Karlsruhe Institute of Technology

Strategies for Out-of-Vocabulary Words in Spoken Term Detection

Bachelor Thesis by

Henrich Kolkhorst

At the Department of Informatics
Institute for Anthropomatics (IFA)

Reviewer:	Prof. Dr. A. Waibel
Second reviewer:	Dr. S. Stüker
Advisor:	Asst. Prof. Dr. F. Metze (CMU)
Second advisor:	Dipl.-Inform. K. Kilgour

Duration: June 1, 2011 – September 30, 2011

This work has been supported by a scholarship of the Baden-Württemberg Stiftung. It has largely been produced at the Language Technologies Institute of Carnegie Mellon University within the interACT exchange program.

Hiermit versichere ich, die vorliegende Arbeit eigenständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet zu haben. Wörtlich oder inhaltlich übernommene Stellen sind als solche kenntlich gemacht und die Regeln zur Sicherung guter wissenschaftlicher Praxis im Karlsruher Institut für Technologie wurden beachtet.

Karlsruhe, den 30. September 2011

Henrich Kolkhorst

Abstract

In this work, a system for the Spoken Term Detection (STD) task has been designed, implemented and evaluated. The system can be used to perform textual queries on speech data. Emphasis has been placed on the ability to detect out-of-vocabulary words (OOVs).
Based on confusion network output by a Speech-to-text (STT) system, an index is created for on-demand querying. Query terms unknown to the underlying STT system are expanded by orthographically or phonetically similar words from the vocabulary. The decision component for judging detection candidates includes rescoring based on acoustic similarity and term-specific thresholds.
The proposed techniques are evaluated on development data from the 2006 NIST STD evaluation and an analysis of different influence factors is performed. The system offers a thorough implementation for the task and achieves competitive results.
On English Broadcast News (BN) and Meeting data, Actual Term-Weighted Values of 0.75 and 0.32 are achieved with the NIST 2006 development term list. On BN OOVs of the DryRun term list, results of 0.36 and 0.40 have been achieved with the orthographic and phone-based query expansion schemes.

Contents

1. Introduction

Over the past decades, large amounts of audio and multimedia data have been produced and, enabled by inexpensive storage space, archived. Therefore, sizable repositories of spoken language, from news shows and recordings of meetings to collections of oral history, are feasible and archives with over 100,000 hours of audio data exist (e.g. recollections of Holocaust survivors and witnesses in [GSO+02]).

Such speech corpora generally contain vast amounts of information. However, this information is not very accessible in an unprocessed, sequential form. Typically, information is more useful, if "random access" to particular parts of information is possible. Consequently, it is desirable to have some sort of index of an audio collection and the ability to answer (textual) queries based on this.

For answering textual queries, it is generally beneficial to also have a textual representation of the audio data in the repository. Automatic Speech Recognition (ASR) in the form of Speech-to-Text (STT) systems can be used to generate transcripts of audio recordings. Subsequently, these transcripts can be indexed and used for text-based search. However, STT systems are error-prone and the quality of the transcripts needs to be considered while performing queries.

Several use cases for such an indexing- and query-system immediately come to mind. For example, indexing recordings of broadcast news enables comparing reports on the same event in different media or at different times, whereas business analysts could use such a system to find all mentions of a particular company. Similarly, indexing of the aforementioned historical recollections eases access for historians. From a linguistic point of view, it might be useful to retrieve all known occurrences of a word for acoustic comparisons.

Several different approaches and names have been used for these and similar tasks. Spoken Document Retrieval (SDR) for example has been used for tasks in which a query should return a list of audio "documents" (for example podcast episodes) that are most "relevant" to the query (but might not contain it at all) [GAV00]. These tasks focus on information retrieval techniques, stressing document relevance and result ranking. Exact detections of words, however, are less important.

Spoken Term Detection (STD) is the process of finding all occurrences of a query in the audio corpus. It has strong focus on distinct detections of words including the exact position in the audio. It is, however, not necessary to find the most relevant or prominent occurrences [FAGD07].

1.1. Challenges in Detecting Out-of-Vocabulary Words

ASR systems typically have a finite vocabulary, i.e. only a certain set of words can be recognized. Although these vocabularies contain tens of thousands of words for the English language, there will generally be e.g. neologisms, names of individuals or foreign words that are not contained in the vocabulary. These words that are not in the vocabulary of a particular ASR system are called out-of-vocabulary words (OOVs). Words contained in the vocabulary, i.e. words that can be recognized by the system, are called in-vocabulary words (IVs). During the transcription process OOV words in the audio are either replaced by similar words or a placeholder symbolizing an unknown word.

In the context of Spoken Term Detection, queries containing OOV words are an important challenge. Since these query words cannot occur in the transcripts, OOV queries will not return any results unless they are handled specifically.

Since vocabularies contain the "most important" words of a language, one might consider OOV queries to be negligible. However, names and rare words are especially popular in searches and have a higher associated information content. In [LMTW00], Logan et al. report that in an audio search engine context, on average 16% of a user query consists of OOVs. Consequently, the ability to handle OOV queries becomes especially important in practical use cases and the performance on out-of-vocabulary words should not be neglected.

1.2. Objectives of this Work

The main objective of this work is the development and evaluation of a system for the Spoken Term Detection Task which is formally defined in chapter 2.

Due to the importance of out-of-vocabulary queries as described in section 1.1, emphasis is placed on their treatment. Using different approaches to detect and handle OOV terms, the performance of the system on OOV queries is measured and optimized.

In order to enable a comparison with similar work in literature, the evaluation is performed on English audio data from the 2006 STD evaluation of the National Institute of Standards and Technology (NIST) [FAGD07]. Different STT systems are used to evaluate the influence of transcription errors on the detection performance.

1.3. Scope of this Work

This work focuses on the detection of terms, given a set of ASR hypotheses. Although different contrastive STT systems are used, no work has been done to tune these systems to the particular STD task. The STD system can be seen as a form of "post-processing" of the speech recognition hypotheses.

Although different languages pose different challenges for STD systems and have varying need for OOV treatment, tests and evaluation of this work have only been performed on English broadcast news and meeting data (for details, see 7.1.1).

According to the task definition in 2.2, no work has been done to infer importance or relevance from the query detections. Only a binary classification into presumably true occurrences and false alerts is used and results are not ranked according to any measure.

Furthermore, this work does not include any user interface or integration into a particular use case. However, design and implementation of the system allow for a fairly easy integration into such scenarios.

1.4. Structure of this Thesis

In chapter 2, a formal task definition is given and fundamentals of both task and work, such as the evaluation criteria, are provided.

Chapter 3 gives an overview over related work and previous results in the Spoken Term Detection task.

Basic details and design considerations of the implemented system are provided in chapter 4, whereas chapters 5 and 6 focus on particular approaches in handling OOV queries and deciding whether possible candidates should be reported as detections.

All approaches proposed in these chapters are tested and evaluated in chapter 7. Chapter 8 summarizes the work and gives a conclusion.

2. Fundamentals

This chapter contains basic information helpful for reading this thesis. In 2.1, a short overview of a typical speech recognition system is given and confusion networks are introduced as a form of recognizer hypotheses.

In 2.2, the task of Spoken Term Detection is defined and evaluation criteria for system performance are described in 2.3.

