An Improved Speech Segmentation Quality Measure: the R-value

Okko Johannes Räsänen[1], Unto Kalervo Laine[1], and Toomas Altosaar[1]

[1]Department of Signal Processing and Acoustics, Helsinki University of Technology, Finland
Okko.Rasanen@tkk.fi, Unto.Laine@tkk.fi, Toomas.Altosaar@tkk.fi

Abstract

Phone segmentation in ASR is usually performed indirectly by Viterbi decoding of HMM output. Direct approaches also exist, e.g., blind speech segmentation algorithms. In either case, performance of automatic speech segmentation algorithms is often measured using automated evaluation algorithms and used to optimize a segmentation system's performance. However, evaluation approaches reported in literature were found to be lacking. Also, we have determined that increases in phone boundary location detection rates are often due to increased over-segmentation levels and not to algorithmic improvements, i.e., by simply adding random boundaries a better hit-rate can be achieved when using current quality measures. Since established measures were found to be insensitive to this type of random boundary insertion, a new R-value quality measure is introduced that indicates how close a segmentation algorithm's performance is to an ideal point of operation.
Index terms: blind speech segmentation, segmentation evaluation.

1. Introduction

Automatic speech segmentation has many applications in speech processing and phonetics, e.g., in automatic speech recognition and automatic annotation of speech corpora. Several methodological approaches to automatic segmentation have therefore been proposed (e.g., [1-9]). In order to develop and test a segmentation algorithm, the properties of the automatically created speech segments need to be analyzed in detail, and therefore automated evaluation methods are required. These methods should provide a fast, independent, and overall estimate of algorithm performance over large amounts of data. This would enable efficient experimentation regarding the effects of different parameter values and would make analysis spanning different annotated corpora more feasible. Moreover, if the output quality can be described using a single reliable measure that is able to indicate the distance and direction from the point of ideal performance, automatic optimization of algorithm parameters would become much more facilitated.

However, evaluation methods described in literature are not self-explanatory and therefore cannot be repeated in an exact manner. No single best — or in any other way — approved method for describing the accuracy has been suggested, the trend being that many authors just adapt some conventional approach exploiting boundary search regions without specifying their use. The ambiguity associated with these evaluation approaches leads to problems when comparisons between approaches are conducted at different sites, and more importantly, the results may become unreliable in terms of the relationship between the real phonetic content of speech and the algorithmic output if the evaluation is not performed with care. Previous publications in this area were therefore comprehensively investigated. As a result, this paper identifies two major problems and offers solutions for them. The first one concerns itself with how correctly detected segment boundaries are computed. The second problem arises from over-segmentation and its corruptive effects on the obtained hit rates.

In order to sufficiently resolve the detected problems in evaluation, a method for correctly counting detected segment boundaries is explicitly defined and a new segmentation quality measure, called the R-value, is proposed.

2. Evaluation methodology

2.1 Evaluation reference

In order to perform automatic evaluation it is necessary to have access to a reliable reference that indicates true segment locations in speech. The convention is to perform a segmental boundary comparison between an automatic method and a manually produced segmentation, since many well-known speech corpora are provided with an annotation created manually by one or more trained phoneticians. While manual segmentation is prone to the variability present in individual judgments, it is often considered as a reliable baseline for quality if it is carefully produced [10].

2.2 Quality measures

When algorithmic output is compared to a reference, a number of measures can be computed. *Insertions* are detected when one or more boundaries created by a segmentation algorithm do not match any reference boundary, or, if there are several generated boundaries in the vicinity of only one reference boundary. *Deletions* are noted when there is a boundary marked in the reference, but the algorithm produces no corresponding boundary. Finally, correctly detected boundaries are considered as *hits*.

By using these measures, the overall segmentation accuracy is usually defined in terms of hit-rate (HR). For some finite section of speech let N_{hit} be the number of boundaries correctly detected and N_{ref} be the total number of boundaries in the reference. HR can then be calculated using equation 1 in table 1 [1]. Another central measure, especially in the case of blind methods, is the over-segmentation (OS) rate, which is the ratio of the total number of detected boundaries N_f to the number of boundaries in the reference N_{ref}(2) [11].

