

Subword-based Indexing for a Minimal False Positive Rate

Research Proposal

Laurens van der Werff
University of Twente
PO Box 217, NL-7500AE
Enschede, The Netherlands
laurens.w@ewi.utwente.nl

Willemijn Heeren
University of Twente
PO Box 217, NL-7500AE
Enschede, The Netherlands
w.f.l.heeren@ewi.utwente.nl

1. INTRODUCTION

The aim of Spoken Document Retrieval (SDR) is to find speech fragments that are relevant to user queries. This is usually accomplished by applying Information Retrieval (IR) methods to an automatic transcription of an audio collection. Although there are many different manners of generating such a transcription and there are also many different techniques for retrieving documents, all popular techniques share the same mismatch: Automatic Speech Recognition (ASR) focuses on frequent terms, while IR typically exploits rare events.

A typical 65k ASR lexicon can describe >96% of the words spoken, depending on the subject and type of the speech. However, our own experiments on a set of Dutch ASR transcripts, a subset of the Spoken Dutch Corpus¹ (CGN), show that >80% of the unique terms that were uttered in this collection were left out in the lexicon, making these terms Out of Vocabulary (OOV). For a dictation-type task this need not be a problem, since the error rate is only affected marginally by the OOV-rate, but when the same transcript is used for SDR, the consequences of the deletions may be much more severe: any queries using OOV terms will fail on this collection. If queries typically use the most discriminative (read: rare) words in a collection, and >80% of the rarest terms are omitted, the problem becomes obvious.

Transcribing the audio in a manner that is not limited to a predefined lexicon may solve this issue. Attempts that have been made so far include the use of phonemes and phoneme-groups (subwords) as basic recognition blocks instead of traditional word-sized units [6, 9]. There seem to be two main disadvantages to such approaches: i) the perplexity of the Language Model (LM) for these units is much higher than for the typical word-based LM, potentially leading to more ASR errors, and ii) (near) homonyms and word boundaries cannot be detected accurately. Both of these

¹<http://lands.let.ru.nl/cgn/ehome.htm>

will cause False Positive (FP) errors in word-spotting scenarios. Despite these issues, subword units seem to be the most promising solution for handling OOV query terms.

If an SDR system is thought of as a concatenation of ASR, IR and a User Interface (UI), then introducing subword units will have consequences for each of these components. Within the ASR, both the choice of subword units and the manner in which they are generated have to be considered. For the IR engine, a hybrid index containing traditional words as well as subword units provides new challenges for resolving queries, and finally the UI must be able to support the user in obtaining maximum benefit from subword-based strategies.

The experiments proposed in this paper have three main goals: i) determine the best way of creating a hybrid index, ii) find a criterion for automatically determining the best retrieval strategy, and iii) estimate the expected reduction in search space that a user can be expected to achieve from expanding a query in order to reduce FPs.

Section 2 will explain the choices that must be made when generating a transcription that is suitable for subword-based indexing, while Section 3 will delve into the issues related to exploiting such an index in order to gain the most benefit from it. Section 4 explains how to measure the potential contribution of a user in making subword based retrieval feasible for large collections and Section 5 lists the experimental setup proposed to evaluate a hybrid SDR system.

2. SUBWORD-BASED TRANSCRIPTION

Attempts made so far with respect to subword-based transcription for SDR vary from using a purely phonetic approach to hybrid systems where both words and subword units are used in conjunction. Any approach based on phonemes inevitably introduces more recognition errors, but only on in-vocabulary terms. The overall expected result therefore is: more errors on $\approx 96\%$ of words, but (potentially) better recognition on $\approx 80\%$ of unique terms. For dictation purposes, focus is on *word* error rate, for SDR *term* coverage is also important, making a subword-based approach potentially beneficial. A hybrid index allows for matching OOV query-terms on subword units, while retaining the original performance on in-vocabulary (INV) query-terms by utilizing the word-based transcription.

Phonetic transcription systems typically use a word-based

speech recognition engine. The word-based output is then converted into phonemes or, instead of using language models and a lexicon to restrict the output to well-formed sentences, it is determined only by the acoustic match between speech and models, thereby ignoring structural properties of the language. Both were examined in [5], where it was determined, using Mean Average Precision (MAP) as a criterion and transcriptions with an error rate of around 45%, that a phonetic transcription that is produced through grapheme-to-phoneme conversion of a word-based transcription performs better in an SDR environment than one that is generated directly.