2.1. Foundations of Automatic Speech Recognition

Automatic Speech recognition (ASR) is the process of automatically creating transcripts from audio files, typically based on statistical pattern recognition. In Large-vocabulary Continous-Speech Recognition (LVCSR), the audio waveform is converted to acoustic feature vectors which are used to find the most probable word sequence. For an overview, see e.g. [You96].

2.1.1. Posterior Probability of Word Recognition

The *posterior probability* of a word sequence W given (preprocessed) acoustic features A can be found with Bayes' theorem as in Equation 2.1. The factor $p(A|W)$ can be seen as an acoustic model score representing the likelihood of the observed pronunciation given the words. $P(W)$ as a language model score represents the probability of the word sequence.

$$P(W|A) = \frac{p(A|W) \cdot P(W)}{p(A)} \tag{2.1}$$

Typically, only the most probable word sequence $\hat{W}$ as defined in Equation 2.2 is needed. Since $p(A)$ is a constant in this context, simplification yields Equation 2.3 which only depends on an acoustic and a language model.

$$\begin{aligned} \hat{W} &= \arg\max_W P(W|A) && (2.2) \\ &= \arg\max_W \frac{p(A|W) \cdot P(W)}{p(A)} \\ &= \arg\max_W p(A|W) \cdot P(W) && (2.3) \end{aligned}$$

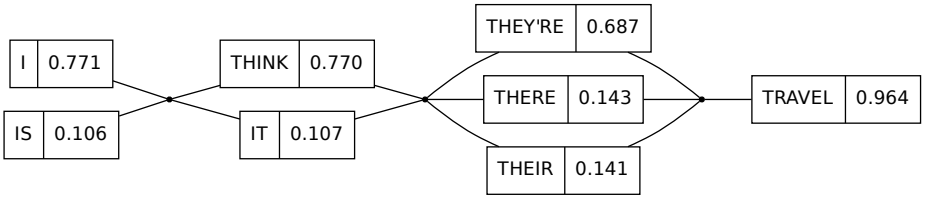


Figure 2.1.: Exemplary confusion network. Each hypothesis word is shown with its probability.

The process of finding $\hat{W}$ is called *Decoding* and consists of finding the best path in a search graph of word hypotheses. Due to computational complexity, only approximations of $\hat{W}$ are calculated in LVCSR context, typically obtained via beam search [You96].

A compressed form of the search graph is called a *lattice*. In the lattice, words are represented as edges and hence finding a word sequence represents finding a path in the lattice.

2.1.2. Confusion Networks

Confusion networks, originally described in [MBS00], are a representation of recognizer lattices which align the different hypotheses for a certain word based on time. Therefore, they "replace global search over a large set of sentence hypotheses with local search over a small set of word candidates" [[MBS00], p. 373].

As seen in Figure 2.1, the confusion network consists of different word clusters that represent different hypotheses of the same audio segment. Possible hypotheses from the confusion network would be for example "I think they're travel" or "Is it their travel" with the first word sequence being the more probable one.

Simultaneous words are grouped together into a cluster and their posterior probability is aggregated over all possible paths in the lattice containing the word at this time. Note that the posterior probabilities of all words in a cluster do not necessarily sum up to 1.

The representation as confusion networks is helpful for indexing and finding words and their posterior probabilities. Additionally, the clustering helps in identifying simultaneous words and hence avoiding duplicates. For more information on confusion networks, see [MBS00].

2.2. The NIST 2006 Spoken Term Detection Task

In 2006, the National Institute for Standards and Technology (NIST) established the task of Spoken Term Detection (STD) to "facilitate research and development of technology for retrieving information from archives of speech data", as stated in the evaluation plan in [Nat06]. Emphasis was placed on well-defined objectives and evaluation criteria.

The objective of STD is the detection of "terms" in large audio corpora. Terms consists of one or multiple words in written form. Unless otherwise noted, the words "term" and "query" will be used synonymously throughout this document.

A detection consists of the start- and end-time of the term occurrence. For diagnostic purposes, all candidates for detections should be reported accompanied by a confidence

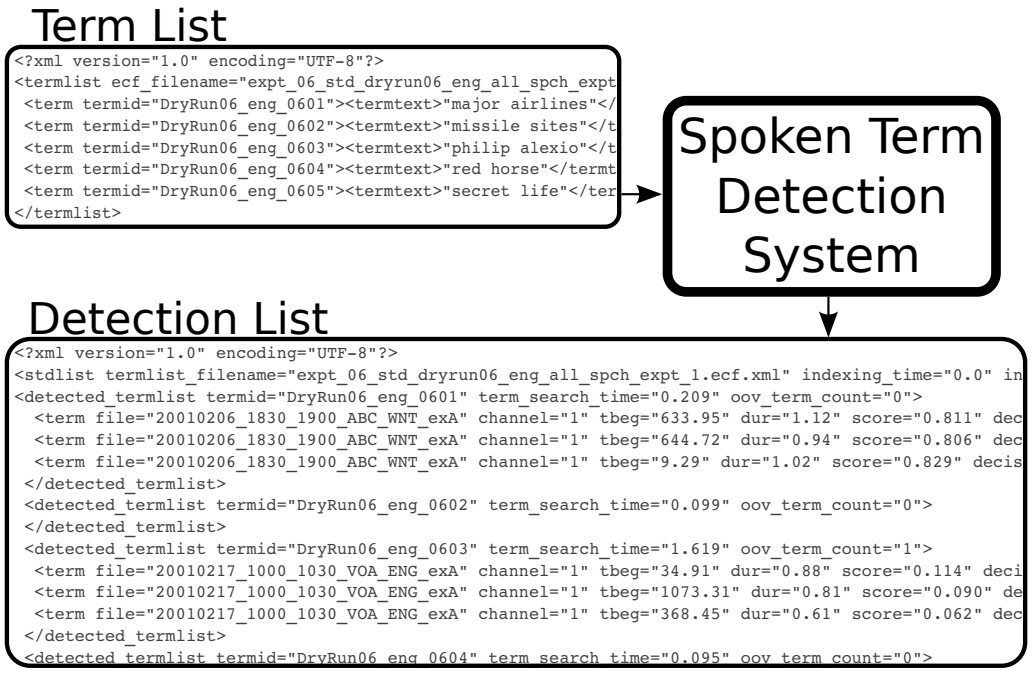


Figure 2.2.: Example of the term detection input and output

score and a binary decision of the system, whether the candidate is believed to be a "true" detection. However, only the candidates with a positive decision are considered in the evaluation process described in section 2.3.

Figure 2.2 shows excerpts from an actual term list and corresponding system output in XML format. Terms are provided in textual form, covering different topics from "missile sites" to "red horse". The expected system output contains the exact location of occurrences within the audio corpus and also includes diagnostic information such as search time or OOV rate.

Since the STD objective is motivated by the presence of large amounts of audio data, the task is split into two separate parts. In the *indexing phase*, the audio data is processed without any knowledge about possible query terms. This processing typically consists of using ASR software to produce transcripts of the audio and storing the transcripts in an index appropriate for the task.
In the *querying phase*, the index (and optionally the audio) can be used to retrieve the occurrences of the elements of a given query list. Generally, the indexing can be considered as an "offline" preprocessing step, whereas the querying should be possible in an "online" manner with only short query times.

Although the amount of data used in the 2006 evaluation is not very large, it is a stated goal of NIST to "simulate the very large data situation" [Nat06]. In order to compare search efficiency, total indexing time and search times for each query had to be reported in the 2006 evaluation.