Precision (3) describes the likelihood of how often the algorithm identifies a correct boundary whenever a boundary is detected. Recall (4) is the same as *HR* (1) except that it is not scaled to be a percentage. In order to describe the performance of an algorithm with one scalar value, the F-value (5) can be computed from precision (3) and recall (4) [12]. False-alarm rates and miss rates are also sometimes used (e.g., [6]) and can be derived directly from the above measures.

Table 1: *The most common quality measures used for segmentation.*

$HR = \frac{N_{hit}}{N_{ref}} * 100$ (1)	$OS = (\frac{N_f}{N_{ref}} - 1) * 100$ (2)
$PRC = \frac{N_{hit}}{N_f}$ (3)	$RCL = \frac{N_{hit}}{N_{ref}}$ (4)
$F = \frac{2.0 * PRC * RCL}{PRC + RCL}$ (5)	

2.3 Counting the hits: the search region method

In order to determine the number of hits, deletions, and insertions, the reference annotation has to be somehow compared to a segmentation algorithm's output. The practice evident in literature is to place a fixed-size search-region around each reference boundary and verify whether the segmentation algorithm has produced any boundaries in these regions. However, a major source of ambiguity exists in literature that concerns the overlapping of search regions [1-9]. A typical definition reads as: *"a boundary is considered to be correctly detected if the hypothesis and the manual transcription are within 20 ms of each other"* without any further specifications (from [2], p. 2; see also, e.g., [1], [3-9]). Situations, in which there are two reference boundaries within 40 ms of each other[1], while the algorithm produces a single boundary in the overlapping region, are not well defined (fig. 1). The manner in which reference and segment output boundaries are paired in these situations and whether re-use of boundaries for several hits is explicitly ruled out leads to different hit-rates. Such subtle differences in interpretation may yield changes in performance as large as 5 % [13] and therefore acts as a large source of inconsistency in the reporting of segmentation algorithm results.

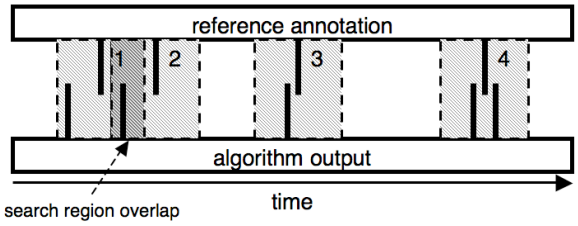

Figure 1: *Example of an overlapping search region causing an ambiguous situation in evaluation. The second algorithmically produced boundary is within two search regions simultaneously, leading to the problem of how to define a matching boundary for each reference boundary.*

A simple method to avoid the overlap problem can be formulated as follows: search regions of a typical fixed size, e.g., ±20 ms, are placed around each reference boundary. If overlapping search regions exist, that is, adjacent regions with their reference boundaries exist closer than 40 ms to each other, then the regions are asymmetrically shrunk to divide the space between two reference boundaries into two equal-width halves (similarly to [11], but now with a maximum search region size; see fig. 2). This prevents ambiguous situations associated with overlapping search regions. Now each region can be searched for algorithmically generated boundaries. Every search region

[1] For example, in TIMIT 21.9 % of all boundaries are closer than 40 ms to each other.

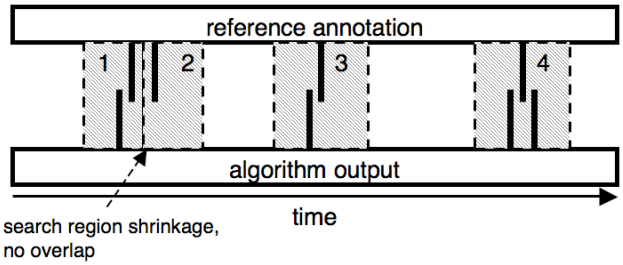

Figure 2: *The overlap of regions is removed by asymmetrically shrinking the search regions of boundary 1 and 2 to a common mid-point (indicated by the arrow). The matching of reference boundaries to algorithmic boundaries now becomes straightforward.*

containing an algorithmically generated boundary is considered as a hit and all additional boundaries are counted as insertions. Empty regions are considered as deletions.

