Spoken Term Detection on German Speech Data

Daniel Schneider, Joachim Köhler

Kurzdresse: Fraunhofer IAIS, 53754 Sankt Augustin
E-Mail: {daniel.schneider,joachim.koehler}@iais.fraunhofer.de
Web: http://mmprec.iais.fraunhofer.de/

Abstract

A growing community of speech researchers investigates new approaches to the Spoken Term Detection (STD) task. The goal of STD is to search for queries in spoken utterances without the constraints of classic keyword spotting or LVCSR approaches. In this contribution we introduce the challenges of the STD task, describe current state of the art approaches and present our system for German STD.

1 Introduction

The term *Spoken Term Detection (STD)* was coined by NIST in 2006 in the scope of the NIST STD evaluation campaign. According to NIST, STD focuses on "technologies that search vast, heterogeneous audio archives for occurrences of spoken terms"[1]. Possible application scenarios for STD range from document retrieval in large audiovisual archives to continuous media monitoring of complex TV and radio data. It contrasts to classic keyword spotting techniques like [18], as it is by definition an open-vocabulary task, i.e., the query terms are not known at indexing time. Vocabulary independence is a major requirement for many interesting applications, and it is especially useful in large corpora with heterogeneous content. The STD community investigates several topics related to this task:

- Using subword models to overcome the vocabulary dependence of classic speech recognizers.
- Applying error-compensation at indexing time to cope with high subword ASR error rates.
- Applying error-compensation at query time to overcome pronunciation variabilities and high subword ASR error rates.
- Fusion of word and subword results.

An important issue in all mentioned topics is the efficiency of the respective STD approach: as STD retrieval is supposed to be executed on demand by actual end users, it must operate in reasonable time even on very large corpora. Depending on the application scenario, different STD approaches can be suitable. While monitoring applications in the security domain might focus on recall, search systems for end users would require more precision-oriented systems. It should be noted that there is a clear boundary between STD and the Spoken Document Retrieval (SDR) community, which aims rather at retrieving complete documents which are relevant for a given query instead of finding exact matches of spoken keywords. In contrast to SDR approaches such as [7], which aim at expanding queries with related terms from the same topic, STD is considered to be topic- and domain-independent.

2 Approaches to STD

Baseline STD systems employ large vocabulary continuous speech recognition (LVCSR) for generating a word transcript, where the query can be searched on the word level. Many systems do not only store the 1-best output of the recognizer, but also competing hypotheses in the form of lattices [15] or word confusion networks [5]. While LVCSR has reached a high level of accuracy in many domains, it is obviously not the most suitable solution for STD due to its inherent dependency on a fixed recognition lexicon. This is a major source for search errors, as the system can never detect queries which contain an out-of-vocabulary (OOV) word. A popular approach to overcome this challenge is to apply subwords instead of words as the decoding unit, where the set of subword units is finite and known a priori [3, 6, 11]. Queries are then broken into subword sequences, which are searched in the subword output of the ASR decoder. Due to less constraining language models, subword systems typically suffer from lower ASR accuracy compared to word-based systems. Moreover, the subword representation of a query contains more (smaller) tokens that must be matched in the subword transcript. If only one of these tokens is incorrect, the matching will fail. As a remedy, error-tolerant indexing and retrieval approaches can be applied, which aim at increasing STD recall at reasonable precision. As in the case of words, lattices can be used to store competing subword sequences [12, 4]. At retrieval time, error-compensating algorithms can be applied to the subword transcript [17] or to the query [8] to cope with subword ASR errors and pronunciation variations. Combining the results from word and subword decoding into hybrid STD systems can further increase the overall retrieval performance [1].

3 German STD at Fraunhofer IAIS

We use a standard word-based LVCSR system as the baseline. In addition to the 1-best transcript, we also generate word lattices including competing recognition hypotheses. To approach the OOV problem, we employ a subword speech recognizer based on syllables in parallel, which are a viable subword unit for German subword STD [14]. In [9], we proposed an efficient approach to subword lattice retrieval of German data, where we exploit the fact that syllable frequencies are Zipf-distributed. We retrieve only lattices that contain the most infrequent query syllable, and start the lattice traversal from these so-called anchor syllables.