2.3. Evaluation Criteria

The STD task has clear evaluation criteria regarding the definition of a term occurrence and the metrics used to judge a detection system. A scoring tool is provided by NIST and all reported results in this document are based on this software. Therefore, the information in this section is based on the NIST publication [FAGD07] unless noted otherwise.

For evaluation purposes, the occurrence of a term is defined based on a reference transcript of the audio. A term occurs in the corpus, whenever the reference transcript contains the term's word sequence in *orthographically identical* form. For multi-word queries, a certain amount of time is allowed between the end of a detected word and the beginning of its successor (set to 0.5 seconds in the 2006 evaluation). Multiple detections of a single reference occurrence are not allowed.

As an example, the term "thesis" does not match the word "theses" in a transcript since the plural form is spelled differently. Furthermore, no distinction is made based on context. Consequently, the term "suit" matches occurrences of the word both in the context of clothing ("three-piece suit") and in the context of a judicial system ("law suit"). NIST sees these characteristics of the task evaluation "as being a relatively minor distortion of the objective" [Nat06].

2.3.1. Trial-based Evaluation

In a detection task, systems try to detect events in a certain number of trials. System performance can generally be characterized by two types of correct system decisions and two types of system errors as shown in Table 2.1. It is desirable to maximize the number of correct detections ("true positives") of events (term occurrences in the STD task) while minimizing the number of spurious detections ("false positives") and missed events ("false negatives").

Typically, all of these types can be treated as discrete quantities, such as the number of missed term occurrences in the STD task. However, the number of trials, and consequently the number of correct rejections, have no canonical definition in the context of continuous speech where the true number of words is unknown for the system. Possible solutions would be to use for example the number of words in the reference transcript or to use properties of the audio data.

For the STD evaluation metrics, NIST chose the latter variant. The number of trials "is defined somewhat arbitrarily to be proportional to the number of seconds of speech" with one word per second in the 2006 evaluation. Therefore, the number of possible false alerts, also known as non-target trials $N_{NonTarget}$, of a term t as defined in Equation 2.4 is dependent on T_{speech}, the length of the audio data in seconds, and $N_{true}(t)$, the number of actual occurrences of a term t in the reference [FAGD07].

$$N_{NonTarget}(t) = T_{speech} - N_{true}(t) \tag{2.4}$$

2.3.2. Detection-Error-Tradeoff Curves

Based on the availability of a confidence score for each detection and the assumption of a global threshold θ (identical for all terms), NIST defines Detection-Error-Tradeoff (DET) curves as plots of miss probability (Equation 2.5) versus false alert probability (Equation 2.6) parametrized by the global threshold [FAGD07].

	true	**false**
positive	correct detection	false alert
negative	correct rejection	miss

Table 2.1.: Elements of system performance on detection tasks. The true/false columns describe whether a decision is correct and the positive/negative rows describe whether a system detection occurred

$$P_{Miss}(t,\theta) = 1 - \frac{N_{correct}(t,\theta)}{N_{true}(t)} \tag{2.5}$$

$$P_{FalseAlert}(t,\theta) = \frac{N_{spurious}(t,\theta)}{N_{NonTarget}(t)} \tag{2.6}$$

A lower value of P_{Miss} represents a higher recall of the system and a lower value of $P_{FalseAlert}$ represents a higher precision. Consequently, both values should be minimized and DET curves should be "close to the origin" of the coordinate system. An example DET curve can be found in Figure 7.5.

2.3.3. Actual Term-Weighted Value

Whereas DET curves give a general overview over a system's performance and ignore the binary decision of the detection system, NIST defines the Actual Term-Weighted Value (ATWV) as a "single performance metric [...] for [...] optimization" [FAGD07] and it is used as the primary evaluation metric to compare the quality of STD systems.

Given a list of query terms $\mathcal{T}$, the ATWV as defined in Equation 2.7 represents a weighted combination of miss probabilities and false alert probabilities, averaged over all terms. For the ATWV-calculation, the binary decision of the system for each candidate detection is used. Therefore, the ATWV does not depend on a threshold θ, as opposed the definitions for the DET analysis above.

$$ATWV(\mathcal{T}) = 1 - \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\left(P_{Miss}\left(t\right) + \beta \cdot P_{FalseAlert}\left(t\right)\right) \tag{2.7}$$

$$= \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\left(\frac{N_{correct}\left(t\right)}{N_{true}\left(t\right)} - \beta \cdot \frac{N_{spurious}\left(t\right)}{N_{NonTarget}\left(t\right)}\right) \tag{2.8}$$

In 2.7, the weight β is term-independent, taking into account a static prior probability of a term and cost-value-ratio for spurious versus correct detections (for details, see [FAGD07]). In the 2006 evaluation, $\beta = 999.9$.

Since all terms contribute equally to the ATWV score, it is not biased toward terms with large amounts of occurrences. It should, however, be noted that terms without any occurrences in the reference data ($N_{true}(t) = 0$) need to be excluded.

Based on the parameters of the 2006 evaluation, the benefit of an additional correct detection is much higher than the cost of a false alert. Therefore, significant amounts of false alerts are tolerable when optimizing for the ATWV metric.

2.3.4. Maximum Term-Weighted Value

Additional to the ATWV, NIST also defines the Maximum Term-Weighted Value (MTWV) as the maximum ATWV achievable by a system using global thresholds.

$$MTWV(\mathcal{T}) = \max_{\theta\in[0,1]} \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\left(\frac{N_{correct}\left(t,\theta\right)}{N_{true}\left(t,\theta\right)} - \beta \cdot \frac{N_{spurious}\left(t,\theta\right)}{N_{NonTarget}\left(t,\theta\right)}\right) \tag{2.9}$$

The definition in 2.9 is very similar to the definition of the ATWV. The actual system decision is, however, neglected and the best value over all possible thresholds is calculated

(hence the dependence on θ in the terms). Although the "best" threshold is chosen, this is not an oracle decision since candidate scores need to be higher than the threshold to be counted as a detection.

At first, one might expect that the MTWV is always higher than the ATWV. This is, however, not the case since identical thresholds over all terms are assumed as opposed to term-specific thresholds as described in 6.2.2.

2.4. Summary

This chapter introduced basic concepts of ASR, the NIST 2006 Spoken Term Detection task and evaluation criteria for detection systems.

The task requires exact detections of terms within the audio data and relies on reference transcripts for judging the system performance. Hence, orthographic details of ASR output heavily influence STD results.

The Actual Term-Weighted Value is the primary evaluation metric for Spoken Term Detection. Rewards for correct detections are much higher than penalties for false alerts. Detection-Error trade-off curves and the Maximum Term-Weighted Value can be used as secondary performance indicators of a system.

3. Related Work

Several tasks have been specified in the area of searching in audio data and information retrieval from speech corpora, both in individual publications and in large evaluations.

Although work on searching in audio data based on spoken queries has been performed for over 30 years (e.g. template matching approaches to word spotting in [CR76]) and is still a research topic (e.g. STD with spoken queries in [CL10]), this overview will focus on fairly recent work using textual queries to search large audio databases.

Although the name "Spoken Term Detection" itself grew primarily due to the NIST 2006 evaluation, basic components and approaches were adapted from the area of Spoken Document Retrieval (SDR) and the corresponding evaluations within the NIST Text REtrieval Conference (TREC).