3. Stochastic over-segmentation

One notable aspect of the search region approach is that a relatively large proportion of the signal timeline becomes covered with search regions, since normal rate speech contains about ten phones per second. For example, with the TIMIT corpus, 45 % of all audio material already falls into some search region when a ±20 ms search region is specified (fig. 3). This is one reason why larger search regions become questionable for evaluation purposes. In some publications even larger search regions up to ±100 ms have been used [14-15]. Even with only a 70 ms search region (± 35 ms) around each TIMIT phone boundary, 68 % of the timeline becomes covered causing many sporadically generated phone boundaries to be classified as correct. This clearly would permit a very poor segmentation algorithm to fare well since any generated boundary would only have an approximately 1/3 chance of falling outside of a hit window. Obviously, improvements are required in how segmentation algorithm performance is measured.

As explained above, expanding timeline coverage is a potential problem in the evaluation of any generic segmentation algorithm since the probability that a randomly inserted boundary hits a search region increases when more of the timeline is covered. To demonstrate the interdependence of accuracy and over-segmentation, a stochastic segmentation experiment was performed. In this experiment boundaries were generated at entirely random temporal locations with a stochastic process. By counting the number of boundaries hitting the search regions, the hit-rate started to increase along with the over-segmentation value. This movement can be seen when plotted in the segmentation performance plane (hit-rate vs. over-segmentation) near the bottom of figure 4 as the dashed line labeled "*Theoretical stochastic process*". Note that a segmentation algorithm that would match the reference annotation perfectly would be considered to be performing ideally and would have its operating point marked at the 100 % hit-rate and 0 % over-segmentation levels. This point of ideal operation is referred to as the *target-point* (TP).

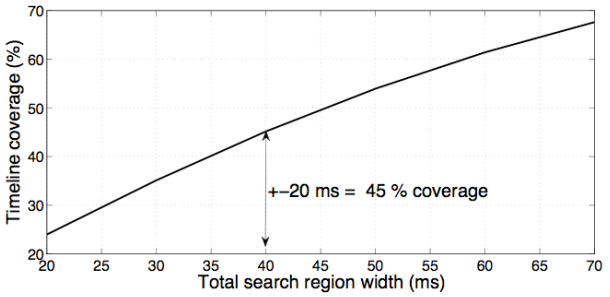

Figure 3: *Search region timeline coverage in TIMIT material as a function of search region width (overlapping sections of adjacent search regions are not included twice in the analysis).*

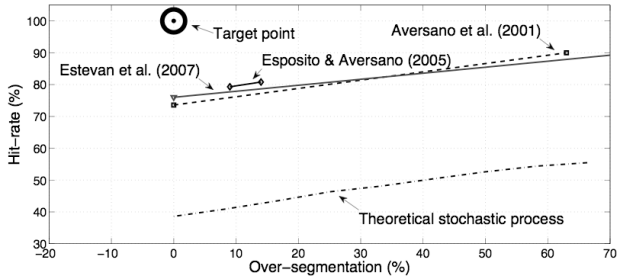

Figure 4: *Use of the segmentation performance plane to display results from different segmentation algorithms ([1-2], [6]) as well as the result obtained using random insertion of segment boundaries. Note that the target-point indicates the point at which a segmentation algorithm would perform ideally, i.e., the algorithmic output would match the reference annotation according to a defined search region.*

The increase of *HR* as a function of *OS* in this process has significant similarity to the state-of-the-art segmentation results reported in literature ([1-2], [6-7]), i.e., the rate at which hit-rate increases (by allowing more over-segmentation) does not seem to be any larger than what can be obtained by randomly inserting boundaries.

This observation leads to the conclusion that the segmentation results that have been reported in literature with relatively high over-segmentation values indicate very little about the internal characteristics of an algorithm. If the increase in accuracy (as is the case with higher levels of OS) starts to align itself with the theoretical stochastic process curve in fig. 4 (i.e., parallel in direction indicating a similar slope), the capability of the algorithm to provide information about meaningful phonetic changes in the signal becomes negligible. Therefore, results from an entirely random segment generation process that assumes no knowledge of the underlying speech signal can be used to effectively define a zero-level segmentation quality baseline.

4. R-value as a measure for segmentation quality

Optimizing the operation of a speech segmentation algorithm is often a tradeoff between hit-rate and over-segmentation (or inversely, false-alarm rate and miss-rate). In order to find a suitable operating point, a proper balance between these two measures needs to be determined. The previously introduced F-value (5) is one possible way to describe overall performance of an algorithm with a single value. However, the F-value is prone to stochastic hit-rate increases due to the over-segmentation problem described in section 3. In order to describe performance using a single value that is also sensitive to over-segmentation, a novel measure was developed.

