Comparing Segmentation Strategies for Efficient Video Passage Retrieval

Christian Wartena
Hochschule Hannover - University of Applied Sciences and Arts
Department of Information and Communication
Expo Plaza 12, 30539 Hannover, Germany
Christian.Wartena@fh-hannover.de

Abstract

We compare the effect of different text segmentation strategies on speech based passage retrieval of video. Passage retrieval has mainly been studied to improve document retrieval and to enable question answering. In these domains best results were obtained using passages defined by the paragraph structure of the source documents or by using arbitrary overlapping passages. For the retrieval of relevant passages in a video, using speech transcripts, no author defined segmentation is available. We compare retrieval results from 4 different types of segments based on the speech channel of the video: fixed length segments, a sliding window, semantically coherent segments and prosodic segments. We evaluated the methods on the corpus of the MediaEval 2011 Rich Speech Retrieval task. Our main conclusion is that the retrieval results highly depend on the right choice for the segment length. However, results using the segmentation into semantically coherent parts depend much less on the segment length. Especially, the quality of fixed length and sliding window segmentation drops fast when the segment length increases, while quality of the semantically coherent segments is much more stable. Thus, if coherent segments are defined, longer segments can be used and consequently less segments have to be considered at retrieval time.

1. Introduction

Video content represents a fast growing part of the total amount of internet content. Audio-visual content is not restricted to entertainment, but includes also video lectures, instructional videos, interviews, documentaries and so on. Users in many cases do not watch these videos linearly but watch just selected fragments ([17]). Thus we need methods to browse and search within a video ([18]) and to find the relevant parts of a video. The retrieval of the relevant fragments or jump-in points is called passage retrieval. Passage retrieval raises a number of interesting questions like the relation between the passages and the video as a whole and the questions of determining the right segment boundaries and the relevant segments given some information need. It is this latter question that we address in this paper.

For passage retrieval of written texts often the formatting of the text, as defined by the author is used as a base for defining passages ([3, 5, 15]). Either the chapters or paragraphs are used directly as retrieval units or paragraphs are merged into larger units. Adjacent paragraphs are merged either because their topics are very similar or in other approaches simply in order to define passages of more or less constant length. For segmentation of video this typ of formatting information usually is not available.

For retrieval and browsing video lectures usually the accompanying slides or textbook are used as sole or additional source to segment the video stream (see e.g. [20, 8, 12]). In general, however, we do not have slides and slide transitions and we have to rely on the information that can directly be obtained form the audio and the video signal. In the following we will focus on segmentation and retrieval based on the audio signal, especially the (automatic) speech transcripts.

There are basically three possibilities for the segmentation of a speech transcript: (1) using prosodic information, like intonation, pauses and speaker turns; (2) using advanced segmentation techniques to find lexically (and semantically) coherent segments; or (3) using segments of fixed length. For the fixed length segments there are two variants. Either the document is split up in segments of (almost) equal length, or a sliding window is used, defining many overlapping segments. Probably by the lack of good evaluation corpora, we find hardly any literature comparing the effectiveness of these methods for passage retrieval directly. Indirectly, passage retrieval is evaluated by its use for question answering and for improving document retrieval. Results in these domains suggest that the sliding window strategy performs best.

In the current study we compare the four segmentation strategies for passage retrieval on a video corpus of the Me-

diaEval 2011 Rich Speech Retrieval task ([7]). Thus we do not evaluate the quality of the segments directly, but indirectly by their usefulness in a passage retrieval task. In fact we even don't consider the whole segment for evaluation but only evaluate the start time of the segment, as as it is the user himself who decides to pursue the listening or not.

We find that fixed length (either overlapping or non-overlapping) passages give best results. However, these results depend on the right estimation of the optimal segment length. Results of the coherent segment strategy are almost as good, but depend less on the choice of the correct parameter. Especially, the results remain stable for longer segments. Thus, the coherent segments give the possibility to work with longer and consequently less segments. In the present paper we focus on a quantitative analysis of the effects of segment length for a number of segmentation strategies. A more qualitative study analyzing different aspects of the text and the segments can be found in [1].