In the 1998 TREC-7 evaluation, a SDR track consisted of finding relevant audio documents for a topic. Topics were defined by questions such as "How and where is nuclear waste stored in New Mexico?". In [GAV00], NIST researchers consider SDR as somewhat of a "solved problem" due to the fast advances in research, with only question answering and obtaining additional non-verbal information of speech being significant challenges. Therefore, they argue that the additional effort for a SDR track is no longer necessary.

In the overview paper [CHS08], however, Chelba et al. criticize that the recognizers in the SDR evaluation were heavily tuned to the domain and the reported word error rates (WERs) between 10% and 15% are generally unrealistic.

In the SDR tasks, emphasis was placed on information retrieval, such as vector space models, word proximities and language modeling for documents, as described in [CHS08], whereas these components are not necessary for the "more technical" task of Spoken Term Detection.

Generally, textual search on audio documents has been performed long before the 2006 evaluation, but mostly with a focus on information retrieval or integrated into larger systems such as a system for indexing and browsing audio recordings of broadcast news in [MKL+00]. Hence, the Spoken Term Detection task can be seen as as an individual evaluation for a search component used in larger systems.

3.1. General Approaches in the 2006 Evaluation

Major research institutions competed in the NIST 2006 Spoken Term Detection Task, e.g. BBN [MKK+07], IBM [MRS07] and SRI [VSS+07]. Generally, all of these participants

	Broadcast News		Telephone		Meeting	
Participant	ATWV	WER	ATWV	WER	ATWV	WER
BBN	–		**0.8335**	14.9%	–	
IBM	**0.8485**	12.7%	0.7392	19.6%	0.2365	47.4%
SRI	0.8238	23.2%[a]	0.6652	17.4%	**0.2553**	44.2%
TUB	0.3890	30.3%	0.1598		0.0500	60.3%

[a] incl. segments without references, the actual value should be approx. 11% [VSS+07]

Table 3.1.: Results of the NIST 2006 English STD Evaluation and corresponding word error rates on the development data (Excerpt)

used word-based ASR for in-vocabulary query terms and mainly differed in their index structure, approach to out-of-vocabulary words and decision strategies.

BBN participated in the conversational telephone speech (CTS) evaluation in English, Arabic and Mandarin. As reported in [MKK+07], their speech-to-text engine produces both word lattices for retrieval of in-vocabulary terms and 1-best phonetic transcripts for out-of-vocabulary queries.
BBN also developed the decision strategy based on term-specific thresholds (see section 3.3) for in-vocabulary terms, whereas for out-of-vocabulary terms, the top k candidates are returned, with k being a term-independent constant.

IBM who participated in broadcast news (BN), conversational telephone speech and conference meetings (CONFMTG) in English, also follows the approach of combining word-based recognizer output with phonetic transcripts (phonetic transcripts are not used in the meeting part). In their system, described in [MRS07], lattices are transformed in confusion networks (see subsections 3.1.2 and 2.1.2) which are searched for in-vocabulary query terms.
Details on the handling of OOV queries and the decision process are given in subsections 3.2 and 3.3.

SRI also participated in the English BN, CTS and CONFMTG evaluations. Their system as described in [VSS+07] solely uses speech recognizer lattices for search. Out-of-vocabulary words are converted into phones as an intermediate step to get the most similar word in the ASR output. For the binary decision, SRI originally used different approaches including multi-layer perceptrons (see 3.3) and also published scores using term-specific thresholds.

3.1.1. Results of 2006 Evaluation

The results of the evaluation as reported in [FAGD07] and posted on the NIST website[1] can be found in Table 3.1, accompanied by the word error rate on the 2006 development data as reported in the participants' respective publications. Comparing the results, it is evident that recognizer performance as indicated by the word error rate is a major influence factor for the ATWV. In CTS and CONFMTG, the participants with the lowest WER also reach the highest ATWV. In BN, this cannot be directly inferred since SRI did not exclude sections without references from the WER calculation. Although each of BBN, IBM and SRI is leading in one of the domains, it seems as if the neural network-based scoring approach of SRI performed worse than IBM's ranking and score boosting—SRI's error rates in CTS and presumably BN are better than IBM's while the ATWV is lower. IBM achieved the highest ATWV of 0.8485 on the Broadcast News data.

[1] `www.nist.gov/speech/tests/std/`

3.1.2. Lattice-based Approaches

Using an ASR system to convert audio data into textual information raises the question, which form of output should be used as a bases for term search. Although the 1-best transcript already represents a textual form of the audio, experiments in [CSA07] have shown that on lecture data, using information from the lattice (a graph containing alternative word hypotheses) a word error rate (WER) of 30% can be obtained while the single-best transcript results in a WER of 55%. Hence it is evident that indexing lattice information as opposed to the 1-best hypothesis is helpful and improvements in ATWV scores have been shown, e.g. in [VSS+07].

Typically, lattices can contain multiple different paths containing the same detected word for a time segment, therefore a search for unique occurrences or subsequent words in lattices is nontrivial. In [CA05], Chelba et al. propose Position Specific Posterior Lattices (PSPLs) as a lattice representation appropriate for indexing and report that in the context of SDR in the lecture domain, PSPLs achieve a 20% higher value in Mean Average Precision (MAP) compared with single-best transcripts.

Confusion Networks, as proposed by Mangu et al. in [MBS00] and described in 2.1.2, offer an alternative way of storing time-aligned word hypotheses based on lattice information. They offer posterior probabilities and rank amongst alternatives for simultaneous words and have been used both in the context of Spoken Document Retrieval [MCH06] and Spoken Term Detection [MRS07].

3.2. Handling of Out-of-Vocabulary Terms

As mentioned in 1.1, Out-of-Vocabulary words pose a challenge to purely word-based detection approaches, but are also very relevant as user queries.

Several different techniques for handling the problem of out-of-vocabulary query terms have been proposed. In [NZ00], Ng and Zue look into using phone-based subword-units for Spoken Document Retrieval. Using overlapping phone sequences, phonetic classes, non-overlapping phone multigrams and syllable units, they investigate the respective SDR performance on clean transcripts and STT-based transcripts of Broadcast News data. Overlapping phone sequences perform best on the data, roughly 10% better than long multigram and syllable units. Although these subwords units should enable detection of OOV queries, no results on OOV data were reported.

In [LVTM05], Logan et al. examine OOV queries and propose the usage of particles, i.e. variable-length phoneme sequences learned on a corpus, as an indexing unit. As an additional approach, they generate the "most confusable" in-vocabulary words for an OOV query and use them as an expanded query. In SDR experiments on news data, they found that the particle system performed worse than the acoustic query expansion, but with a lower false alarm rate.

In [WFTK08], Wang et al. compare phone-based and grapheme-based STT engines, the latter one similar to work in [KSS03], on English and Spanish in a STD task. While the phone-based system performed better than the grapheme based one on English overall, the latter achieved better results in Spanish. Taking into account the higher word error rate of the grapheme-based system in English, they found that it performed better than the phone-based one at same word error rates.

For the 2006 STD evaluation, BBN and IBM used phonetic transcripts to detect OOV words. Whereas BBN uses the STT engine to directly produce single-best phonetic transcripts [MKK+07], IBM used a STT engine based on sub-word units for recognition and subsequently converted the sub-word lattices into phonetic lattices [SRM06].

BBN performs queries with at least one out-of-vocabulary word by first predicting a pronunciation for the query terms as a reference. Afterwards, possible alignments with the phonetic transcript are generated and accepted if the phonetic edit distance between alignment and reference are less than a certain fraction of the reference length.