An alternative to a phonetic transcription is the use of larger subword units, i.e., word-fragments. This may allow for some form of language modeling, while also simplifying the generation of an index. Especially when multiple transcription alternatives are generated (e.g., by using lattices), larger units may provide some advantages. The combination of subword-sized units and a hybrid lattice transcription (containing both word and subword units) was put into practice in [9], which showed encouraging results for word-spotting performance on OOV terms.

In the experiments proposed here, word-based ASR lattices will be used as a basis for generating the hybrid subword-/word-based index. Since most of the speech can be transcribed adequately by a word-based ASR system, making a subword transcription of all the data may not be beneficial. Ideally, only those areas that are likely to contain the terms that are resolved by the IR component as a subword-based query, see Section 3, should be converted into a subword representation. Therefore, instead of making a full subword-based transcription of each utterance, only those areas of the lattice that show the most evidence for potential gain will be converted, for example because of certain combinations of acoustic and linguistic posterior probabilities [1].

3. QUERIES AND THE HYBRID INDEX

An index that enables searching at both the acoustic level (phonemes or word-fragments) and the traditional word level allows for some extra flexibility that should be exploited optimally. Using the subword index only for OOV query terms, while using the word-based index for in-vocabulary (INV) terms is the approach chosen in [3]. FPs are only introduced for query terms that would previously have given no results at all, while ‘normal’ performance can be maintained for all other query terms.

However, as was shown in [9], there can be a benefit to using subword based queries for INV terms as well, especially when the language model (LM) that was used for generating the transcription is a mismatch to the domain of the speech. This can be explained through the fact that some LM probabilities may be incorrect due to insufficient training data, leading to more recognition errors on such terms. This would most likely be the case for infrequent terms that may then become practically OOV, as was shown in [2].

In our experiments we will attempt to find a criterion to automatically divide the (potential) query terms into three groups: 1. INV terms that are best served with traditional word-based retrieval, 2. OOV terms that can only generate

results using subword-based retrieval, and 3. INV terms for which ASR performance is so poor that a combination of subword and word-based retrieval is expected to give the best results.

4. THE ROLE OF THE USER

The reason why FPs are so harmful can be understood from the fact that from a user perspective discarding non-relevant results requires actually listening to audio fragments instead of just glancing over pieces of text. The problem with FPs is generally amplified in settings with large corpora, such as cultural heritage collections. For instance, the ‘Radio Rijnmond’ archive that is kept at the Municipal Archives of Rotterdam spans approximately 20,000 hours. For an SDR system that has an FP-rate of 0.5 per hour (which in general is quite low) and an average amount of 20 true hits per OOV query term (provided the term was uttered at all in the corpus), the average number of FPs would outnumber the correct hits by 500:1, and then only if all of the uttered instances of the OOV query term were actually found! Reducing the FP-rate may improve the user experience for smaller collections, but a purely subword-based retrieval is probably not very useful for users of large (>100 hours) collections.

However, users are quite capable of expanding the OOV query with an INV term. This enables the IR engine to first reduce the search space to those documents that contain the INV term and then perform the subword-based query on only these documents. Key in such a scenario is the success that can be expected from the reduction of the search space through the use of an additional INV query term.

In our experiments, all queries shall be formed from a combination of (at least) two terms, one of which should be resolved using word-based retrieval, while the other should be done using the subword-based approach. This will enable us to determine the average rate of reduction in search space that can be expected from using a strategically chosen additional INV query-term. We will assume that the average reduction from the ten best performing INV terms for each OOV term is representative of what an average user can expect to attain.

5. PROPOSED EXPERIMENTS

The three main goals of the experiments are: i) determine the best way of creating a hybrid index, variables will be the amount of speech that is transcribed into subword units and the choice of the subword units themselves, ii) find a criterion for automatically determining the best retrieval strategy for each query term in accordance with the groups described in Section 3, and iii) estimate the expected reduction in search space that a user can expect to achieve from adding an INV term to an OOV query.

5.1 Experimental Setup

Experiments will be carried out primarily on CGN, since this collection has been partially transcribed at the phonetic level and fully at the word level, allowing for a direct evaluation of the lattice-derived subword transcript. Additionally, it is divided into several types of speech, such as conversational telephone speech, broadcast news, sports comments, read

speech, etc., allowing for a comparison of results on these different types.

Automatic transcription of the speech data will be performed using an HMM-based ASR system [7] delivering output that contains word-based lattices. These will then be converted into a sausage structure [4] to enable easier conversion into a hybrid index.

