An Improved Algorithm for the Automatic Segmentation of Speech Corpora

Tom Laureys, Kris Demuynck, Jacques Duchateau and Patrick Wambacq

Katholieke Universiteit Leuven, ESAT - PSI
Kasteelpark Arenberg 10
3001 Leuven (Belgium)
{tom.laureys,kris.demuynck,jacques.duchateau,patrick.wambacq}@esat.kuleuven.ac.be

Abstract

In this paper we describe an improved algorithm for the automatic segmentation of speech corpora. Apart from their usefulness in several speech technology domains, segmentations provide easy access to speech corpora by using time stamps to couple the orthographic transcription to the speech signal. The segmentation tool we propose is based on the Forward-Backward algorithm. The Forward-Backward method not only produces more accurate segmentation results than the traditionally used Viterbi method, it also provides us with a confidence interval for each of the generated boundaries. These confidence intervals allow us to perform some advanced post-processing operations, leading to further improvement of the quality of automatic segmentations.

1. Introduction

This paper describes a novel approach to the automatic segmentation of speech corpora. Automatic segmentations of speech, on phoneme level (eg. TIMIT) or word level (eg. CGN, Switchboard), are a standard annotation within speech corpora. In segmented speech corpora, the phonemes or words are coupled to their corresponding segment in the speech signal by means of time stamps. Speech technologists apply segmentations in the bootstrapping process for training acoustic ASR models, in the development of TTS systems and within speech research in general. For other users, segmentations provide fast and easy access to audio fragments of words or phoneme sequences in the corpus. Some speech corpora provide only automatic segmentations, thus requiring a highly accurate segmentation algorithm. In other corpora speech segmentations are also checked manually. Since this manual procedure is time-consuming and thus expensive, it is important to base the manual work on an already accurate automatic segmentation in order to speed up the verification process. So in both cases an automatic segmentation procedure producing high-quality output is necessary.

In the literature several systems for automatically generating speech data segmentations have been described. Most of them have been applied to databases for TTS systems. Some of the methods described in the literature are based on specific acoustic cues or features for the segmentation task (Vorstermans et al., 1996; van Santen and Sproat, 1999; Husson, 1999) focusing for instance on transient behaviour or specific differences between phoneme classes. Others use general features and acoustic modelling which are common in ASR (Ljolje and Riley, 1991; Beringer and Schiel, 1999). The method we present here is of the latter type.

Speech segmentation systems typically take both the speech signal and its phonemic transcription as input. The phonemic transcription may be generated manually (as is the case in our experiments) or may be automatically derived from the orthographic transcription and a phonemic dictionary.

In this paper we focus on automatic segmentation deduced from the Forward-Backward algorithm. Forward-Backward segmentation outperforms Viterbi segmentation in two important respects. First, we performed a set of experiments in which automatic segmentations (based on Viterbi and Forward-Backward respectively) were compared to manually checked reference segmentations, showing that the Forward-Backward algorithm generates more accurate segmentations than the Viterbi algorithm. Second, Forward-Backward segmentation provides us with a confidence interval for each generated boundary (Laureys et al., 2001). The application of confidence intervals to automatic segmentation is novel and has a potentially wide-ranging applicability. In view of the topic of this paper, we successfully applied the confidence intervals to an advanced post-processing procedure which further improves the automatic segmentations.

The paper is organised as follows. In section 2., we explain how an automatic segmentation is generated and describe both the Viterbi algorithm and the Forward-Backward algorithm. In addition, we focus on the confidence intervals and on how we applied them to obtain more accurate segmentations. Section 3. presents the set-up for our experiments and discusses the results. We end with conclusions and suggestions for future research.