The theoretical goal of segmentation is to achieve operation around the *target-point* (TP) that is located at the 100 % hit-rate and 0 % over-segmentation levels as compared to a reference. The basis of the new measure is the algorithm's distance from TP and not the (hit-rate) gain achieved by over-segmentation. On the segmentation performance plane, a distance r_1 from the segmentation result to TP can be derived (6). Additionally, to appreciate the value of under-segmentation compared to over-segmentation in an algorithm (i.e., less false positives), another distance r_2 (7) is measured from the segmentation result perpendicularly to the ideal zero-insertion line y = x – 100 (fig. 5). This line is the left-side theoretical limit for possible results in this space and extends from –100 % over-segmentation and 0 % hit-rate, to the 100 % hit-rate level with 0 % over-segmentation (e.g., with a 50 % hit-rate, over-segmentation needs to be –50 % in order to avoid any insertions). The distances r_1 and r_2 are then added together and normalized to have a maximum value of 1 at the target-point (8). This new distance measure, referred to as the *R-value*, decreases as the distance to the target grows, i.e., similarly as the F-value does, but is critical towards over-segmentation.

$$r_1 = \sqrt{(100 - HR)^2 + (OS)^2} \tag{6}$$

$$r_2 = \frac{-OS + HR - 100}{\sqrt{2}} \tag{7}$$

$$R = 1 - \frac{abs(r_1) + abs(r_2)}{200} \tag{8}$$

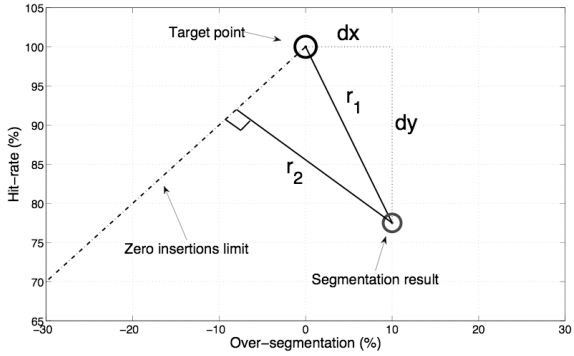

Figure 5: *R-distance is calculated by summing distance r_1 (distance from the segmentation algorithm's operating point "Segmentation result" to the target-point), with r_2 (distance from "Segmentation result" to the ideal zero-insertions limit), and then normalized according to (8).*

Figure 6 shows the behavior of F- and R-values in the segmentation performance plane using equal value curves. Dashed lines indicate how hit-rate increases as a function of over-segmentation due to the stochastic generation of boundaries. As can be seen, the F-value behaves in a linear manner when compared to the R-value. High over-segmentation rates are more severely penalized when using the R-value as compared to the F-value, and the R-value drops dramatically above OS = 0 % when OS is increased unless the accuracy is actually increasing more rapidly even with the generally detrimental effect of increased random insertions.

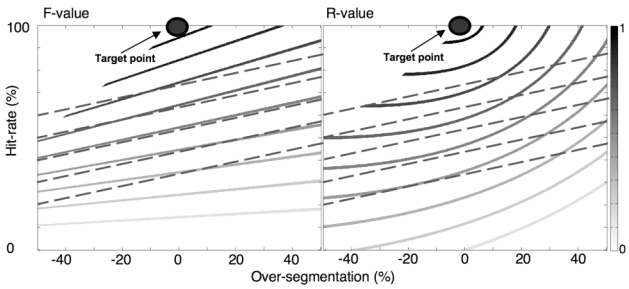

Figure 6: *F-value and R-value equal value curves in hit-rate vs. over-segmentation coordinates. Superimposed dashed lines indicate hit-rate increase as a function of over-segmentation in stochastic boundary generation: five different elevated offsets for the theoretical stochastic process are shown. The F-value curves (solid lines) are nearly parallel with the effects of stochastic boundary generation (dashed lines) indicating the F-value's strong correlation with over-segmentation and thus its weakness as a reliable quality measure. On the other hand, the R-value measure is aware of the level of over-segmentation and can be used to direct the segmentation algorithm towards the target-point.*

Thus, the R-value not only measures the quality of a segmentation algorithm but can also be used to automatically direct the automatic segmentation process towards a goal, e.g., the target-point. It should also be noted that distance r_1 alone can be used to find the optimal operating point from an operating curve of an algorithm in terms of equally weighted HR and OS rates.