Lattices focus on coping with ASR errors, which might occur due to a mismatch between the test data and the acoustic or language models of the ASR decoder, or due to a tight beam during decoding. However, we recently observed that there is an additional error source in subword STD, which cannot be covered by lattice retrieval [10]: in

[1] http://www.itl.nist.gov/iad/mig//tests/std/

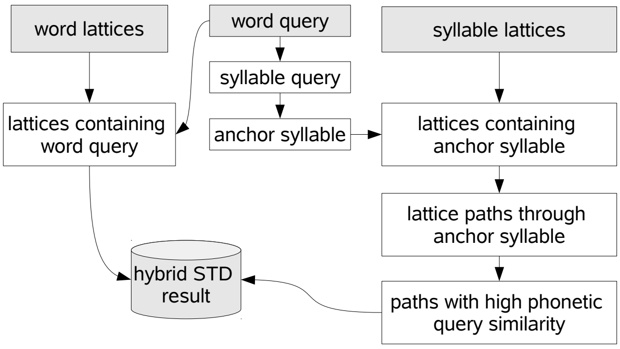
word lattices
word query
syllable query
anchor syllable
syllable lattices
lattices containing word query
lattices containing anchor syllable
lattice paths through anchor syllable
hybrid STD result
paths with high phonetic query similarity

Figure 1: Workflow for hybrid fuzzy lattice STD.

the case of subword decoding, the search algorithm must also be able to cope with pronunciation variations on the subword level. Pronunciation variation occurs if the actual spoken subword sequence differs from the canonical subword sequence provided by the grapheme-to-phoneme conversion, e.g., by syllable-final phone deletion. Even though the subword transcription is correct from an ASR point of view, it is likely that neither the 1-best transcript nor the lattice will contain the canonical transcription of a subword with a an uncommon pronunciation. For example, consider the German word *und* and its canonical syllabification *U_n_t_*. If the speaker omits the final *t_* phoneme, the correct subword ASR transcription is *U_n_*, hence an exact search for *und* will fail. Therefore it is reasonable to allow for such variations when retrieving from subword lattices. For each query we identify all lattices which contain the least frequent query syllable, which we assume to be non-mutated by pronunciation variation. Each path in each lattice that traverses that syllable is then matched against the query syllable sequence with a phonetic minimum edit-distance as described in [14]. If the score is above a threshold we accept the path as a hit. Finally, we merge the result sets from both word and subword retrieval into a hybrid result in order to maximize recall. Figure 1 illustrates the workflow for a full hybrid fuzzy lattice run.

4 Evaluation of STD systems

A wide range of evaluation metrics exist for assessing the quality of a STD system. While the standard metrics *recall* and *precision* are widely used in the information retrieval community, NIST has proposed additional STD metrics, which aim at removing the influence of individual query frequencies from the evaluation. Let Q be the set of actually occurring queries that shall be detected by the STD system. From the result set of a certain STD system, we can estimate the probability of missing a certain query $q \in Q$ with

$$p_{\text{miss}}(q) = 1 - \frac{TP(q)}{\text{reference occurrences of } q} \qquad (1)$$

where $TP(q)$ is the number of correct hits produced by the system for q. As defined by NIST in the STD evaluation plan, the probability that a given system produces a false alarm for a certain query q can be estimated with

Table 1: Evaluation Data Subsets. 'Other' includes segments with background noise and multiple characteristics.