In the IBM system, OOV query terms are converted to phones by a Joint Maximum Entropy N-gram Model [Che03] and searched for in the phonetic index. In the system, phones are treated similar to words, which means that phone insertions are allowed as long as a maximum time gap (0.2 seconds between phones as opposed to 0.5 seconds between words) between two adjacent phones is maintained.

Whereas the last two approaches allow a soft match between phonetic transcripts and a single reference pronunciation, in [WKF09] Wang et al. propose a stochastic pronunciation modeling allowing multiple reference pronunciations including different confidences of the respective pronunciations. Using this method yields to an improvement of 17% over an ATWV value of 0.28 with a single reference pronunciation.

Instead of using a phone edit distance to obtain the similarity between two pronunciations, [CP07] proposes higher order confusions based on phone N-grams. Recall and precision were improved by using higher order confusions when searching a phone n-gram index.

An alternative approach to OOV detection in [RSM+09] is the use of both word and fragment-based sub-words units in the ASR vocabulary and using fuzzy search. Results on the 2006 corpus showed that such a flat hybrid system is more robust at higher word error rates than a purely word-based one.

A flat hybrid model has also been proposed by Akbacak et al. in [AVS08] for OOV detection in the STD task. They use "graphones" as a sub-word unit which have been introduced by [BN05] and are based on the assumption of an underlying base unit for orthographic representation and pronunciation of a word, consisting of pairs of short sequences of letters phonemes. About half of the original OOVs can be recovered by adding graphones and improvements for in-vocabulary words are also reported. On a 60k word vocabulary with a word error rate of 14%, they achieve an OOV-ATWV of 0.25.

In [PSR09], Parada et al. use a Query-by-Example system to detect OOV words in (textual) Spoken Term Detection. Using a two-step process, they first perform an STD search using grapheme-to-phoneme conversions for query terms and subsequently retrieve audio samples of query terms by lattice cuts of high scoring results. In a second step, acoustic lattices are searched to find occurrences corresponding to the samples. They report improvements in ATWV when adding the second-pass results to the purely text-based search.

3.3. Decision Strategies

An essential part of a STD system is the binary decision whether the system expects a detection candidate to be an actual occurrence. Several different approaches have been developed for and after the 2006 evaluation.

In [MKK+07], Miller et al. introduce term-specific thresholds (TSTs) as a decision process tuned to the ATWV metric. The concept—described in detail in section 6.2— adapts the detection threshold for each term to the respective candidate set and generally results in lower thresholds for rare terms and vice versa.

In [VSS+07], Vergyri et al. compare global thresholds, term-specific thresholds and a neural network-based regression as decision strategies and find that TSTs give the best results and generalization. In the 2006 evaluation they relied on the neural network based

approach which worked well on the development set but did not generalize well to the evaluation meeting data. In the workshop presentation, overtraining is given as a possible cause [VSGW06].

IBM [MRS07] used a different scoring and decision strategy in their 2006 system. While term-specific thresholds are dependent only on the posterior probabilities, IBM uses the rank of a word detection (within all hypotheses starting at the same time) as an explicit factor in their score. For phonetic matches, where they only use single-best transcripts, the score is based on the proximity of detected phones (for details, see [MRS07, section 3.4.2]).
For multi-word queries, the multiplied word scores are boosted by an empirically determined exponent of $1/n$ to account for the claimed lower probability of false alerts on multi-word queries. Generally, IBM uses global thresholds for each audio category for their final decision.

In [CS09], Can et al. propose a thresholding method based on the estimation of an exponential mixture model of the class (correct/incorrect) of a detection. Evaluations on Turkish news programs show that this approach can achieve a higher maximum precision than—but is typically outperformed by—term-specific thresholds.

Typically, the posterior scores (confidences) of the STT engine, which do not depend on the query term, are used as the sole feature for the decision process. In [WKFB09], Wang et al. propose a different approach using term-dependent confidence measures. They formulate the idea of term-specific thresholds as term-specific confidence transformations (with a fixed global threshold) instead. By modifying these transformation functions with a multilayer perceptron or a support-vector machine, optionally taking into account the estimated false alarm and occurrence rate, improvements were shown in ATWV on individual-microphone meeting data when using a phoneme-based recognition system. However, the impact on a (generally better performing) word-based ASR system is reported as "very little".

In [CLYL10], Chen et al. propose an approach for reranking query results in a STD/SDR-like task. Based on acoustic similarity among candidates calculated with dynamic time warping, they use a pseudo-relevance feedback technique to rerank results and achieve improvements in Mean Average Precision (MAP). In [CCL+11], a graph-based version of this approach is adapted, building on similar work in video reranking in [HKC07]. Using MAP as an evaluation metric—hence considering the result rank—they report improvements of 20% over a baseline only based on posterior scores. A similar approach has been adapted for this work and is described in 6.1.

3.4. Summary

Spoken Term Detection has been a popular research topic since the 2006 NIST evaluation, but has also been used before as a component in larger systems. Typically, an ASR system is used to transcribe the audio and generate lattice representations which, in case of in-vocabulary terms, are searched for occurrences.

For out-of-vocabulary terms, different strategies have been used. Many are based on predicting a pronunciation for the term and searching phonetic transcripts for a similar phoneme sequence. Other approaches use STT engines based on sub-word units like fragments or graphones to detect OOV terms.

As a decision strategy, term-specific thresholds have been found to be well-performing in the STD task. They offer substantial improvements over global thresholds.

4. Design and Implementation Framework

This chapter describes both the architecture and implementation details of the Spoken Term Detection system developed for this thesis.
After an overview of the general architecture, the different steps in processing the audio, indexing word hypotheses and querying the database are explained in the following sections.

4.1. Basic Setup and Design Considerations

The Spoken Term Detection system can be split up into parts for indexing and for querying the index. The top half of Figure 4.1 shows the indexing part. The audio data is processed by a speech recognizer which produces *confusion networks* (cf. subsection 2.1.2). In the next step, these confusion networks are used by the indexing engine to insert all word hypotheses into the CouchDB database which processes them for the query engine (cf. 4.3).
During the querying (the bottom half of Figure 4.1) all terms of a given term list are searched for in the database, all candidates are scored and, with a binary decision, output to the detection list.

In order to ease modifications and reduce coupling, the developed system has been modularized along functional borders. The rough structure corresponds to Figure 4.1, with finer substructures in the more complex components, e.g. for the approaches in 5.3 or 6.1.

For indexing and storing, CouchDB[1] is used. As an "off-the-shelve" open-source software, it is reasonably well documented and the use of HTTP as a protocol allows for easy access from different software environments or hosts.

TCL has been used as the primary language for implementation. This allows easy integration into the existing ASR environment. Parts of the rescoring approach described in section 6.1 have been implemented in C within the Janus ASR system to reuse existing audio and feature infrastructure. Small portions of JavaScript and Scheme have been used for defining the CouchDB views (4.3) and accessing the Festival text-to-speech system (5.3), respectively.

