MERGING SEARCH SPACES FOR SUBWORD SPOKEN TERM DETECTION

Timo Mertens[1], *Daniel Schneider*[2] *and Joachim Köhler*[2]

[1]Department of Electronics and Telecommunications, NTNU, Trondheim, Norway
[2]Fraunhofer IAIS, Schloss Birlinghoven, 53754 Sankt Augustin, Germany
timo.mertens@iet.ntnu.no, {daniel.schneider,joachim.koehler}@iais.fraunhofer.de

Abstract

We describe how complementary search spaces, addressed by two different methods used in Spoken Term Detection (STD), can be merged for German subword STD. We propose fuzzy-search techniques on lattices to narrow the gap between subword and word retrieval. The first technique is based on an edit-distance, where no *a priori* knowledge about confusions is employed. Additionally, we propose a weighting method which explicitly models pronunciation variation on a subword level and thus improves robustness against false positives. Recall is improved by 6% absolute when retrieving on the merged search space rather than using an exact lattice search. By modeling subword pronunciation variation, we increase recall in a high-precision setting by 2% absolute compared to the edit-distance method.

Index Terms: subword speech recognition, spoken term detection, pronunciation variation

1. Introduction

As vast amounts of media can be stored digitally, methods are required to make them searchable. For data containing speech, a challenging task is Spoken Term Detection (STD) in which all occurrences of a word or phrase have to be located precisely in a database.

A straightforward approach to STD is to use a large vocabulary continuous speech recognizer (LVCSR) to transcribe the speech parts of the data. However, this implies that if a word is not in the vocabulary it cannot be transcribed, and thus cannot be found upon search time. This phenomenon is known as the out-of-vocabulary (OOV) problem and is especially prominent in languages that make use of compounding or employ complex morphological patterns to create new words. Examples of such languages are Finnish, German and Turkish [1, 2, 3]. One way to alleviate this problem is to use subwords as the recognition unit. In contrast to words, the inventory of subwords, e.g. syllables, is finite and hence known a priori. Their compositional nature allows for retrieval of compound terms that otherwise could not be found on the word level. However, a decrease in recall compared to word STD has been observed [2, 4].

The 1-best transcription given by the LVCSR will contain other misrecognitions besides those caused by OOVs. This is especially the case for spontaneous speech, which is often unplanned and may contain a high number of false starts. Several STD techniques have emerged to cope with these recognition errors. One approach is retrieval on word or subword lattices (or lattice-like variants) instead of the 1-best transcription [5]. In [4] we compared this approach to an edit-distance search on the 1-best transcription [6], which matches the query with the 1-best transcription and calculates a distance according to a similarity measure. Like other fuzzy approaches to subword STD [7], this method can tolerate a degree of mismatch between the recognition output and the query, and is thus able to find keywords in erroneous transcriptions. On a German STD task, we found that on a subword level, each approach covers a different area of the search space, and comes with relative advantages and disadvantages.

In this contribution we present ways of merging the search spaces covered by the two previously mentioned retrieval approaches. The aim is to exploit the complementary nature of subword lattice and subword fuzzy-search by combining them into a single, unified search method. Further, we investigate how explicit modeling of pronunciation variation on a subword level can improve robustness of the retrieval method.

2. Subword Spoken Term Detection

As this contribution investigates subword search spaces, the choice of recognition unit is important. Several intermediate units between phonemes and words have been proposed [6, 8, 9]. In [4] we found that syllables are a viable subword unit for German STD. Their size makes them suitable in terms of retrieval efficiency since smaller subword units usually increase the complexity of the search algorithm. Furthermore, syllables present a good trade-off between the number of distinct units in the lexicon and stability in terms of recognition context. In the remainder of this work we use syllables as our subword indexing unit.

The recognized 1-best subword hypothesis often deviates from the true subword representation of the speech. Different STD subword retrieval algorithms are available to cope with these deviations. We present a short description of the search spaces, followed by the algorithms that address them.

2.1. Subword Search Spaces

ASR Error Search Space: Unsurprisingly, speech recognition errors are a fundamental problem in STD. Errors are caused by a wide range of factors, such as inadequate language and acoustic models, difficult speech material, or an incomplete recognition vocabulary. Although the latter can be addressed by subword retrieval, other automatic speech recognition (ASR) errors cannot be easily remedied. The ASR 1-best transcription often contains incorrect substitutions, insertions or deletions of hypotheses because of the uncertainty of the acoustic and language models, which prevents an exact search on the 1-best transcript.

