Assessing Segmentations: Two Methods for Confidence Scoring Automatic HMM-Based Word Segmentations[1]

Tom Laureys, Kris Demuynck, Jacques Duchateau, Patrick Wambacq
Katholieke Universiteit Leuven - ESAT
Kasteelpark Arenberg 10 3001 Leuven, Belgium
{tom.laureys, kris.demuynck, jacques.duchateau, patrick.wambacq}@esat.kuleuven.ac.be

Alina Bogan-Marta
University of Oradea – Department of Computers
Armatei Romane 5 3700 Oradea, Romania
alinab@univ.uoradea.ro

Abstract

The Dutch-Flemish project Spoken Dutch Corpus (1998-2003) aims at the development of an annotated corpus of 10 million spoken words. In order to make the speech data easily accessible, a word segmentation couples the orthographic transcription to the speech signal by means of time stamps. Generally, such segmentations are produced manually. Since this manual procedure is a time-consuming effort, we developed an automatic segmentation tool and two methods for assigning confidence scores to the automatically generated boundaries. This way we aim to speed up the manual work considerably.

1. Introduction

In this paper we present a tool used in the development of the Spoken Dutch Corpus, an annotated corpus of 10 million spoken words [5]. For a corpus of speech data to be useful and easily accessible, it is necessary to know (1) *what* words are spoken (orthographic transcription) and (2) *when* each word is spoken. Segmenting the speech data at the word level fulfills the latter requirement: time stamps indicate the onset and end of each uttered word. This makes the retrieval of any desired spoken word in the database straightforward.

Generally, a time-consuming manual procedure is required to produce high-quality segmentations. We describe a tool aimed at reducing the manual effort by the application of algorithms based on ASR (Automatic Speech Recognition). Our tool has two objectives: (1) automatically generate word segmentations for speech data, and (2) put confidence scores on the automatically generated word boundaries. This way we plan to speed up the manual work for the aligner, whose task now comes down to checking only those automatically generated boundaries with low confidence scores.

In the literature several other systems to automatically generate speech data segmentations have been described. Most of them have been applied to databases for TTS

[1] This publication was supported by the Project Spoken Dutch Corpus (CGN), which is funded by the Flemish Government and the Netherlands Organisation for Scientific Research (NWO).

(Text To Speech) systems, thus generating segmentations at the phoneme level. Yet, this is not very different from our tool, which produces word level segmentations. In fact, at an intermediate stage our tool also puts a segment boundary between any two phonemes, but given our specific application only the boundaries between words are eventually retained.

Some of the methods described in the literature are based on specific acoustic cues or features for the segmentation task [3,6,7], focusing for instance on transient behaviour or specific differences between phoneme classes. Others use general features and acoustic modelling which are common in ASR [1,4]. The method we present is of the latter type.

Most methods for the segmentation of speech data rely on a phonetic transcription. This phonetic transcription may be generated manually (as is the case in our experiments) or can be automatically derived from the orthographic transcription and a phonetic dictionary.

This paper focuses on an improved algorithm for determining the position of the word boundary and on the assignment of a confidence score to each word boundary. The confidence scoring of segmentations is novel and can be useful for the type of application we are working on as well as for TTS systems (eg. segments with a high confidence score on both segment boundaries could be preferred in the segment selection).

The paper is organised as follows. In section 2, we present the automatic HMM-based segmentation tool. The two confidence scoring methods are described in section 3, followed by a discussion of the experiments in section 4. We end with conclusions and suggestions for future research.

2. Automatic segmentation of speech

The automatic segmentation system is based on a two-step process. First, the phones in the input[2] are coupled to their respective acoustic Hidden Markov Models (HMMs). Then, the Viterbi algorithm finds the best assignment of speech data to the acoustic model of the complete phone sequence.

HMMs describe the basic units of speech (the phones) as a sequence of states (figure 1). The acoustic properties of the states s_i are modelled by means of observation density functions $f_i(y) = f_{Y|s_i}(y)$, y being the feature vector that describes a given speech frame at 10 msec intervals. The duration and possible order of the states is governed by the transition probabilities between the states $a_{ij} = P(s^{(t+1)} = s_j \mid s^{(t)} = s_i)$, $s^{(t)}$ being the HMM state at moment t. HMM phone models typically have three states and a simple left-to-right topology as illustrated in figure 1.