Data Set	Utterances	Amount	Query Occurrences
plan_clean	1364	55 min	228
spont_clean	2861	115 min	378
dialect_clean	318	13 min	42
other	12609	521 min	1648
all	17152	704 min	2296

$$p_{\text{FA}}(q) = \frac{FA(q)}{\text{possible number of trials}} \qquad (2)$$

where $FA(q)$ is the number of false alarms produced by the STD system for q. The number of possible trials can be approximated with the total length of the corpus in seconds. Averaging over all terms we obtain two adjusted indicators for the two aspects *completeness* and *correctness*:

$$p_{\text{miss}} = \frac{1}{|Q|} \sum_{q \in Q} p_{\text{miss}}(q) \qquad (3)$$

$$p_{\text{FA}} = \frac{1}{|Q|} \sum_{q \in Q} p_{\text{FA}}(q) \qquad (4)$$

In order to obtain a single estimate for measuring the overall system performance, NIST proposed to use the *average term-weighted value ATWV*, which is estimated using

$$ATWV = 1 - \frac{1}{|Q|} \sum_{q \in Q} p_{\text{miss}}(q) + \beta \cdot p_{\text{FA}(q)} \qquad (5)$$

where the false alarm probabilty of a certain term is weighted with the constant cost β.

No standard evaluation data set exists for German STD on broadcast news data. For the results presented below, we used DiSCo, a new corpus for German Broadcast data, which was designed to reflect the acoustic characteristics of German TV programs [2]. The evaluation set consists of 15 hours of speech segments. Table 1 describes some interesting subsets which are used in the evaluation below. All listed subsets only show *one* characteristic, i.e., planned clean speech has no background noise and no dialect speech etc. Besides planned clean speech, we look at spontaneous clean speech, and clean speech from speakers with dialect.

A set of 501 queries was automatically generated, yielding 2736 query occurrences in the whole corpus. Typically, error-compensating STD approaches show poor performance when applied to short queries with a limited number of phonemes [13]. For the evaluation at hand, we focus only on queries with at least 6 phonemes, still yielding 2296 query occurrences in the data set. Figure 2 compares the relative query frequencies depending on the query length observed in the DiSCo corpus with the WDR/DW corpus, a data set with similar characteristics [14]. The graph illustrates that the query length distribution is similar across the corpora, and that only a few queries are below the chosen length limit.

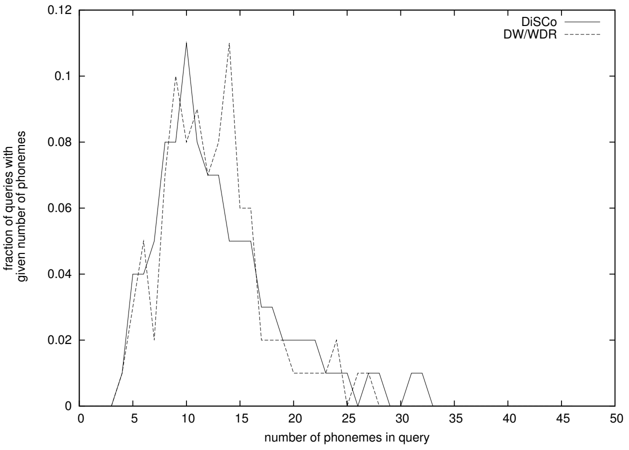

Figure 2: Relative query frequency, depending on number of phonemes in query.

5 Experimental Results

In this section, we present the experimental ASR and STD results of the system described in section 3 using the evaluation setup from section 4.

5.1 ASR results

The same ASR baseline as described in [10] was used to generate the 1-best transcripts and the word and syllable lattices. Both syllable and word ASR share the same acoustic phoneme models, which were trained with 80,000 utterances of manually transcribed broadcast speech. The word trigram language model was trained on a German newswire corpus of 158 million words, and a large 200k dictionary was used for word decoding. Syllable ASR was performed using a 10k syllable dictionary, and a 4-gram syllable language model trained on the syllabified version of the large newswire corpus. Table 2 shows the ASR results for the individual DiSCo subsets. All in all, the system produces an absolute difference of 12.1% in WER between the simplest subset and the average on the complete evaluation corpus. The results are in line with existing evaluations of the different challenges on English data [16]. The word and syllable decoding results show similar error rates across all subsets. The SER is slightly lower due to decompounding of reference compounds in the word ASR output (e.g., the compound *Fußballschiedsrichter* transcribed as *Fußball Schiedsrichter* yields a substitution and an insertion error during evaluation of the word ASR, while no errors occur on the syllable level). The results show that spontaneous and especially dialect speech present a major obstacle for the recognition system. This is not surprising, since these speech types often exhibit less clear pronunciations. Moreover, the system was trained mainly on planned non-dialect speech. The OOV rates are rather low for a German system, and they stem from the large word decoding lexicon.