2. Automatic Segmentation of Speech

Automatic segmentation of speech is based on the following process. First, the phonemes in the input phonemic transcription are coupled to their respective acoustic Hidden Markov Models (HMMs). The acoustic properties of the HMM states s_i are modelled by means of observation density functions $f_i(x) = f(x \mid s_i)$, x being the feature vector that describes a given speech frame at 10 ms intervals. The duration and possible order of the states are governed by the transition probabilities between those states $a_{ij} = p(s_j \mid s_i)$. The HMM phoneme models typically have three states and a left-to-right topology, as illustrated in figure 1.

Once the acoustic properties of the different phones have been encoded in statistical models, sentence models are generated by concatenating all relevant phoneme mod-

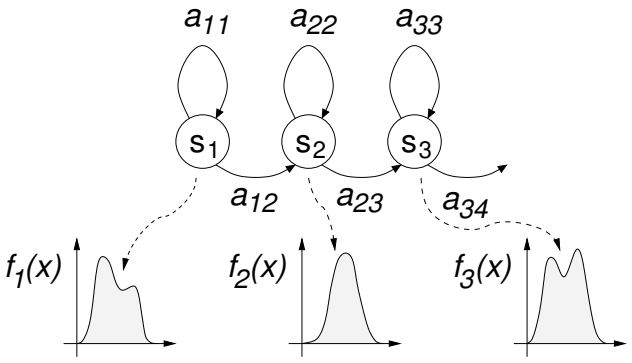

Figure 1: A phoneme model with three states and a left-to-right topology.

els (see figure 2). Next, the speech data are assigned (hard or soft, by respectively Viterbi or Forward-Backward) to the acoustic model of the complete phoneme sequence, still adhering to the left-to-right constraints of the model.

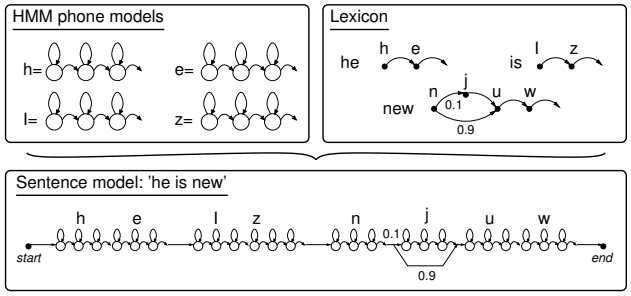

Figure 2: A model for a complete sentence made by concatenating the appropriate phoneme models.

2.1. Viterbi Segmentation

Traditionally, the Viterbi algorithm is used to find the single best path through the model given the observed speech signal x_1^T (the sequence of feature vectors corresponding to the speech signal):

$$s_i^T = \arg\max_{s_i^T \subset S} \prod_{i=1}^{T} f(x_i \mid s_i) p(s_i \mid s_{i-1}),$$

with s_i^T a sequence of HMM states (one state for each time frame) that is consistent with the sentence model S, T being the number of time frames. Thus, the Viterbi algorithm results in the segmentation which reaches maximum likelihood for the given feature vectors.

2.2. Forward-Backward Segmentation

Unfortunately, Viterbi's maximum likelihood assignment is only sub-optimal in that it does not necessarily generate the boundary that is closest to the position where the boundary is expected given the speech signal and the acoustic models. That is, the Viterbi algorithm only provides us with an approximation of the quantity that is really looked for. This is illustrated in figure 3. The Viterbi algorithm generates the boundary corresponding to (1), whereas the optimal boundary corresponds to (2).

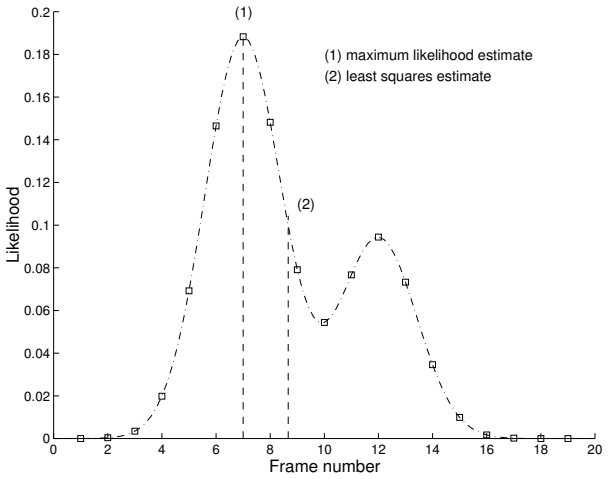

Figure 3: Maximum likelihood vs. least squares estimate for boundary.