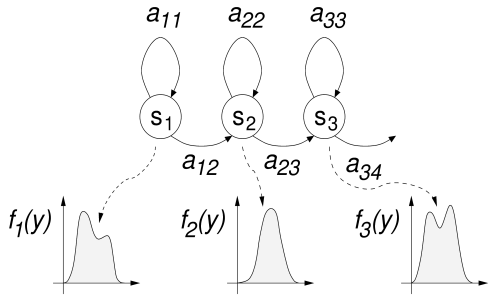

Figure 1: A 3-state HMM with a left-to-right topology.

[2] Remember that our segmentation is based on a (manual) phonetic transcription.

Once the acoustic properties of the different phones have been encoded in statistical models, sentence models are generated by concatenating all relevant phone models (figure 2). Next, the Viterbi algorithm is used to find the best path through the model given the observed speech signal (the sequence of feature vectors corresponding to the speech signal). Based on this path, the best assignment of speech to the different states can be derived, while still adhering to the left-to-right constraints of the model.

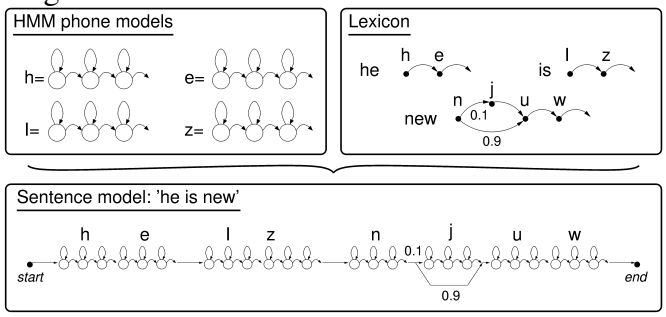

Figure 2: A model for a complete sentence made by concatenating the appropriate phone models.

3. Two methods for confidence scoring

3.1. Method 1: acoustic match of the surrounding phones

Since the Viterbi algorithm searches for the best assignment of speech frames to the different states in the sentence model, serious mis-alignments only occur if (1) the phonetic transcription of the sentence contains inaccuracies, or if (2) the acoustic models do not adequately model certain speech effects. An example of the first type of errors is the occurrence of non-speech events that are not adequately reflected in the phonetic transcription (eg. coughing, laughing). The main cause of errors of the second type is the minimal duration constraint of the phones (30 msec) imposed by the simple 3-state left-to-right HMM. This may particularly lead to a mis-alignment when a sequence of phones is uttered very fast.

In both cases, we expect that the data immediately left and right of the hypothesized word boundary will not match well with the corresponding boundary phones (states). To detect these types of errors the normalised acoustic (log) likelihood of the speech signal given the surrounding phones is calculated:

$$\log(\ell(Y \mid t,l,r,s_i)) = \frac{1}{l+r}\sum_{i=t-l}^{t+r-1}\log\left(\frac{f(y_i \mid s_i)}{f(y_i)}\right),$$

$$f(y_i) = \sum_s f(y_i \mid s)P(s),$$

with l and r the duration (number of frames) of the phones left and right of the boundary respectively, s_i the best matching state (according to the Viterbi alignment) for frame i, $P(s)$ the a-priori probability of the state s, and t the boundary (a frame number) for which the confidence score is calculated.

In our experiments a window of 1 phone left and 1 phone right yielded the best results, but the above formula can be easily modified to incorporate more phones on either side. Since the conditional density functions $f(y_i \mid s)$ are weighted sums of Gaussian mixtures, the unconditional density $f(y_i)$ is a weighted sum of Gaussians as well. A method for a fast evaluation of both the conditional and unconditional densities can be found in [2].