Table 2: ASR results

Data Set	WER	SER	OOV rate
plan_clean	26.4	22.5	1.46 %
spont_clean	33.5	30.2	0.69 %
dialect_clean	51.2	50.0	1.56 %
all	38.5	34.9	1.38 %

5.2 STD results

STD systems typically have quite a number of parameters that can be configured, due to the numerous components involved. For the evaluation at hand, we decided to use a fixed set of lattices which are already highly pruned at indexing time based on the node posteriors, i.e., we did not exploit the node posterior at retrieval time as in [9], but assume that all paths in the pruned lattice are equally probable. Moreover, for the approximate search on lattice paths, we used the same fixed similarity threshold in all experiments, which was obtained in earlier experiments on a different development data set [14].

Table 3 compares the performance of various STD approaches to searching a classic 1-best word transcript. First we observe that by using the pruned word lattice instead of the simple 1-best transcription, the ATWV is increased by 0.04. A similar increase can be observed when moving from syllable 1-best to syllable lattice. Allowing for approximate matches on lattice paths further increases the ATWV.

All subword approaches have ATWV values below the word baseline. A reason is the higher length of subword queries: a subword query for a single query word with *n* syllables requires that a path with all *n* syllables is found by the system, i.e., the system has to compensate errors on all *n* syllables. Note however, that the subword results were obtained without any word level lexicon or language model. We combined the two approaches into hybrids, and give the results for two different configurations which can be useful in two different scenarios. The hybrid 1-best variant merges the result sets from word and syllable 1-best search, respectively. The advantage is that the approach can be easily realized with any standard text indexing system, and that retrieval efficiency is comparable to text retrieval, i.e., the approach scales to virtually any available corpus size. For example, extrapolating the observations from the DiSCo corpus, a corpus of 10,000 hours will contain about 100 million running words or 300 million syllables, which does not pose a problem for standard text retrieval systems. However, not all applications require such large data sets. This includes media monitoring, where only a subset of the data set is searched at a time, or smaller audiovisual databases. The results indicate that the hybrid lattice approach outperforms all other approaches in terms of retrieval accuracy, and it should be used in such scenarios.

Looking at the individual subsets, we observe that the additional gain from the approximate search depends on the complexity of the speech. While there is no increase in ATWV compared to the 1-best hybrid on planned speech, the fuzzy lattice hybrid outperforms both the word and 1-best hybrid baselines on spontaneous and dialect speech, where the deviation of the ASR result from the canonical transcription is higher.

Table 3: Comparison of various STD approaches on the complete corpus.

Setup	p_{miss}	ATWV
Word 1-best	0.40	0.59
Word Lattice	0.36	0.63
Syll 1-best	0.54	0.46
Syll Lattice	0.50	0.49
Syll Fuzzy Lattice	0.47	0.52
Hybrid 1-best	0.35	0.65
Hybrid Fuzzy Lattice	**0.30**	**0.68**

Table 4: Comparison of various STD approaches on subsets of the corpus.

Data Set	ATWV		
	Word 1-best	**Hybrid 1-best**	**Hybrid Fuzzy Lattice**
plan_clean	0.78	0.84	**0.84**
spont_clean	0.70	0.77	**0.81**
dialect_clean	0.61	0.65	**0.70**
all	0.59	0.65	**0.68**

6 Conclusion

STD is an appealing research topic with approaches to speech search beyond searching the word transcript. In this contribution, we presented results using our fuzzy lattice word-syllable hybrid on a new German evaluation corpus. The evaluation shows that compared to classic LVCSR word transcription, using a full-fledged STD system increases retrieval performance on a complex broadcast corpus. Contributions from other STD research groups indicate that some of the findings can be transferred to STD in other languages (e.g., [17]). However, setting up a cross-lingual evaluation would require not only comparable corpora and query sets, but also systems which can be applied in multiple languages.