To find the best possible estimate of the boundary in a least squares sense the probability function of each boundary must be calculated:

$$P(b|S, x_1^T) = \frac{f(x_1^b|S_l) f(x_{b+1}^T|S_r)}{f(x_1^T|S)},$$

with

$$f(x_a^b|S_x) = \sum_{s_a^b \subset S_x} \prod_{i=a}^{b} f(x_i|s_i)^{1/\beta} p(s_i|s_{i-1})^{1/\beta}.$$

In the above equations, sentence S (the sequence of HMM states) is divided in part S_l left and part S_r right of the boundary of interest. The extra parameter β compensates for the ill-matched assumption made by HMMs that the observations x_i are independent. The optimal value for β in our experiments was 10, but its exact value was not at all critical (a broad optimum). The same compensation factor—with approximately the same value—can be found in recognition systems (Demuynck, 2001) as well as in confidence scoring of recognized words (Wessel et al., 2001). In these two cases, its function is to balance the contributions of acoustic model and language model to the total likelihood of the sentence.

Calculating the density functions for all boundaries in a sentence can be done efficiently with the Forward-Backward algorithm. Given the probability density function of each boundary, the least squares estimate now equals:

$$E\{b\} = \sum_{b=1}^{T} P(b|S, x_1^T)\, b.$$

2.3. Confidence Intervals for Segmentation

Since the Forward-Backward algorithm computes the probability density function for each boundary, we can regard the variance of this function as a confidence interval for the respective boundary:

$$\mathrm{Var}(b) = \sum_{b=1}^{T} p(b \mid S, x_1^T)(b - E\{b\})^2.$$

We assume that these confidence intervals can be useful in several types of applications. For example, in TTS systems segments with small confidence intervals on both segment boundaries could be preferred in the segment selection. In the context of this paper's topic, we successfully incorporated the intervals in a post-processing procedure aimed at removing biases between automatic and manual segmentations.

The forementioned biases are dependent on the classes of the phonemes left and right of the boundary, and can be attributed to the fact that humans use different cues than HMMs for finding the boundary between consecutive phonemes (van Santen and Sproat, 1999). For the transition to a vowel, for example, the average difference between automatic and manual segmentation can be more than halved when compensating for these biases. An equally big improvement can be obtained for the transitions to noise. In section 3. we will show that optimal compensation for these biases can be calculated as a function of confidence intervals.

3. Experiments

3.1. Description

The automatic segmentation and the confidence intervals were evaluated on part of the read aloud data in the currently developed Spoken Dutch Corpus (CGN) (Oostdijk, 2000). Read aloud text accounts for the 'cleanest' speech within the corpus. The automatic segmentation started from manually created chunks (2 to 6 seconds of speech bounded by silence) provided by the corpus. In our experiments the test set consisted of 13958 words, resulting in 17774 boundaries since pauses exceeding 50 ms were also part of the segmentation.

We based the evaluation of the automatic word segmentations on a comparison with corresponding manual word segmentations. The choice for evaluating on word segmentations was motivated by the function of segmentations within the Spoken Dutch Corpus project: providing easy access to the words in the speech corpus.

First, the words in the test set were segmented manually by two persons. They were instructed to use audible cues only and to position boundaries so that each word would sound acoustically acceptable in isolation, i.e. could be played back without hearing (part of) the phonemes of the preceding or following word. Shared phonemes at the boundary (e.g. he is_sad) were split in the middle, except for shared plosives (e.g. stop_please), which were isolated altogether. Noticeable pauses ($>$ 50 ms) were segmented in the same way as words, thus producing empty chunks.