[1] `http://couchdb.apache.org/`

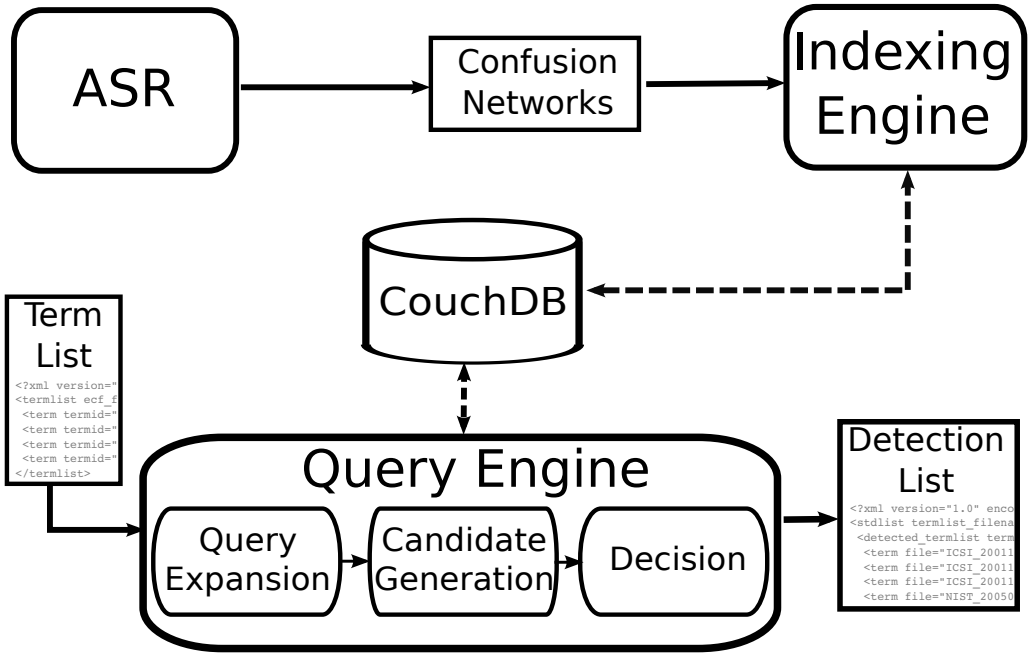


Figure 4.1.: Basic Setup of Spoken Term Detection system with major components

4.2. Decoding and Indexing

The audio data is processed by the Ibis Decoder of the Janus Speech Recognition Toolkit [SMFW01]. Details concerning the different decoder setups can be found in section 7.1. After decoding of an utterance, the hypotheses are transformed into confusion networks (cf. 2.1.2) and subsequently inserted into the database.

4.3. Database Setup

CouchDB is used for storing and indexing the word hypotheses of the confusion networks. CouchDB can be accessed via the HyperText Transport Protocol (HTTP). This enables direct access to the database from within Janus instances and ensures compatibility with a wide variety of programming environments.

CouchDB is a non-relational database that is based on documents, each consisting of key-value pairs. In the STD setup, each document represents a single hypothesis word alongside metadata such as start and end time and posterior probability.

Instead of supporting dynamic queries in a particular query language, CouchDB offers a MapReduce-based framework, in which custom JavaScript functions can be used to build "views" on the data. Each view consists of key-value mappings that are indexed by the key. Since the database serves only as storage in the STD task, only two views are defined, both consisting only of a map function. One view, $V_w(word)$, uses the hypothesis word as a key and is used to retrieve candidate sets for the first word of a query. For subsequent words in multi-word queries, an additional view, $V_{wu}(word, utt)$, consisting of the word-utterance pairs as keys is used. It allows searching for words in specific utterances.

4.4. Querying

Given a search term, the view V_w is queried for occurrences of the term's first word in the complete corpus. Results contain the start and end times of the word detections, an ASR posterior score, the original file name and the utterance name from the Janus database.

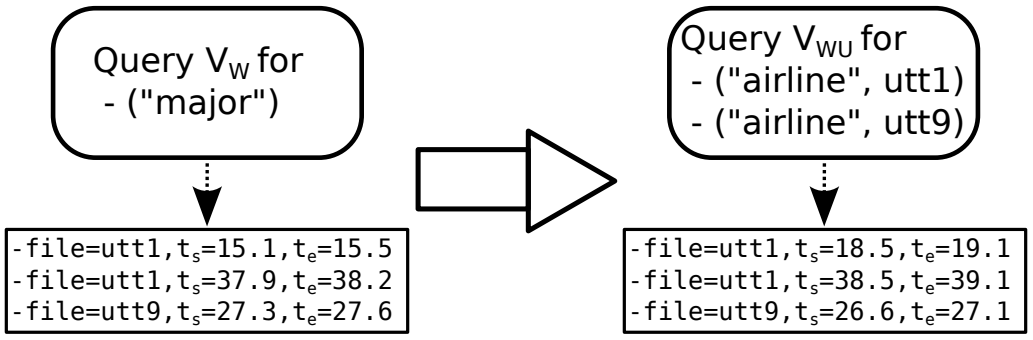


Figure 4.2.: Example of a query process for the term "major airline"

4.4.1. Treatment of Multi-Word Queries

Multi-word queries are treated as a sequence of single word queries. According to the evaluation criteria in [Nat06], the time between a query word w_i and its successor w_{i+1} must not be longer than $T_{gap} = 0.5s$ in the reference transcript. Since the STT system might have recognized other words in such a time period, no general assumptions about direct succession in the confusion networks can be made.

Therefore, for each hit $h_k = (w_i, start_k, end_k, utt_k, posterior_k)$ of word w_i (retrieved by querying view $V_w(w_i)$), the view $V_{wu}(w_{i+1}, utt_k)$ is queried for all occurrences $h_l = (w_{i+1}, start_l, end_l, utt_l, posterior_l)$ of the subsequent query words in the same utterance. If the word w_{i+1} occurs within the allowed time, i.e. $start_l - end_k \leq T_{gap}$, the word is matched and the querying continues with w_{i+2} or a new detection is reported if w_{i+1} is the last word of the query. If $start_l - end_k > T_{gap}$, the possible detection is rejected.

4.4.2. Example

Figure 4.2 shows an exemplary querying process for the two-word term "major airline". First, all occurrences for the first word are retrieved and as a second step, follow-up words are found. The first result of the second query does not start within 0.5 seconds of the end of an occurrence of the first word and is rejected. The second results of the queries match and hence form a detection candidate. The third results occur within 0.5 seconds of each other, but in the wrong order and are consequently rejected. Hence, one possible term occurrence is detected.

4.5. Summary

The implemented STD system is divided into two parts, an indexing and a query engine. The first one processes confusion networks and inserts them into a document-oriented database while the second one can be used to query this database for term occurrences.

Querying is performed on a word level, disregarding possibly inserted words and finding word sequences based on the maximal time gap allowed in the evaluation.

5. Query Expansion for Out-of-Vocabulary Words

The term *query expansion* describes the process of adding related terms to the original query in order to increase the result set. The NIST Spoken Term Detection task uses the exact orthographic transcription of words to identify hits, therefore a query expansion using e.g. stemmed word forms or synonyms would actually be counterproductive. For in-vocabulary words, query expansion typically leads to more false alerts. However, some query expansions with normalized forms can be beneficial and expansion with similar words can be used to find out-of-vocabulary words.

Generally, the original query and expanded queries should be treated differently in the decision process since query expansions are potentially error-prone. For example when replacing an OOV word with a similar sounding one, the degree of similarity should be taken into account.

Hence, the score of a detected candidate c for a query term w with ASR confidence $P(c)$ is obtained by $Score(c) = P(c) \cdot Sim(w, c)$. The function $Sim(w, c)$ represents the quality of the query expansion (the similarity to the actual term) and depends on the type of query expansion. Different variants will be mentioned in the following subsections and experimental results of the strategies are presented in chapter 7.