Literatur

[1] M. Akbacak, D. Vergyri, and A. Stolcke. Open-vocabulary spoken term detection using graphone-based hybrid recognition systems. In *Proceedings of ICASSP*, pages 5240–5243, 2008.

[2] Doris Baum, Daniel Schneider, Rolf Bardeli, Jochen Schwenninger, Barbara Samlowski, Thomas Winkler, and Joachim Köhler. DiSCo - a German evaluation corpus for challenging problems in the broadcast domain. In *Proceedings of LREC 2010*, 2010.

[3] Maximilian Bisani and Hermann Ney. Open vocabulary speech recognition with flat hybrid models. In *Proceedings of Interspeech*, pages 725–728, 2005.

[4] Ciprian Chelba, Jorge Silva, and Alex Acero. Soft indexing of speech content for search in spoken documents. *Computer Speech & Language*, 21(3):458–478, 2007.

[5] Dilek Hakkani-Tür, Frédéric Béchet, Giuseppe Riccardi, and Gokhan Tur. Beyond ASR 1-best: Using word confusion networks in spoken language understanding. *Computer Speech & Language*, 20(4):495 – 514, 2006.

[6] Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja, and Janne Pylkkänen. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech & Language*, 20(4):515–541, 2006.

[7] Pierre Jourlin, Sue E. Johnson, Karen Spärck Jones, and Philip C. Woodland. Spoken document representations for probabilistic retrieval. *Speech Commun.*, 32(1-2):21–36, 2000.

[8] Jonathan Mamou and Bhuvana Ramabhadran. Phonetic query expansion for spoken document retrieval. In *Proceedings of Interspeech*, pages 2106–2109, 2008.

[9] Timo Mertens and Daniel Schneider. Efficient subword lattice retrieval for German spoken term detection. In *Proceedings of ICASSP*, pages 4885–4888, 2009.

[10] Timo Mertens, Daniel Schneider, and Joachim Köhler. Merging search spaces for subword spoken term detection. In *Proceedings of Interspeech*, pages 2127–2130, 2009.

[11] S. Parlak and M. Saraclar. Spoken term detection for Turkish broadcast news. In *Proceedings of ICASSP*, pages 5244–5247, 2008.

[12] Murat Saraclar and Richard Sproat. Lattice-based search for spoken utterance retrieval. In *Proceedings of HLT-NAACL*, pages 129–136, 2004.

[13] Daniel Schneider, Timo Mertens, Martha Larson, and Joachim Köhler. Contextual verification for open vocabulary spoken term detection. In *Proceedings of Interspeech*, 2010.

[14] Daniel Schneider, Jochen Schon, and Stefan Eickeler. Towards large scale vocabulary independent spoken term detection. In *SIGIR SSCS Workshop*, pages 34–41, Singapore, July 2008. ACM.

[15] F. Seide, Peng Yu, and Yu Shi. Towards spoken-document retrieval for the enterprise: Approximate word-lattice indexing with text indexers. In *Proceedings of IEEE ASRU*, pages 629–634, 2007.

[16] Alex Waibel, Hua Yu, Martin Westphal, Hagen Soltau, Tanja Schultz, Thomas Schaaf, Yue Pan, Florian Metze, and Michael Bett. Advances in meeting recognition. In *Proceedings of HLT-NAACL*, pages 11–13, 2001.

[17] Roy Wallace, Robbie Vogt, and Sridha Sridharan. Spoken term detection using fast phonetic decoding. In *Proceedings of ICASSP*, pages 4881–4884, 2009.

[18] J. G. Wilpon, L. R. Rabiner, C. H. Lee, and E. R. Goldman. Automatic recognition of keywords in unconstrained speech using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(11):1870–1878, 1990.