Pronunciation Search Space: Although subword ASR errors are mainly due to misrecognitions of the acoustic segments, there are cases where the recognition is indeed correct but only the pronunciation of the segment in question differs from the canonical form. Pronunciation variation is usually addressed on the word level by incorporating multiple pronunciations in the word lexicon. On a subword level, however, the situation cannot be handled in the lexicon since the transcription of the syllable already reflects its pronunciation. An example of a pronunciation variation is the German conjunction *und* (and). Canonically, the monosyllabic word is realized with the unvoiced plosive */t/* in Coda position. In spontaneous speech, however, the plosive can be dropped, producing [U_n_]. Hence, even if the subword ASR is correct when comparing the ASR output to the spoken subword sequence, an exact search for the canonical representation [U_n_t_] will fail.

2.2. Baseline Retrieval Methods

In this Section we introduce the baseline retrieval methods compared in [4].

Lattice Retrieval: During decoding, the ASR considers competing hypotheses of what was actually spoken. Making these hypotheses available to the retrieval algorithm via lattices leads to a robust coverage of the ASR error search space. Many variants exist for increasing the retrieval efficiency [10, 11, 12]. Lattice retrieval usually achieves a moderate gain in recall while maintaining a high degree of precision. The advantage is that evidence from the decoder is used during retrieval, i.e., confidence scores and competing hypotheses are consulted to assess the quality of a hit. Even though lattices cope with ASR errors efficiently, they do not model subword pronunciation variation, as described in Section 2.1.

Fuzzy-search on 1-best: Fuzzy syllable search relies on approximating the distance between the query and each position in the 1-best transcription using a sliding window [2, 6]. If the distance is above a prescribed threshold, δ, then the position is accepted as a hit. The computation of the distance can be considered as a two-stage Levenshtein distance. An edit-distance between the query and the windowed syllable sequence is calculated. For the substitution cost, another edit distance between the substituted syllables is estimated on the phoneme level, such that the substitution of syllables with a similar phonetic representation is cheaper. Fuzzy-search disregards acoustic information given by the speech signal and does not explicitly model either of the described error phenomena. This leads to a high increase in recall but comes with a significant loss in precision.

3. Merging of Search Spaces

As described above, subword retrieval needs to cope with two complementary search spaces. In [4] we found that in a German STD task, 14 out of 551 query occurrences that were not found with fuzzy-search could be retrieved on lattices. On the other hand, fuzzy-search was able to find 29 correct results (including subword pronunciation variations) which were not found on lattices. The focus of this contribution is to merge the search spaces with a novel retrieval approach such that the strengths of the baseline retrieval approaches are combined.

We propose the following novel retrieval methods:

- *Edit-distance Lattice Retrieval*: Apply the fuzzy-search algorithm proposed in [6] on syllable lattices. A similar approach has been successfully applied to phone lattices for English STD [13].
- *Pronunciaton Lattice Retrieval* : Use statistical pronunciation models to score the deviation between lattice path and query.

While the first technique uses a simple edit-distance to compare the query to a syllable sequence, the second makes use of a priori knowledge about possible subword confusions.

3.1. Edit-distance Lattice Search

In order to speed-up lattice retrieval, we exploit the Zipfian distribution of syllable occurrences. As in [4], we reduce the set of lattices in the collection to those which contain the least frequent syllable of the query. All paths traversing this anchor syllable in the lattice are then matched against the query, using the edit-distance approach described in Section 2.2, and a distance is obtained. If the distance is above δ, the hit is accepted. Due to the nature of the fuzzy-search we expect an improvement in recall over both lattice and fuzzy retrieval in isolation. Precision should be higher than for standard fuzzy-search, but possibly lower than for lattice retrieval.

3.2. Pronunciation Lattice Search

This approach uses statistical pronunciation models to assess the quality of lattice paths w.r.t. the query. We argue that by directly simulating subword pronunciation variation, we should be able to address the pronunciation search space robustly in addition to the search space covered by the lattice.

Since we are investigating syllable retrieval, we would like to have a statistical model that predicts likely pronunciation confusions for each possible syllable. Such a model could then be incorporated in the retrieval algorithm, making it possible to explicitly include knowledge about pronunciation variation on a subword level. Ideally, each model would be trained on observed pronunciation variations for the syllable. In practice, however, it is difficult to derive such statistics for all syllables (there are $\sim$10k distinct syllables in our dictionary).

Therefore, we propose a smaller confusion unit: *position-specific cluster confusions* (PSCC). This unit exploits the natural structure of syllables, in which three adjunct phone clusters occur: Onset, Nucleus and Coda clusters. Each of these positions can contain phone clusters of varying lengths, depending on the phonotactics of the language. In German speech, these phone clusters are realized differently depending on their position in the syllable. A canonical */t/*, for example, might be dropped in Coda position, but not in Onset position. Hence, PSCCs represent position-specific confusion evidence for phone clusters, and can be used to generalize pronunciation variation from a phone cluster level to a syllable level by training statistical models for each dictionary syllable.