5. Conclusions

Serious weaknesses in the manner in which the performance of speech segmentation algorithms are currently measured were shown. It was found that random insertion of segment boundaries increases hit-rate as a function of over-segmentation due to a large covering of the search regions in the speech timeline. This increase in hit-rate correlates significantly with the high over-segmentation results reported in literature, suggesting that at higher over-segmentation rates a stochastic process starts to dominate the results instead of the capabilities of the tested algorithm. Therefore, a novel measure called the *R-measure* was introduced that can be applied to the evaluation of any automatic speech segmentation algorithm. The R-measure increases towards the ideal target-point (100 % hit-rate and 0 % over-segmentation) and is much more sensitive to increases in over-segmentation levels than previously used measures like the F-value. In combination with the search region method explicitly defined in section 2.3, this measure provides for an independent and more relevant quality score for automatic segmentation of speech.

Acknowledgements

This research is funded as part of the EU FP6 FET project Acquisition of Communication and Recognition Skills (ACORNS), contract no. FP6-034362. The authors would like to thank Douglas O'Shaughnessy of INRS for his valuable comments.

References

[1] G. Aversano, A. Esposito, A. Esposito, and M. Marinaro, "A New Text-Independent Method for Phoneme Segmentation," *Proc. IEEE international Workshop on Circuits and Systems*, vol. 2, pp. 516-519, 2001.

[2] Y. P. Estevan, V. Wan, and O. Scharenborg, "Finding Maximum Margin Segments in Speech," *Proc. ICASSP '07*, vol. 4, pp. 937-940, 2007.

[3] A. Sarkar and T. V. Sreenivas, "Automatic speech segmentation using average level crossing rate information, " *Proc. ICASSP '05*, vol. 1, pp. 397-400, 2005.

[4] K. Sjölander, "An HMM-based system for automatic segmentation and alignment of speech," *Proc. Fonetik 2003*, Umeå University, Dept of Philosophy and Linguistics, pp. 93-96, 2003.

[5] M. Sharma and R. Mammone, ""Blind" speech segmentation: automatic segmentation of speech without linguistic knowledge," *Proc. ICSLP'96*, vol. 2, pp. 1237-1240, 1996.

[6] A. Esposito and G. Aversano, "Text Independent Methods for Speech Segmentation," In G. Chollet et al. (eds.): *Nonlinear Speech Modeling, Lecture Notes in Computer Science*, Springer Verlag, vol. 3445, pp. 261-290, 2005.

[7] G. Almpanidis and C. Kotropoulos, "Phonemic segmentation using the generalized Gamma distribution and small sample Bayesian information criterion." *Speech Communication,* vol. 50, pp. 38-55, 2008.

[8] O. Scharenborg, M. Ernestus, and V. Wan, "Segmentation of speech: Child's play?", *Proc. Interspeech'07*, pp. 1953-1956, 2008.

[9] A. Cherniz, M. Torres, H. Rufiner, and A. Esposito, "Multiresolution Analysis applied to Text-Independent Phone Segmentation." *Journal of Physics: Conference Series,* vol. 90, 012083, 2007

[10] M.-B. Wesenick and A. Kipp, "Estimating the Quality of Phonetic Transcriptions and Segmentations of Speech Signals," *Proc. ICSLP'96*, pp. 129-132, 1996.

[11] B. Petek, O. Andersen, and P. Dalsgaard, "On the Robust Automatic Segmentation of Spontaneous Speech," *Proc. ICSLP'96*, pp. 913-916, 1996.

[12] J. Ajmera, I. McCowan, and H. Bourlard, "Robust Speaker Change Detection," *IEEE Signal Processing Letters*, vol. 11, no. 8, pp. 649-651, 2004.

[13] O. J. Räsänen, "Speech Segmentation and Clustering Methods for a New Speech Recognition Architecture," Master's thesis, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, http://lib.tkk.fi/Dipl/2007/urn010123.pdf, 2007.

[14] K. Demuynck and T. Laureys, "A Comparison of Different Approaches to Automatic Speech Segmentation," *Proc. 5th International Conference on Text, Speech and Dialogue (TSD)*, pp. 277-284, 2002.

[15] K. Kvale, "Segmentation and Labeling of Speech," Ph.D. Dissertation, The Norwegian Institute of Technology, 1993.