The remainder of the paper is organized as follows. In section 2 we discuss related work. In section 3 we introduce the data we have used to test the segmentation strategies. Subsequently we sketch our general approach for passage retrieval. In section 4 we discuss the compared segmentation methods in more detail and show how they are applied to the used data set. Section 5 gives the results of the experiment. We finish the paper with a discussion and outlook for future work.

2. Related Work

Segmentation of spontaneous or planned speech has been studied mainly for lecture videos. The quality of the segmentation of these videos is usually accessed by a comparison with an available ground truth ([20, 8, 12]). We are not aware of any evaluation of segmentation strategies in the context of passage retrieval for this type of data. Moreover most research on lecture video segmentation uses additional sources of information.

Passage retrieval for written text has received a lot of attention. However, in most work passage retrieval is used to improve document retrieval ([3],[5],[10]), to improve query expansion ([19]) or it is used as a intermediate result for question answering ([13],[15]).

Hearst and Plaunt ([3]) introduce the text tiling algorithm that defines lexically coherent segments. They base document retrieval in various ways on passage retrieval. They report that the text tiling strategy outperforms fixed lengths segmentation. However, no significant differences are found with retrieval results based on the paragraph structure of the documents. Kaszkiel and Zobel ([5]) also compare effectiveness of different segmentation strategies for document retrieval based on passage retrieval. They introduce a further segmentation strategy with overlapping segments, that does not only use a sliding window but also considers windows of different size. Kaszkiel and Zobel find that this strategy, that they call arbitrary segmentation, gives best results. Arbitrary segmentation gives also best results for question answering in experiments described by Tiedemann ([15]). In this study no lexically coherent segmentation is evaluated. In another study Tiedemann and Mur ([16]) include text tiling but do not consider arbitrary segmentation. Now the sliding window supports the question answering task best.

The various studies all indicate that paragraph structure, if available, works very well. Best results are generally obtained with very flexible and redundant segmentation: the sliding window or arbitrary segments approach. Lexically coherent segments seem to have no advantages, but it should be noted that in all cases the text tiling algorithm ([2]) was used and that no variations of the granularity with which the algorithm should work were investigated. In the following we we use another algorithm to build lexically coherent segments in which the number of segments is an explicit parameter. Like for the fixed length segments and the sliding window we can then vary the (average) segment length.

3. Experimental Setup

3.1. The MediaEval Dataset

We carry out experiments on a corpus with Creative Commons content collected from blip.tv. This data set was used for multiple tasks in the MediaEval benchmark ([7]). The collection contains 1974 episodes (247 development and 1727 test) comprising a total of ca. 350 hours of data. We have used the development set to test our algorithms. In the following we will report only on the test set. The spoken channel is a mixture of planned and spontaneous speech. Each episode is accompanied by automatic speech recognition (ASR) transcripts provided by CNRS-LIMSI and Vocapia Research ([6]) and also by metadata (descriptions, title and tags), added by the uploader. In the following we focus on segmentation of ASR transcripts and do not use the metadata to improve the retrieval.

The 2011 Rich Speech Retrieval task provided 80 queries (30 development and 50 test), each with both full and short forms. The full form consists of a user description of the target video segment (e.g., 'This is a clip from a George Carlin special in which he comments on why he does not vote.' and 'Andrew Magloughlen talks how Google can help advance government tech.'). The short form consists of a query formulated to be directed at a general Web search engine (e.g., 'Voting Opinions' and 'Google government projects'). For our study of different segmentation methods we use only the short queries.

3.2. Approach

There are basically two approaches to passage retrieval ([13]): Either all possible passages are ranked directly, or initially documents are retrieved and subsequently the most relevant passages within these documents are searched. We use the first approach here.

Before segmentation and ranking all words are stemmed and stop words are removed. Mark Hepple's [4] part-of-speech (POS) tagger is used to tag and lemmatize all words. We remove all closed class words (i.e., prepositions, articles, auxiliaries, particles, etc.). To compensate for POS tagging errors, we additionally remove stop words (standard Lucene search engine stop word lists). Word and sentence segmentation, POS-tagging and term selection are implemented as a UIMA (http://uima.apache.org) analysis pipeline. The ASR-transcripts of the test set (1727 videos) contain approximately 3,07 million words. After filtering and stop word removal 1,27 million words remain. this roughly gives a rate of a bit more than 1 content word per second. The average length of a video is 1782 recognized or 735 content words.