3.2. Method 2: probability changes when moving segment boundary

The second method is based directly on the statistical framework used in speech recognition. For every boundary to be checked, the change in the likelihood of the best possible path (Viterbi alignment) for the observed speech signal is calculated in function of the boundary position t:

$$\log(\ell(Y \mid t,s)) = \sum_{i=0}^{t-1} \log(f(y_i \mid s_i^l)) + \sum_{i=t}^{N} \log(f(y_i \mid s_i^r)),$$

with N the total number of frames in the sentence, and $s_i^{l/r}$ the best matching state (according to the Viterbi alignment) for frame i for the left/right part of the sentence, i.e. before/after the boundary for which the confidence score is calculated.

These functions can be efficiently computed by the forward-backward algorithm, which is also used for training the HMM-parameters in speech recognition. Once these curves are known for all word boundaries, the probability of having the boundary between a given pair of words at frame t can be calculated as follows:

$$P(T = t \mid s,Y) = \frac{\ell(Y \mid t,s)^{1/\beta}}{\sum_{t'} \ell(Y \mid t',s)^{1/\beta}}$$

The fuzzy factor β compensates for the ill-matched assumption made by HMMs that the observations y_i are independent. The same compensation factor can be found in the confidence scoring of recognized words [8]. In both cases the optimal value for β is somewhere between 10 and 20. Based on this boundary density function $P(t \mid s,Y)$, the variance on the boundary positions can be calculated as well, which provides us with another potential confidence measure.

The position t at which the above density function has its maximum corresponds to the boundary found by the Viterbi algorithm. Since this maximum position is rather arbitrary when the density function has no clear peak, one could argue that the mean value $\sum_t t \cdot P(t \mid s,Y)$ is a better estimate for the real boundary position.

4. Experiments

The automatic alignment and the confidence scores were evaluated on part of the read aloud data in the Spoken Dutch Corpus. Since read aloud text accounts for the 'cleanest' speech within the corpus, it was obviously the best starting point for testing and comparing our segmentation and confidence scoring techniques. In our experiments, the test set consisted of 13958 words in 2544 chunks[3].[4]

The words in the test set were first aligned manually by two aligners. They were instructed to use audible cues only and to position boundaries so that each aligned word would sound acoustically acceptable in isolation, i.e. could be played back without hearing (part of) the phones of the preceding or following word. Shared phonemes at the boundary (e.g. he i*s_s*ad) were split in the middle, except for shared plosives (e.g. sto*p_p*lease), which were isolated altogether. Noticeable pauses (> 50 msec) were aligned in the same way as words, thus producing empty chunks. Based on a subset of the data (3155 words) manually aligned

[3] A chunk is a sequence of words surrounded by longer pauses (> 1 sec). Chunks typically have a length of about 3 seconds and are put in by orthographic and phonetic transcribers to have a first rough segmentation of the data.

[4] Note that for our experiments the acoustic models for the segmentation were estimated on a *different* acoustic database consisting of *different* speakers.

by both aligners, we found that in 90% of the cases mutual inter-aligner deviations were less than 35 msec. Yet, for transitions to or from silence the 90% limit on the deviations between the two manual alignments doubled to 70 msec.

We based the evaluation of the automatic alignments on a comparison with the corresponding manual alignments. More specifically, the number of boundaries for which the deviation between automatic and manual alignment exceeded 35, 70 and 100 msec were counted (top part of table 1). To evaluate the confidence scores, the number of non-detected deviations that exceeded 35, 70 and 100 msec were counted if 50%, respectively 30% of the boundaries with the lowest confidence score were to be checked manually (bottom part of table 1).

confidence level	**nr. of times the time deviation exceeds**			**total nr. of boundaries**
	35 msec	**70 msec**	**100 msec**	
automatic alignment				
/	2184	552	229	17774
confidence scoring: acoustic likelihood				
50%	840	132	32	8887
30%	1234	213	60	5332
confidence scoring: P(T=t|s,Y)				
50%	531	86	16	8887
30%	910	147	36	5332
confidence scoring: variance(t; ℓ (Y | t, s))				
50%	520	75	11	8887
30%	923	128	24	5332

Table 1: The results of the automatic alignment and confidence scores.