It should be noted that all of these approaches operate on the same confusion network output. Since no adaptation or retraining of the ASR system is necessary, the STD system can be easily adapted to different setups and languages.

5.1. Normalization

For example, the hyphenation of terms in the recognizer output is typically not completely conforming to dictionaries and query input. Query terms such as “brother-in-law” are often recognized as multiple words (“brother in law”). Therefore, the STD system expands hyphenated terms into their components and uses results of both query terms as output. It can be configured to penalize results from the expanded query. Typically, however, $Sim_{norm} \equiv 1$ is used for the normalizations unless otherwise noted.

Similarly, the system can be configured to normalize apostrophes. Although this can occasionally avoid OOV words (e.g. “form’s”), it can also add additional ones (e.g. “he’s”).

Levenshtein Similarity

alexis	0.8333	alessio	0.6923	alexei	0.6667	alex	0.6
alexi	0.8182	flexion	0.6923	alex's	0.6667	aleo	0.6

Dice coefficient

alexi	0.8889	alex	0.7500	flexion	0.7273	exiles	0.6
alexis	0.8000	lexical	0.7273	exile	0.6667	exiled	0.6

Figure 5.1.: Most similar words for the out-of-vocabulary search term "alexio" based on 55k vocabulary

Additionally, the system can be configured to use GLM files containing mappings between different orthographic forms of the same words, e.g. expanding the title "dr" to the word "doctor" which is typically found in the recognizer results.

Depending on the dictionary used by the speech recognizer and the orthographic format of the queries, it can be necessary to perform additional normalizations. For example, the NIST 2006 term lists spell acronyms as a series of one-letter-words ("u. s. a.", "t. v."), whereas the Quaero ASR system (see section 7.1.2) uses a single word spelling ("USA", "TV").

5.2. Orthographic Query Expansion

As mentioned in section 3.2, there are several approaches to find occurences of query terms that are not in the recognizer dictionary. One such approach is identifying words in the recognizer vocabulary which are orthographically similar to the query word.

Orthographic similarity can be defined in a multitude of different ways. Two arguably common ones are string similarities based on Levenshtein distance and the Dice coefficient. Based on these similiarity measures, additional searches are performed and subsequently their results are combined.

5.2.1. Levenshtein distance

The Levenshtein distance between two strings is an edit distance based on the minimum number of substitions, insertions or deletions necessary to transform one string into the other (5.1). As opposed to the approach in 5.3, uniform substitution penalties are used due to the difficulty of estimating the confusability of single letters. This distance measure can then be used to retrieve the words of the vocabulary most "similar" to a particular out-of-vocabulary query term .

$$D_{lev}(s_1, s_2) = N_{sub} + N_{ins} + N_{del} \tag{5.1}$$

For the decision process, however, this distance measure cannot be directly integrated with the posterior probabilities into a final score. Additionally, the Levenshtein distance itself is highly dependent on the length of words compared and should therefore be normalized. One possibility is to subtract the number of character edit operations necessary for transforming both strings into each other from the number of total characters and normalize them based on the total length, as seen in Equation 5.2 where $|\ldots|$ represents the number of characters in a string.

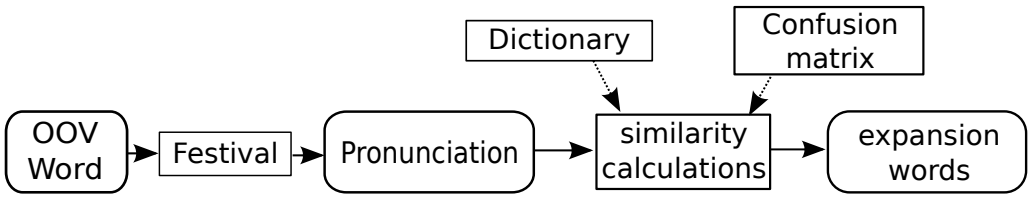


Figure 5.2.: Phone-based Query Expansion

$$Sim_{lev}(s_1, s_2) = \frac{|s_1| + |s_2| - 2 \cdot D_{lev}(s_1, s_2)}{|s_1| + |s_2|} \tag{5.2}$$

An example of word similarities can be found in Figure 5.1.

5.2.2. Dice coefficient

An alternative similarity measure based on the orthographic word representation is the dice coefficient as seen in Equation 5.3. It emphasizes the number of shared character bigrams of two words. Apart from the bigram structure, the order of characters is, however, discarded.

$$Sim_{dice}(s_1, s_2) = \frac{2 \cdot |B_1 \cap B_2|}{|B_1| + |B_2|} \qquad B_i = \{c_j c_{j+1} | c_j \textit{ is } j^{th} \textit{ character of } w_i\} \tag{5.3}$$

An example of word similarity based on the dice coefficient can be found in Figure 5.1.

5.3. Phone-based Query Expansion

An alternative to using orthographically similar words for query expansion is the utilization of the phonetic representations of words to find phonetically similar ones. All words in the recognizer vocabulary are accompanied by their respective pronunciation. The (presumptive) pronunciation of out-of-vocabulary words can be obtained by using speech synthesis tools such as Festival ([TBC98]) which include letter-to-sound rules for unknown words.

The general process is depicted in Figure 5.2. As a first step in the phoneme-based query expansion approach, a pronunciation is generated by Festival. Subsequently, the similarity of this reference pronunciation and the pronunciation of every vocabulary word is calculated and the k most similar words are used for query expansion.

Since the speech synthesis component of Festival is not needed, the pronunciation can be retrieved by a `lex.lookup` command and possibly necessary conversions from the Festival phone set to the speech recognizer phone set are performed subsequently.

To calculate the similarity Sim_{phone} between two pronunciations, a dynamic programming-based algorithm is used. The probabilities of substituting two phones are based on the confusion matrices in [CWSC04] and the probability of phone insertions and deletions can be parameterized.

It should be noted that, opposed to the approach in section 5.2, the phone-based query expansion is not language-independent. It relies on the availability of letter-to-sound rules (included within Festival) and phone confusion estimates for the particular language or dialect.

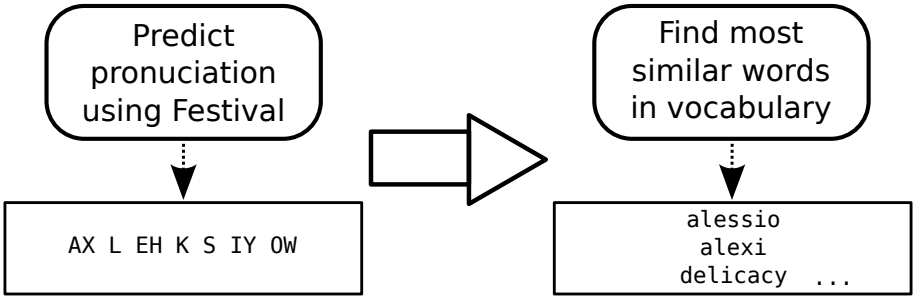


Figure 5.3.: Example of phone-based Query Expansion for the OOV word “alexio”

Figure 5.3 shows an example for phone-based query expansion. First a pronunciation for the word is predicted by Festival's letter-to-sound rules and, if necessary, mapped to the recognizer phone set. Subsequently, the similarity of this predicted pronunciation to all words in the recognizer dictionary is calculated and the top results are returned and used for query expansion.