The training procedure of the PSCC-based syllable models is as follows. First we produce a syllable ASR transcription of a training corpus, and obtain the canonical syllable transcription using grapheme-to-phoneme conversion of the reference transcript. We obtain syllable confusion pairs by aligning both transcriptions using a minimal edit-distance. We then find PSCCs by breaking down the canonical syllable and its substitution into phone clusters. Cluster boundaries are found by grouping consonant and vowel clusters, taking into account language specific phonotactic constraints. Then, PSCC substitution counts are inferred from the syllable substitutions. The probability of confusing a canonical phone cluster C_{can} with another phone cluster C_{conf} is given by

$$P(C_{\text{conf}}|C_{\text{can}}) = \frac{\text{count}(\text{align}(C_{\text{can}}, C_{\text{conf}}))}{\text{count}(C_{\text{can}})} \qquad (1)$$

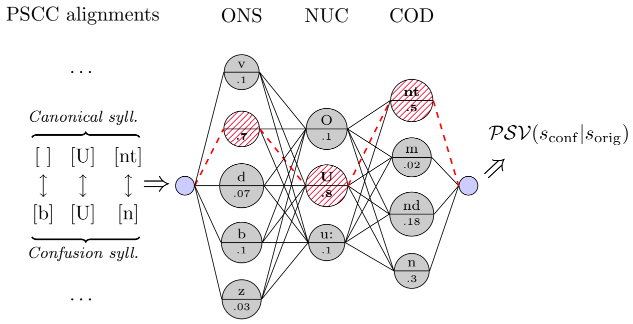

Figure 1: *Stat. model for the canonical syllable* und [U_n_t_].

where $\text{count}(\text{align}(\cdot\,,\cdot))$ denotes the number of substitutions between the given clusters during the aforementioned alignment. From the PSCC statistics, we generate a syllable confusion model for each canonical syllable. First, the canonical syllable is segmented into its position-specific phone clusters. For each canonical phone cluster we obtain the most likely confusions given by the PSCC statistics estimated during training. Then we construct a sub-syllable lattice with three alignment positions, corresponding to Onset, Nucleus and Coda clusters, where a node represents a phone cluster together with its confusion probability $P(C_{\text{conf}}|C_{\text{can}})$. We populate each position in the lattice with the canonical cluster and its confusions. The process is illustrated in Fig. 1, where, at each lattice position, a red node denotes an original cluster and the others its PSCCs.

In our lattice retrieval framework, we want to apply these models to find out whether a syllable on a lattice path is a pronunciation variation of a canonical query syllable. We first construct the sub-syllable pronunciation model for the lattice syllable as described above. Given that the model contains a PSCC path with the canonical PSCC sequence from the query, the pronunciation variation score ($\mathcal{PVS}$) is computed according to

$$\mathcal{PVS}(s_{\text{latt}}|s_{\text{query}}) = \prod_{N} P(C_{\text{latt}}|C_{\text{query}}) \tag{2}$$

where s_{latt} is the lattice syllable, s_{query} is the canonical query syllable and N corresponds to all three nodes on the path through the sub-syllable lattice of the model for s_{query}.

For a canonical query $q = q_1 \dots q_N$ with syllables q_i and a path $P = p_1 \dots p_N$ through the lattice with lattice syllables p_i, we can now estimate a pronunciation score $\mathcal{S}$:

$$\mathcal{S}(P|q) = \prod_{i=1}^{N} \mathcal{PVS}(p_{\text{i}}|q_{\text{i}}) \tag{3}$$

Again, if $\mathcal{S}$ is above δ, the path is accepted as a hit.

Edit-distance lattice retrieval does not model a single error phenomenon, but simulates both ASR and pronunciation errors. Using PSCC probabilities, however, moves away from the rather uncontrolled strategy used by the edit-distance paradigm towards explicitly modeling pronunciation variation, thereby reducing the amount of false alarms at high recall levels.

4. Experimental Methodology

We use the same data as described in [2, 4] to evaluate the proposed novel retrieval methods, namely 3.5 hours of manually segmented German speech data with both planned and spontaneous speech. We use 213 queries containing single- and multiple words with 551 occurrences in the data.