Then, the automatic segmentations were evaluated by counting the number of boundaries for which the deviation between automatic and manual segmentation exceeded thresholds of 35, 70 and 100 ms. To evaluate the confidence intervals, the number of non-detected deviations that exceeded 35, 70 and 100 ms were counted if 50% or 30% of the boundaries with the largest confidence intervals would have been checked manually.

no of deviations exceeding			**no of**
35 ms	**70 ms**	**100 ms**	**boundaries**
Viterbi segmentation (base segmentation)			
2184	552	229	17774
Forward-Backward segmentation			
2102	490	182	17774
rel. improvement w.r.t. base segmentation			
3.8%	11.2%	20.5%	17774

Table 1: Viterbi vs. Forward-Backward.

For the experiments we used the large vocabulary continuous speech recognition system developed by the ESAT-PSI speech group at the K.U.Leuven. The system's context-dependent acoustic models (partially tied gaussians) were estimated on a database with 6 hours of dictated speech in Flemish Dutch. The speakers in this database did not occur in the test data for the experiments. A detailed overview of the context-dependent acoustic modeling can be found in (Duchateau, 1998), the search module is described in (Demuynck, 2001; Demuynck et al., 2000).

3.2. Discussion

Table 1 compares Viterbi and Forward-Backward segmentations. It shows that Forward-Backward segmentation has a significant relative improvement over Viterbi segmentation. It is important to notice that especially the number of large errors ($>$ 100 ms) is reduced by more than 20%.

As explained in section 2.3., there are some (phoneme-dependent) biases between automatic and manual segmentations. We discerned 9 phoneme classes in total and analysed the biases on the boundary position between each pair of classes. Those biases in the Forward-Backward segmentations were removed in a post-processing step, as will be explained next.

In a first step, we shifted the boundaries purely on the basis of the average biases. As is shown in the top part of table 2, this again resulted in a more accurate segmentation.In a second step, we compensated for the biases in a more advanced way. The bottom part of table 2 shows results for shifting the boundary while taking into account the confidence interval for this boundary. More precisely, the bias was estimated as a function on the boundary's confidence interval. This function was determined empirically with a polynomial fit on a test set. The improvement shows that confidence intervals are useful when determining the optimal boundary shift: large intervals (large variances) typically correspond to large shifts.

The majority of the remaining random (or hard to predict) deviations in the automatic post-processed segmentations are transitions to and from noise and transitions to unvoiced plosives (45%, 11% and 15% of the remaining 35 ms errors respectively). Since these boundaries also show large variation between the corresponding manual segmentations of different correctors, we cannot expect an auto-

no of deviations exceeding			no of
35 ms	70 ms	100 ms	boundaries
after post-processing, excl. conf. intervals			
1969	390	163	17774
rel. improvement w.r.t. base segm., excl. conf. intervals			
9.8%	29.3%	28.8%	17774
after post-processing, incl. conf. intervals			
1928	359	147	17774
rel. improvement w.r.t. base segm., incl. conf. intervals			
11.7%	35.0%	35.8%	17774

Table 2: Results after post-processing.

confidence level	no of non-detected deviations exceeding			no of boundaries
	35 ms	70 ms	100 ms	
confidence intervals				
50%	439	58	11	8887
30%	805	108	23	5332

Table 3: Evaluation of confidence intervals.

matic system to give more consistent results.

Finally, table 3 evaluates the confidence intervals in the way described in section 3.1.. As such, these results are not yet good enough to speed up the work of a manual segmenter by flagging only the least probable boundaries. Even if 50% of the boundaries were to be verified, 7% of the 100 ms deviations and 23% of the 35 ms deviations would still be overlooked. Yet, as shown above, the confidence intervals are accurate enough for advanced post-processing.