For ranking we use BM25 [14]. Since fragments may overlap, we calculate idf (Eq. 1) on the basis of the sentence, the basic organizational unit of the speech channel,

$$\mathrm{idf}(t) \quad = \quad \log \frac{N - df_t + 0.5}{df_t + 0.5}. \qquad (1)$$

Here, N is the total number of fragments, and df_t is the number of sentences in which term t occurs. The weight of each term in each fragment-document is given by $w(d,t)$,

$$w(d,t) \quad = \quad \mathrm{idf}(t) \frac{(k+1) * f_{dt}}{f_{dt} + k * (1 - b + b * \frac{l_d}{\mathrm{avgdl}})}, \quad (2)$$

where f_{dt} is the number of occurrences of term t in document d, l_d is the length of d, and $avgdl$ is the average document length. In our experiments, we set $k = 2$ and $b = 0.75$, based on optimization of results on the development set. The retrieval status value (RSV) of a document for query consisting of more than one word is defined as,

$$w(d,Q) \quad = \quad \sum_{t \in Q} w(d,t). \qquad (3)$$

We create an initial ranking by ordering all fragments by their RSV values (Eq. 3). In order to generate our final results list, we remove all fragments with a starting time within a window of 60 seconds of a higher ranked fragment.

3.3. Evaluation Metric

Results are evaluated in terms of mean generalized average precision (mGAP) [9], which generalizes the calculation of the average precision of hypothesized jump-in points in relation to ground truth points by imposing a symmetric step-wise linearly decaying penalty function within a window of tolerance. In the following we use a 60s tolerance window and a 10s granularity step used for counting the penalty for the distance from the actual jump-in point within the window. Since RSR is a known-item task, the metric is effectively a mean generalized reciprocal rank (mGRR).

4. Segmentation

If no additional information is available, like written plots, accompanying slides, etc., we can distinguish four basic ways to segment the speech transcripts. Segmentation can be based on prosodic features like intonation and pauses. The simplest method to segment the transcript is by simply splitting it up into segments of equal length. This method also has the advantage that the subsequent ranking algorithm does not have to deal with problems arising from length differences. A variant of this fixed length method uses overlapping segments. Finally more advance methods can be used that try to identify lexically and semantically coherent segments. In the following we will present the exact variants of these approaches that we have used.

For all used methods we have to determine the length of the segments or the number of segments for a video. For the retrieval task as described above long segments clearly have two disadvantages: longer segments have a higher risk of covering several subtopics and thus give a lower score on each of the included subtopics. In the second place, long segments run the risk that they include the relevant fragment but that the beginning of the segment is nevertheless too far away from the jump-in point that should be found. Short segments on the other hand might get high rankings based on just a view words. Furthermore, short segments make the retrieval process more costly. The ideal length should be learned on a test set. Here we are however not interested in determining the optimal length, but rather in studying the behavior of the retrieval under changing lengths.

4.1. Prosodic Segmentation

The data set we have used (see section 3.1) is distributed with transcripts from automatic speech recognition. These transcripts are divided into fragments based on prosodic information. Apparently fragment boundaries are assumed at each speaker turn and at longer silences. Exact thresholds are not given. We use these fragments as retrieval units in the first condition. There are 77 878 fragments in the test set with an average length of 16.3 content words per fragment.

4.2. Fixed Length Segmentation

Before segmenting we lemmatize the speech transcripts and remove all stop words and non-content words (see section 3.2). We count the length of a passage in terms of content words rather than in terms of recognized words. The ASR-transcript also provides sentence boundaries. Since a sentence reasonably is the smallest unit to be retrieved, we respect these boundaries. Thus, the actual length of each segment might be a few words longer or shorter. The segmentation algorithm always chooses the sequence of sentences with smallest absolute difference between the actual and the targeted length.

