap in <silence> in Figure 1) were excluded from the numerator as well as from the denominator.

A gold standard for the alignment precision was established by measuring the frame-by-frame overlap between the manual segmentations by the first and second authors.

Results

The results of the measurements of alignment precision using the frame-by-frame measure are shown in Table 1 together with the gold standard of overlap between manual segmentations.

These figures show that the overall precision of the automatic segmentation was fairly high. The temporal overlap between automatic and manual segmentation was 90 to 95% depending on the type of speech material and the human transcriber. That is, the precision was lower in the spontaneous dialogue than in the read-aloud monologue. Moreover, the alignment precision was only marginally lower than the gold standard.

Table 1. The alignment precision in the read-aloud monologue and spontaneous dialogue data measured using the frame-by-frame measure of overlap. The rows labeled NALIGN vs 1st and 2nd auth show the alignment precision relative to the manual segmentations by the first and second authors. Gold standard is the overlap between the manual segmentations.

	Read-aloud monologue	Spontaneous dialogue
NALIGN vs 1 st auth	95.9%	93.4%
NALIGN vs 2 nd auth	95.4%	90.4%
Gold standard	96.6%	94.9%

Error analysis

Qualitative examination of the discrepancies between manual and automatic segmentations revealed that there were several recurring types of errors related to the silences inserted by the aligner. There were several cases where the aligner underestimated the duration of words preceding a silence, and consequently overestimated the duration of the following silence (as in the word 'försvann' in Figure 1). There were also a number of inaccuracies involving erroneous insertions of silences. These errors occurred in the beginning or in the end of words, during voiceless stops or during portions with creaky voice. Thus, the aligner also underestimated the duration of words with stops or creaky voice at the word edges. There were also a few cases where word labels were wrongly placed relative to silences, typically a word placed before where it ought to have been after the silence.

Finally, the overall impression given by the error analysis was that words adjacent to a silence represent the worst cases with respect to alignment precision.

Words adjacent to silences

The silences inserted by the aligner thus account for several recurring error types. Furthermore, these silences are frequent. About 32% of the words in the read-aloud monologue and 38% of the words in the spontaneous dialogue are either preceded, followed or both preceded and followed by silences. To assess the influence of silences, the precision (here using the word-by-word overlap between the aligner and the manual segmentation by the second author) was measured separately for words preceded and followed by silences in the automatic segmentation, and for words neither preceded nor followed by silences. These results are shown in Table 2.

Table 2. The precision (using the word-by-word overlap) in words preceded or followed by silences, and in words neither preceded nor followed by silences, in the read-aloud monologue and spontaneous dialogue data.

	Read-aloud monologue	Spontaneous dialogue
Silence before	91.4%	87.3%
Silence after	92.4%	85.3%
No adjacent silence	95.3%	92.1%

These analyses revealed that the precision was about 3 to 5% lower in words adjacent to a silence compared to words neither preceded nor followed by a silence. The most substantial effect was found in the spontaneous dialogue. However, there were no clear-cut differences between the words followed and those preceded by silences. Furthermore, the words adjacent to silences account for 51% of the total error in the read-aloud monologue, and for 61% of the total error in the spontaneous dialogue.

Discussion and conclusions

This study has shown that the overall word level precision of the automatic aligner is fairly high. In fact, the precision is only marginally lower than the gold standard obtained by comparing the segmentations by two human segmenters using perceptual segmentation criteria. Thus, the loss of precision is minor, while the gain in time and effort is considerable. As can be

expected, automatic segmentation is more difficult in spontaneous speech than in read-aloud speech.

We have also isolated an important source of the remaining error, namely the transitions to and from silences. These transitions result in several systematic errors and represent a substantial portion of the total error. Improving the precision in words adjacent to silences would therefore improve the overall precision considerably. A reason for these systematic errors is probably that the Hidden Markov Models represent average spectra of phones, as well as of silences. Thus, the aligner boundaries between phones and silences will be at an intermediate position on the slope between the two, often leaving audible traces of the phone in the silence. Combining the HMM method with detection of periodicity or a threshold for sound energy is probably a reasonable path to follow to improve the situation.

Although this evaluation has only dealt with the precision of the word level segmentation by the aligner, we expect a similar precision for the phoneme level segmentation. The evaluation is based on around 6800 words, thus the starting points of 6800 word initial segments and the endpoints of 6800 word final segments have been evaluated. The fact that only segments at the edges of words have been evaluated together with the observations of systematic errors related to silences between words may even result in a higher phoneme level precision. Future work could verify this by explicit investigations of phoneme level precision.

To conclude, this study has shown that the automatic segmentation algorithm NALIGN is well suited as a basis for large quantitative studies already in its present shape. The few systematic errors that occur leave room for future improvements. Currently NALIGN is used within the GROG project for duration and intonation modeling (e.g. Heldner, Edlund, & Björkenstam, 2004; Heldner & Megyesi, 2003), and within the NICE project for automatic generation of the database used in a unit selection synthesizer (e.g. Sjölander & Gustafson, forthcoming).

Acknowledgements

The authors are grateful to the Swedish Radio (SR) and the Swedish Library of Talking Books and Braille (TPB) for making the speech data available. This work was carried out within the GROG project (Gräns och gruppering – Struk-

turering av talet i olika kommunikativa situationer) supported by The Swedish Research Council (VR) 2002–2004. The research was carried out at CTT, the Centre for Speech Technology, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations.

Notes

¹ *The classification of speaking styles was far from clear-cut. The read-aloud monologue contained examples of dialogues with one speaker acting more than one character, and the spontaneous dialogue contained longer stretches of uninterrupted or monologue-like speech. It is also worth noting that the interviewee, i.e. the politician, did most of the talking in this recording.*

References

- Carlson R., Granström B., Heldner M., House D., Megyesi B., Strangert E., and Swerts M. (2002) Boundaries and groupings - the structuring of speech in different communicative situations: a description of the GROG project. *THM-QPSR 44*, 65-68.
- Heldner M., Edlund J., and Björkenstam T. (2004) Automatically extracted F0 features as acoustic correlates of prosodic boundaries. In *Proceedings of Fonetik 2004*. Stockholm.
- Heldner M., and Megyesi B. (2003) Exploring the prosody-syntax interface in conversations. In *Proceedings ICPhS 2003* (pp. 2501-2504). Barcelona.
- Sjölander K. (2003) An HMM-based system for automatic segmentation and alignment of speech. In *Proceedings of Fonetik 2003* (pp. 93-96). Umeå.
- Sjölander K., and Gustafson J. (forthcoming) Voice creation for conversational fairy-tale characters. In *To appear in proceedings of the 5th ISCA Speech Synthesis Workshop*. Pittsburgh.