Table 2 shows the results when the mean position of a word boundary is used instead of the best position.

confidence level	**nr. of times the time deviation exceeds**			**total nr. of boundaries**
	35 msec	**70 msec**	**100 msec**	
automatic alignment (mean boundary positions instead of best)				
/	2102	490	182	17774
confidence scoring: variance(t; ℓ (Y | t, s))				
50%	439	58	11	8887
30%	805	108	23	5332

Table 2: The results of the corrected automatic alignment and confidence scores.

Human inter-aligner agreement is higher than the agreement between automatic and manual alignments. Moreover, manual verification showed that some of the deviations in the automatic alignment no longer conform to the objectives (acoustic acceptability of words in isolation). So manual verification is still necessary.

Further, at this point the precision and recall figures of the confidence measures, even for the best method, are not yet good enough to speed up the manual verification by marking (and limiting the aligner's verification task to) the least probable boundaries. Yet, starting from a good automatic alignment (even without highly accurate confidence scores) still significantly reduces the amount of manual work.

A further analysis brought forward some predictable differences (biases) between the automatic alignment and the manual alignment. A first bias is due to the algorithms for extracting features from the speech signal. These features use time derivatives and thus include information on how the signal will change in the future, causing a fixed shift in the boundary position of approximately 10 msec. In both table 1 and table 2, this bias was

removed. The remaining biases (not removed in the tables) are dependent on the phonetic classes of the phones 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 phones [6]. For the transition to a vowel, for example, the average difference between automatic and manual alignment can be more than halved when compensating for these biases. An equally big improvement can be obtained for the transitions to noise. Our analysis also showed that the confidence scores, especially the variance measure proposed in 3.2, are useful when determining the optimal shift of the boundary: large variances typically correspond to large shifts.

Apart from the biases mentioned above, the majority of the remaining random (or hard to predict) deviations are transitions to and from noise and transitions to unvoiced plosives (45%, 11% and 15% of the remaining 35 msec errors respectively). Since these boundaries also show large inter-aligner variation, we cannot expect an automatic system to give more consistent results.

5. Conclusions and future work

In this paper we presented a tool for the automatic segmentation of speech and two methods for confidence scoring such segmentations. We found that providing human aligners with an accurate automatic base segmentation significantly reduces the amount of manual work. Further, our primary motivation for generating confidence scores, namely reducing the manual work to a verification of only the least probable boundaries, has not been fulfilled so far. First results are nevertheless promising and we think there are opportunities for further research. On the other hand, current confidence measures generated by method 2 (section 3.2) have already turned out to be useful when compensating for the biases between HMM-based and manual segmentations.

In the future, we will focus our research on further optimisation of our confidence scoring methods, on the implementation of accurate post-processing tools for the automatic segmentation (removing biases), and on the application of the proposed techniques to other subcomponents of the Spoken Dutch Corpus (face-to-face conversations, broadcast material, …) which are generally considered harder to process automatically.

References

[1] Beringer, N. and Schiel, F.: "Independent Automatic Segmentation of Speech by Pronunciation Modeling", Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS-99), 1999, 1653-1656.

[2] Demuynck, K.: “Extracting, Modelling and Combining Information in Speech Recognition”, PhD thesis ESAT-KULeuven, 2001.

[3] Husson, J.: "Evaluation of a Segmentation System based on Multi-Level Lattices", Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech-99), 1999, Volume I: 471-474.

[4] Ljolje, A. and Riley, M.: "Automatic Segmentation and Labeling of Speech", Proceedings of the 16th IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP-91), 1991, Volume I: 473-476.

[5] Oostdijk, N.: “The Spoken Dutch Corpus”, The ELRA Newsletter, 5-2 (2000), 4-8.

[6] van Santen, J. and Sproat, R.: "High-Accuracy Automatic Segmentation", Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech-99), 1999, Volume VI: 2809-2912.

[7] Vorstermans, A., Martens, J. and Van Coile, B.: "Automatic Segmentation and Labelling of Multi-Lingual Speech Data", Speech Communication, 19-4 (1996), 271-293.

[8] Wessel, F., Schlüter, R., Macherey, K. and Ney, H.: “Confidence Measures for Large Vocabulary Speech Recognition”, IEEE Transactions on Speech and Audio Processing, 9-3 (2001), 288-298.