The test set comprises 199 140 sentences. On average each sentence has 15 recognized words and 6.4 content words. Due to the discontinuities in spontaneous speech, a lot of very short sentences are hypothesized, resulting in a large variance in sentence length.

4.3. Sliding Window Segmentation

The sliding window method uses fixed length segments as well. The first segment is the same as in the fixed length approach. In order to find the next possible segment, the first sentence of the segment is removed, and one sentence at the end is added. If this new segment is longer than the target length, more sentences at the beginning are removed as long as the absolute difference with the target length decreases. If the segment is too short, in the same manner more sentences are added at the end. In case the target length is close to the average sentence length, the sliding window segmentation becomes almost the same as the fixed length segmentation.

4.4. Lexically Coherent Segmentation

The fixed length segments do not take into account the structure of the video. Ideally a segmentation corresponds to rhetorical or topical structure of the video. Such a segmentation then could give better results than fixed length segmentation if human annotators tend to choose the beginnings of these 'natural segments' as jump-in points.

A lot of research has been done into automatic segmentation of texts and speech transcripts. The basic idea is always to find regions that are lexically (and hence semantically) coherent. Lexically coherent passages can be understood as passages with a vocabulary that is distinct from adjacent regions or distinct from the overall vocabulary of the text. These regions usually tend to correspond very well to regions with a distinct subtopic. However, if we have spontaneous informal speech with smooth transitions of subtopics it is not that evident that always really natural segments are found, especially if we try to find very short segments. Probably the most well-known method implementing this ideas is Hearst's text tiling algorithm ([2]).

The method for segmentation that we have used is the minimum cut model from Malioutov and Barzilay ([11]). This algorithm is based on sentence similarity. A cut has to be chosen, such that (length normalized) sum of the similarities between sentences to the right and to the left of the cut is minimal. If a text has to be split up in more than two segments the sum of the (normalized) cut values has to be minimal. In the original algorithm Malioutov and Barzilay do not use sentences but word sequences of fixed length. In our implementation we however stick to the sentences proposed by the speech recognizer. This raises the problem that a very short sentence between two long sentences is very likely to cause a break. To avoid a this effect we use a relatively strong smoothing of word frequencies between adjacent sentences. We use the smoothing as proposed by Malioutov and Barzilay that is defined as

$$\overline{s}_i \quad = \quad \Sigma_{j=i}^{i+k} e^{-\alpha(j-i)} s_j, \tag{4}$$

where s_i are vectors of tf.idf values, and α is the parameter that controls the degree of smoothing. In our experiments we have set $\alpha = 1$. For the computation of the tf.idf we use the document frequency of a word in the whole test set.

Moreover we want to avoid that short sentences of one or two words with one common word, are tight much stronger together than long sentences with much more words in common. Thus we do not use cosine similarity but use the inner product of the tf-idf vectors as a similarity measure.

In order to speed up the segmentation we do not consider segments shorter than two sentences and segments longer than half of the whole video. The algorithm finds an optimal segmentation into a given number of segments. Thus the number of segments (or equivalently the average segment length) is a parameter of the algorithm like in the fixed length and sliding window segmentation. However, we always split each video up in at least three segments. The segmentation algorithm cannot do anything useful if the targeted segment length becomes to small. Thus, we did not use this method for very short segments.

5. Results

The retrieval results for the 50 questions from the MediaEval 2011 Rich Speech Retrieval task for all four segmentation methods and for varying segment lengths are given in Figure 1. The results are given for the actual average segment length, not for the targeted segment length. Segment lengths vary slightly because segments have to begin and end at sentence boundaries. For the longer fixed length segments the average length is strongly influenced by the last segment of each video, that often is much smaller that the other segments.