The open-source decoder Julius[1] was applied for generating 1-best transcriptions and lattices on a syllable level, resulting in a syllable error rate of 29.1%. We used the acoustic and language model setup described in [4] with an increased acoustic training set of 50 hours and a 4-gram syllable language model. For the PSCC generation we aim to reduce the effect of ASR errors by using the ASR training data for estimating the pronunciation variation models. We re-recognize the acoustic model training data and then align the syllables of the reference with the syllables in the recognition transcripts. Because the syllables in the reference transcriptions were obtained automatically from a word-level reference, the syllable sequences are canonical in terms of their pronunciation. The recognized syllables, however, are based on the true pronunciation of the spoken utterance and can thus differ from the reference in the terms of subword pronunciations. As the re-recognition is not perfect, however, some ASR confusions will additionally influence the pronunciation models.

We use the standard metrics precision, recall and actual term-weighted value (ATWV) to evaluate the methods.

5. Results

In this section we present the evaluation results for the novel retrieval approaches. First, the baseline results, i.e., exact lattice and 1-best edit-distance search in isolation, are reported. Following this, we evaluate the first merging approach of edit-distance and lattice retrieval, edit-distance lattice retrieval. Finally, we analyze the results of the PSCC-weighted retrieval method.

In Table 1 we summarize the results of the baseline methods. We see that exact search on the 1-best syllable transcript is very precise, but lacks coverage. Exact syllable lattice search does not achieve the same recall level as edit-distance search on the 1-best transcript, but remains at a high degree of precision. Edit-distance search on the 1-best generates the highest recall but is rather imprecise, as it simulates possible ASR errors without additional knowledge from the actual ASR decoding process. Because the default NIST ATWV weighting favors recall, the edit-distance 1-best method achieves the highest score.

Fig. 2 illustrates the results for edit-distance search on 1-best and on lattices with varying δ. Edit-distance search on lattices gives consistent recall improvements at equal precision compared to the 1-best approach. Hence, the new retrieval approach merges the available search spaces successfully and achieves the highest recall values on a subword level observed so far in our experiments. Since lattice retrieval uses an exact anchor lookup, only a subset of all possible paths through the lattice is made available for subsequent edit-distance search. This restricts the search space initially, but also eliminates the high false-positive rate of 1-best edit-distance search. Furthermore, the complexity of retrieval on these pre-filtered lattice paths is considerably lower than for standard edit-distance search on the whole 1-best transcript. Fig. 3 compares the behavior of edit-distance lattice search and PSCC lattice retrieval, restricted to results with precision over 85%. One can see that there is indeed a considerable boost in recall in this particular setting when using PSCCs. Where the rather uncontrolled variation strategy of the edit-distance variant produces a considerable amount of false alarms with small additional recall gains, the PSCCs use a more coherent and restricted way of modeling pronunciation variation. Unlike the edit-distance approach, PSCCs concen-

[1] http://julius.sourceforge.jp/en

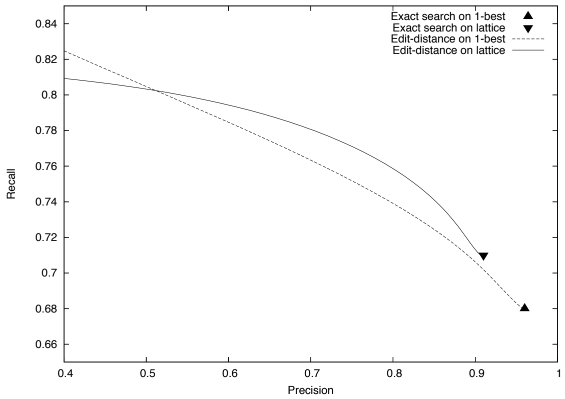

Figure 2: *Edit-distance search on 1-best and lattice paths with varying δ.*

Table 1: Baseline results.

Approach	Recall	Precision	ATWV
Exact 1-best	0.68	0.96	0.65
Exact lattice	0.71	0.91	0.68
Edit-distance 1-best	0.74	0.81	0.71

trate on the pronunciation search space and leave the ASR error search space to the lattice, leading to fewer false alarms.

However, as δ decreases for both approaches, we observe that the recall gain of the PSCC approach levels off. The additional gain of the edit-distance method is mainly caused by the remaining ASR errors which are not covered by the lattice, and which are simulated by the edit-distance search. Nonetheless, by simulating pronunciation variation in addition to the exact lattice, an increase of 3% absolute in recall is achieved. Table 2 summarizes the results tuned towards optimal ATWV.

6. Conclusion

We have presented ways of merging the available search spaces for subword STD. The motivation was to use the complementary advantages of well-known retrieval approaches such that the coverage of the overall search space is improved. We proposed a novel weighting scheme exploiting the phonetic cluster structure of syllables. Thereby we explicitly model pronunciation variation on a syllable level with the goal of having more control over the allowed variation in the lattice.