5.2 Evaluation

A typical Information Retrieval (IR) performance measure such as MAP is flawed when it is applied to SDR. The reason for this is that text-based IR and SDR find their challenges at different parts of the lexical spectrum. Text-based IR actually performs quite well when query terms are rare or highly discriminative, since even a simple word-spotting algorithm is likely to return most relevant documents and produce few non-relevant ones. This contrasts with SDR, where any OOV or rare term may produce no relevant documents and/or large numbers of non-relevant ones (FPs) due to errors from the ASR stage. When this is combined with the knowledge that browsing through audio results is, for now at least, much more time-consuming than browsing through text results, it becomes clear that MAP is probably too forgiving towards FPs. This becomes especially problematic when looking at the subword-based approaches to dealing with OOV terms that are the subject of this research proposal. Additionally, MAP is highly dependent on the ranking of the results and therefore on the calculated relevance of terms in each document. Having the manner of score-calculation as an extra variable within these experiments is not desirable. Calculation of MAP also requires ground-truth judgments, severely limiting the queries that can be evaluated within such a context.

Instead, performance will be evaluated through a word spotting task. This allows for a complete set of queries, containing every possible combination of INV and OOV query terms. The Figure-of-Merit (FOM) and FP-rate can be used to judge the impact of choice of subword, while the algorithm for determining which fragments to convert into subwords can be evaluated directly from the reference transcription and the list of query terms that are resolved through the subword-based index.

Evaluation of ASR output for use in SDR is typically carried out using traditional ASR performance measures, such as Word Error Rate (WER) and Out-of-Vocabulary (OOV) rate. These measures are not ideally suited to the task even for a word-based transcription [8], but cannot be applied at all when the transcription is a (hybrid) lattice containing subword units. WER and OOV rate will therefore be given only as an indication of word based 1-best ASR performance but will not be used as an optimization criterion.

Dividing all potential query terms into the three groups mentioned in Section 3 can be done by simply measuring FOM/FP-rate for single word queries of each word in the reference transcription and determining which strategy (word or subword) performs best. Then an algorithm should be implemented that can make such a division automatically, based on word or LM properties.

It is then relatively straightforward to determine which (top-10) INV terms are the best ones to use for reducing the search space for the subword-based query terms. The average reduction in search space from these terms will be taken as a measure of the potential success a user may achieve from using an additional term based on his knowledge of the subject.

6. CONCLUSION

A subword-based approach to SDR may seem like the best way of handling the OOV issue, but the disadvantages as for the user experience may very well be underestimated. Any system making use of a subword-based transcription should at least ensure it is no more inconvenient than is strictly necessary, while maintaining the potential benefits. In order to do this, it is important that not only the subword-based index itself is optimized, but that also the retrieval strategy takes the properties of query terms into account. Optimization should be done based on performance that can be achieved under practically relevant conditions, making it essential to incorporate a strategy that a user is likely to employ when faced with unsatisfactory results.

7. REFERENCES

- [1] G. Evermann and P. Woodland. Large vocabulary decoding and confidence estimation using wordposterior probabilities. In *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 3, pages 1655–1658, 2000.
- [2] G. Jones, K. Zhang, E. Newman, and A. M. Lam-Adenisa. Examining the contributions of automatic speech transcriptions and metadata sources for searching spontaneous conversational speech. In *Proc. of the SIGIR 2007 workshop on Searching Spontaneous Conversational Speech*, pages 29–36, 2007.
- [3] B. Logan, P. Moreno, and O. Deshmukh. Word and sub-word indexing approaches for reducing the effects of oov queries on spoken audio. In *Proc. of the 2nd int. conf. on Human Language Technology Research*, pages 31–35, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.
- [4] L. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition: Word error minimization and other applications of confusion networks. *Computer Speech and Language*, 14(4):373–400, Oct 2000.
- [5] C. Ng, R. Wilkinson, and J. Zobel. Experiments in spoken document retrieval using phonetic n-grams. *Speech Communication*, 32(1–2):61–77, Sept. 2000.
- [6] K. Ng. *Subword-based Approaches for Spoken Document Retrieval*. PhD thesis, MIT, 2000.
- [7] B. Pellom. Sonic: The university of colorado continuous speech recognizer. Technical report, University of Colorado, 2001.
- [8] L. van der Werff and W. Heeren. Evaluating ASR output for information retrieval. In *Proc. of the SIGIR 2007 workshop on Searching Spontaneous Conversational Speech*, pages 7–14, 2007.
- [9] P. Yu and F. Seide. A hybrid word / phoneme-based approach for improved vocabulary-independent search in spontaneous speech. In *Proc. ICSLP2004*, pages 293–296.