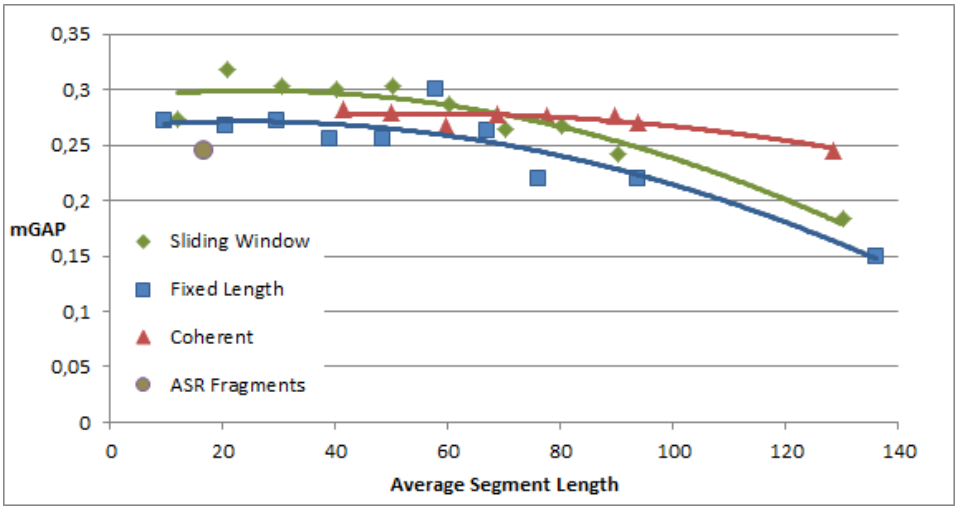

Figure 1. Results (observed values and trend lines) for the MediaEval2011 RSR task using 4 different segmentation methods with varying segment lengths.

Most pairwise differences are not significant. We used the Wilcoxon signed-rank test to compute significance. At the significance level of 0.05 the difference between the sliding window and the fixed length segments for length of 20 words is significant as well as the difference between the coherent and the fixed length segments for the longest segment length (130 and 136, resp.). Also the drop of mGAP between the values at e.g. 30 and 50 of the fixed length segments to the low values in the end is highly significant. The same holds for the sliding window. In contrast the pairwise differences between the results of the coherent segment strategy are not even significant at the level of 0.1.

For the fixed length and lexically coherent segmentation the number of segments that has to be ranked directly corresponds to the segment length. The range is for the fixed length segments is from 130 496 to 9 325. For the sliding window the number of segments ranges from 130 496 for the shortest segments to 99 865 for the longest segments.

6. Discussion

The results obtained with the fragments from the ASR are reasonable, but not as good as those achieved with the other segmentation strategies. The best results are obtained using the sliding window. However, for most segment lengths the differences are not significant and the price in terms of number of segments that has to be considered for retrieval is high. The results of the fixed lengths segmentation is almost as good, but seems to be very sensitive to the exact value of the segment length. The problem here is, that in a number of cases a passage containing the correct jump-in point is ranked very high, but that the beginning of the passage is too far away from this jump in point. The beginning of the passage is not determined by a change of vocabulary (like for the lexically coherent segments) nor by an optimal match (like for the sliding window), but by a rigid division into equal length segments. The segmentation method using the minimum cut model of Malioutov and Barzilay gives also similar results for segments up to a length of about 80 content words (i.e. about 70 seconds). This method does not have the disadvantages of the other methods: it seems much less dependent on the the exact value of the segment length and it does not leave all the labor to the ranking algorithm. For the longer segments we see that the results obtained with the lexically coherent segments are also more stable and do not drop as fast as the fixed length and sliding window segments. Thus this segmentation strategy allows to work with much less segments for retrieval.

Our results suggest that there is a clear advantage of using sophisticated segmentation methods for passage retrieval. This is somewhat surprising, since in most research on passage retrieval the advanced segmentation methods did not significantly perform better than other methods. As noted before most research on passage retrieval was done to improve document retrieval. In those studies the results are evaluated by the relevance of the retrieved documents only. A correct prediction of the jump-in point or of the relevant passage is not necessary. It seems to be exactly for the matching of the jump-in points that the semantic segmentation performs better than the fixed length strategy.

In the present study we tested only one method for non-trivial segmentation. Also the used method could be improved by including information about relations between

words in the computation of sentence similarities: often passages are not coherent because the same words are used all over the passage, but because the words in the passage are related to each other. Thus further improvement of the results for this segmentation approach can be expected. For future work we also plan to include shot boundaries.