The actual merging method should be selected depending on the application. While edit-distance retrieval outperforms the other methods in terms of possible recall gain, PSCC retrieval proved to be more robust against FP in a high-precision setting.

Using sub-syllable phone clusters to predict pronunciation variation proved to be successful. A possible improvement would be to combine confidence measures from the lattice with PSCC scores to address false-positive more robustly. It is reasonable to expect that PSCCs are also useful in other applications where subword pronunciation variation needs to be modeled, such as applying recent approaches to phonetic query expansion [14] to German syllable STD.

7. Acknowledgments

The work for this contribution was partly supported by the VITALAS project (EU IST-FP6) and the SMUDI project under the Research Council of Norway's VERDIKT programme.

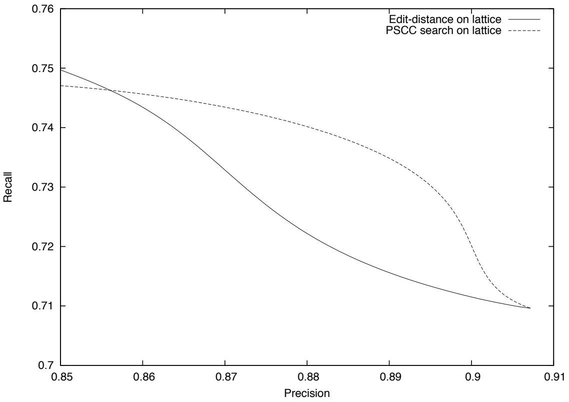

Figure 3: *Edit-distance and PSCC lattice search above 85% precision.*

Table 2: Optimal ATWV results.

Approach	recall	precision	ATWV
Edit-distance lattice	0.77	0.80	0.73
PSCC lattice	0.74	0.89	0.72

8. References

[1] T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pylkkönen, "Unlimited vocabulary speech recognition with morph language models applied to Finnish," *Computer Speech & Language*, vol. 20, no. 4, pp. 515–541, Oct. 2006.

[2] D. Schneider, J. Schon, and S. Eickeler, "Towards Large Scale Vocabulary Independent Spoken Term Detection," in *SSCS, SIGIR, Singapore*. Singapore: ACM, July 2008, pp. 34–41.

[3] S. Parlak and M. Saraclar, "Spoken term detection for Turkish Broadcast news," in *Proc. ICASSP,*, 2008, pp. 5244–5247.

[4] T. Mertens and D. Schneider, "Efficient Subword Lattice Retrieval for German Spoken Term detection," in *Proc. ICASSP*, 2009.

[5] M. Saraclar and R. Sproat, "Lattice-based Search for Spoken Utterance Retrieval," in *Proceedings of HLT-NAACL*, 2004.

[6] M. Larson and S. Eickeler, "Using syllable-based indexing features and language models to improve German spoken document retrieval," in *Proc. Eurospeech*, 2003, pp. 1217 – 1220.

[7] J. Pinto, I. Szöke, S. Prasanna, and H. Hermansky, "Fast approximate spoken term detection from sequence of phonemes," in *Proc. SSCS 2008: Speech search workshop at SIGIR*, 2008, pp. 28–33.

[8] M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraclar, and A. Stolcke, "Morph-based speech recognition and modeling of out-of-vocabulary words across languages," *ACM Transactions on Speech and Language Processing*, vol. 5, no. 1, pp. 1–29, 2007.

[9] M. Bisani and H. Ney, "Open vocabulary speech recognition with flat hybrid models," in *Interspeech*, 2005, pp. 725–728.

[10] C. Chelba, J. Silva, and A. Acero, "Soft indexing of speech content for search in spoken documents," *Computer Speech & Language*, vol. 21, no. 3, pp. 458–478, Jul. 2007.

[11] Y.-C. Pan, H.-L. Chang, B. Chen, and L.-S. Lee, "Subword-based Position Specific Posterior Lattices (S-PSPL) for Indexing Speech Information," in *Proc. Interspeech*, 2007, pp. 318 – 321.

[12] P. Yu, Y. Shi, and F. Seide, "Approximate word-lattice indexing with text indexers: Time-Anchored lattice expansion," in *Proc. ICASSP*, 2008, pp. 5248–5251.

[13] R. Wallace, R. Vogt, and S. Sridharan, "A phonetic approach to the 2006 NIST spoken term detection evaluation," in *Proc. Interspeech*, 2007, pp. 2385 – 2388.

[14] J. Mamou and B. Ramabhadran, "Phonetic query expansion for spoken document retrieval," in *Proc. Interspeech*, 2008.