4. Conclusions and Future Research

We presented an improved algorithm for the automatic segmentation of speech corpora based on the Forward-Backward algorithm. We found that automatic Forward-Backward segmentation produces more accurate results than the more traditionally used Viterbi segmentation. Moreover, the variances of the probability functions (calculated for each boundary by the Forward-Backward algorithm) can be considered confidence intervals on the respective boundaries. Experiments showed that the confidence intervals are useful when compensating for the biases between HMM-based and manual segmentations, thus improving the quality of automatic segmentations even further. Since the post-processed Forward-Backward word segmentations clearly outperform Viterbi word segmentations, we plan to adopt them as word segmentations in the creation of the Flemish part of the Spoken Dutch Corpus, both as the basis for manual verification (performed on part of the data) and as the automatic word segmentations for the complete Flemish corpus part.

Future research will focus on the evaluation of the Forward-Backward segmentation algorithm on 'more challenging' parts of the Spoken Dutch Corpus (containing background noise, overlapping voices, . . .). In addition, research will be conducted on the development of advanced methods for the production of reliable automatic segmentations based on automatically generated phonemic transcriptions (derived from the orthographic transcription). Finally, we will look into a further optimisation and application of confidence intervals for segmentation with the aim of considerably speeding up the segmenter's manual task.

5. Acknowledgements

This publication was supported by the project "Spoken Dutch Corpus" (CGN-project) which is funded by the Flemish Government and the Netherlands Organisation for Scientific Research (NWO).

6. References

N. Beringer and F. Schiel. 1999. Independent automatic segmentation of speech by pronunciation modeling. In *Proc. International Congress of Phonetic Sciences*, pages 1653–1656, San Francisco, U.S.A., August.

K. Demuynck, J. Duchateau, D. Van Compernolle, and P. Wambacq. 2000. An efficient search space representation for large vocabulary continuous speech recognition. *Speech Communication*, 30(1):37–53, January.

K. Demuynck. 2001. *Extracting, Modelling and Combining Information in Speech Recognition*. Ph.D. thesis, K.U.Leuven, ESAT, February. Available from http://www.esat.kuleuven.ac.be/∼spch.

J. Duchateau. 1998. *HMM Based Acoustic Modelling in Large Vocabulary Speech Recognition*. Ph.D. thesis, K.U.Leuven, ESAT, November. Available from http://www.esat.kuleuven.ac.be/∼spch.

J. Husson. 1999. Evaluation of a segmentation system based on multi-level lattices. In *Proc. European Conference on Speech Communication and Technology*, volume I, pages 471–474, Budapest, Hungary, September.

T. Laureys, K. Demuynck, J. Duchateau, P. Wambacq, and A. Bogan-Marta. 2001. Assessing segmentations: Two methods for confidence scoring automatic HMM-based word segmentations. In *Proc. of the 6th International Conference on Engineering of Modern Electric Systems*, pages 116–121, Oradea, Romania, May.

A. Ljolje and M. Riley. 1991. Automatic segmentation and labeling of speech. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 473–476, Toronto, Canada, May.

N. Oostdijk. 2000. The Spoken Dutch Corpus. *The ELRA Newsletter*, 5(2):4–8. Available from http://lands.let.kun.nl/cgn/home.htm.

J. van Santen and R. Sproat. 1999. High-accuracy automatic segmentation. In *Proc. European Conference on Speech Communication and Technology*, volume VI, pages 2809–2812, Budapest, Hungary, September.

A. Vorstermans, J. Martens, and B. Van Coile. 1996. Automatic segmentation and labelling of multi-lingual speech data. *Speech Communication*, 19(4):271–293, October.

F. Wessel, Schlüter R., K. Macherey, and H. Ney. 2001. Confidence measures for large vocabulary speech recognition. *IEEE Transactions on Speech and Audio Processing*, 9(3):288–298, March.