7. Conclusion

We have made some interesting first observations on the usefulness of different strategies for the segmentation of ASR transcripts for video passage retrieval. As more similar data sets will become available, more experiments should be done to substantiate these observations.

Having made this reservation, we found that the results of video passage retrieval do not depend very strongly on the segmentation strategy used. Nevertheless, the segmentation method based on finding lexically coherent segments has some clear advantages: In the first place retrieval results based on this segmentation strategy are not as dependent on the choice of the correct segment length as those using a fixed length segmentation strategy. In the second place it does not produce as many candidate segments for retrieval as the sliding window approach and finally it also gives reasonable results for longer segments.

References

[1] M. Eskevich, G. J. Jones, M. Larson, C. Wartena, R. Aly, T. Verschoor, and R. Ordelman. Comparing retrieval effectiveness of alternative content segmentation methods for internet video search. In *10th International Workshop on Content-Based Multimedia Indexing (CBMI)*, June 2012.

[2] M. Hearst. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 9–16, 1994.

[3] M. Hearst and C. Plaunt. Subtopic structuring for full-length document access. In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 59–68, 1993.

[4] M. Hepple. Independence and commitment: Assumptions for rapid training and execution of rule-based POS taggers. In *ACL*, 2000.

[5] M. Kaszkiel and J. Zobel. Effective ranking with arbitrary passages. *Journal of the American Society of Information Science*, 52(4):344–364, 2001.

[6] L. Lamel and J.-L. Gauvain. Speech processing for audio indexing. In *Advances in Natural Language Processing*, LNCS 5221, pages 4–15. Springer, 2008.

[7] M. Larson, M. Eskevich, R. Ordelman, C. Kofler, S. Schmiedeke, and G. Jones. Overview of MediaEval 2011 Rich Speech Retrieval Task and Genre Tagging Task. In *MediaEval 2011 Working Notes*, 2011.

[8] A. Li, H. und Dong. Hierarchical segmentation of presentation videos through visual and text analysis. In *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology*, 2006.

[9] B. Liu and D. W. Oard. One-sided measures for evaluating ranked retrieval effectiveness with spontaneous conversational speech. In *SIGIR'06*, pages 673–674, 2006.

[10] X. Liu and W. B. Croft. Passage retrieval based on language models. In *Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management (CIKM 2002), November 4-9, 2002*, pages 375–382, 2002.

[11] I. Malioutov and R. Barzilay. Minimum cut model for spoken lecture segmentation. In *Procceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL 2006)*. The Association for Computer Linguistics, 2006.

[12] S. Repp and C. Meinel. Segmentation of lecture videos based on spontaneous speech recognition. In *Proceedings of the 10th IEEE International Symposium on Multimedia*, 2008.

[13] I. Roberts and R. J. Gaizauskas. Evaluating passage retrieval approaches for question answering. In *Advances in Information Retrieval - Proceedings of the 26th European Conference on IR Research (ECIR 2004)*, volume 2997 of *Lecture Notes in Computer Science*, pages 72–84. Springer, 2004.

[14] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)*, pages 232–241, 1994.

[15] J. Tiedemann. Comparing document segmentation strategies for passage retrieval in question answering. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP07)*, 2007.

[16] J. Tiedemann and J. Mur. Simple is best: Experiments with different document segmentation strategies for passage retrieval. In *Proceedings of the 2nd workshop on Information Retrieval for Question Answeringa (IRQA '08)*, pages 17–25, 2008.

[17] Y. van Houten, J. G. Schuurman, and P. Verhagen. Video content foraging. In *Proceedings of the Third International Conference on Image and Video Retrieval (CIVR 2004)*, volume 3115 of *Lecture Notes in Computer Science*, pages 15–23. Springer, 2004.

[18] S. Whittaker, J. Hirschberg, J. Choi, D. Hindle, F. Pereira, and A. Singhal. SCAN: Designing and evaluating user interfaces to support retrieval from speech archives. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, pages 26–33, 1999.

[19] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, 2000.

[20] N. Yamamoto, J. Ogata, and Y. Ariki. Topic segmentation and retrieval system for lecture videos based on spontaneous speech recognition. In *Proceedings of the 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003*, 2003.