5.4. Summary

Queries containing out-of-vocabulary words can be expanded to (multiple) similar words and the search is performed with expanded queries, taking into account the degree of similarity to the search term.

Expansion can be performed based on text normalizations and orthographic or phonetic similarity to in-vocabulary words. The results of the corresponding implementations described in this chapter can be found in chapter 7.

6. Decision Strategies

As depicted in Figure 4.1, the decision process is the final step while performing a term query. It includes possible modifications of the detection confidences as described in section 6.1 and the binary decision process for each detection candidate as described in section 6.2. Results and evaluations of the approaches in this chapter are presented in chapter 7.

6.1. Rescoring Based on Acoustic Similarity

The approach in this section utilizes the acoustic similarities among the detection candidate set of a particular term to perform a random walk over a candidate context graph and update the candidates' confidences.

Generally, the scores of candidate detections are based on the posterior probabilities (confidences) of the STT system. There are, however, approaches that also include information from the candidate set, especially the similarity of different candidates.

Relevance Feedback is a popular technique in information retrieval to refine queries. Based on an initial result set, a user identifies relevant results. Features of these relevant results (e.g. term frequencies) are then used to refine the query or reweigh the results, yielding substantial improvements [SB90].

Pseudo-Relevance Feedback (PRF) is a method to apply the concepts of relevance feedback without the need for user interaction. It assumes that the top results (based on some scoring metric) of a query are the most relevant [CYF+97]. Hence, other rather low-scoring candidates that are very similar to the "relevant" top results should be given a higher score (or better rank in ranked result lists). In [CCL+11], Chen et al. propose a PRF approach to Spoken Term Detection based on acoustic similarity which has been used and adapted for this work.

Using Dynamic Time Warping as a similarity measure (cf. 6.1.1.1), results are rescored based on their acoustic similarity, but also taking into account their prior ASR confidence score. Using a random walk over a candidate context graph adapted from [HKC07], new scores are calculated and integrated into the candidate score.

This rescoring approach is different from the rest of this work in the way that it takes into account the audio underlying candidate detections and hence uses additional information not stored in the CouchDB index.

6.1.1. Acoustic Similarity

The pairwise similarity between all audio segments c_i, c_j from the set of detection candidates $\mathcal{C}$ is calculated based on a distance measure d between the lMEL feature vectors of the respective audio segments as shown in Equation 6.1. Two different distance measures have been implemented, d_{DTW} as described in 6.1.1.1 and $d_{SelfSim}$ as described in 6.1.1.2.

$$Sim(c_i, c_j) = 1 - \frac{d(c_i, c_j) - d_{min}}{d_{max} - d_{min}} \tag{6.1}$$
$$d_{min} = \min_{c_k, c_l \in \mathcal{C}} d(c_k, c_l), \quad d_{max} = \max_{c_k, c_l \in \mathcal{C}} d(c_k, c_l)$$

6.1.1.1. Dynamic Time Warping-based Distance

The distance measure d_{DTW} between two audio segments is based on Dynamic Time Warping (DTW). DTW is based on the frame-wise (Euclidean) distance between feature vectors. However, to allow for time variations, a frame alignment between the two segments is calculated with Dynamic Programming. For details, see e.g. [SC78]. No constraints have been used to find the DTW path.

6.1.1.2. Self-Similarity-based Distance

An additional distance measure $d_{SelfSim}$ between two audio segments is based on the concept of Self-Similarity Matrices (SSMs) as proposed by Muscariello et al. in [MGB11]. SSMs can be seen as "fingerprints" of audio data and for example cancel out constant additive noise within the audio.

Given a candidate audio segment c with n frames, with $(F(c))_i$ being the feature vector of the i-th frame, the Self-Similarity matrix $M_{SSM}(c)$ of this segment is based on the pair-wise Euclidean distance between the frames' feature vectors as shown in Equation 6.2.

$$M_{SSM}(c) = \Big(m_{ij}\Big), \quad m_{ij} = ||(F(c))_i - (F(c))_j||_2, \quad 1 \leq i, j \leq n \tag{6.2}$$

Interpreting this matrix as an image, a feature vector $\Phi(M_{SSM}(c))$ can be calculated based on Histograms of Oriented Gradients (HOGs) that have been introduced by Dalal et al. and shown to be useful in image classification tasks [DT05]. For details on the calculation of these feature vectors, see [DT05] and [MGB11]. Since the size of the feature vectors vary based on matrix dimensions, two SSMs $M_{SSM}(c_i)$, $M_{SSM}(c_j)$ are first time-aligned by Dynamic Time Warping to create two matrices of identical dimensions, $M'_{SSM}(c_i)$ and $M'_{SSM}(c_j)$.

$$d_{SelfSim}(c_i, c_j) = \Big|\Big|\Phi(M'_{SSM}(c_i)) - \Phi(M'_{SSM}(c_j))\Big|\Big|_1 \tag{6.3}$$

The distance between two candidate audio segments is then defined as the 1-norm of the difference of their HOG feature vectors as depicted in 6.3.

6.1.2. Constructing a Candidate Context Graph

As in [CCL+11], a complete directed graph is calculated for the candidate set $\mathcal{C}$. Each candidate $c \in \mathcal{C}$ corresponds to a vertex and the edges are weighted based on the similarity of their incident vertices.

$$A(\mathcal{C}) = \Big(a(c_i, c_j)\Big), \quad a(c_i, c_j) = \frac{Sim(c_i, c_j)}{\sum_{c_k \in \mathcal{C} \setminus \{c_i\}} Sim(c_i, c_k)} \tag{6.4}$$

The weight of a directed edge (c_i, c_j) is the acoustic similarity between the corresponding candidates as defined in 6.1.1, normalized over all outgoing edges of c_i. Equation 6.4 shows the edge weights in the form of the adjacency matrix $A(\mathcal{C})$ of the candidate context graph.

6.1.3. Score Update

The score update $u(c_i)$ for each candidate (vertex) is then defined recursively as the interpolation of a candidate's original confidence score $s_{ASR}(c_i)$ and the other candidates' updated score, weighted by the similarity to the respective other candidate over the incoming edge (c_j, c_i) as seen in Equation 6.5. $s'_{ASR}(c_i) := s_{ASR}(c_i) / \sum_{c_k \in \mathcal{C}} s_{ASR}(c_k)$ is the normalized confidence score and α is an interpolation weight to adjust the influence of the original score.

$$u(c_i) := (1 - \alpha) \cdot s'_{ASR}(c_i) + \alpha \sum_{c_j \in \mathcal{C} \setminus \{c_i\}} a(c_j, c_i) \cdot u(c_j) \tag{6.5}$$

This update formula 6.5 can be seen as a form of Pseudo-Relevance Feedback. Each candidate score is based on the original confidence score, but the scores of candidates that are similar to other "high-scoring" (i.e. top-ranking) candidates are boosted by the second summand.

Based on the observation that $\forall c_i \in \mathcal{C} : \sum_{c_j \in \mathcal{C} \setminus \{c_i\}} a(c_i, c_j) = 1$, the adjacency matrix $A(\mathcal{C})$ of the candidate context graph can also be seen as a stochastic matrix defining path probabilities. The solution to 6.5 can then be found as the stationary probabilities after a random walk [HKC07].

Converting Equation 6.5 in vector form yields Equation 6.6. $\mathbf{u} := (u(c_1), \ldots, u(c_{|\mathcal{C}|}))$ is the vector of all candidates' score updates, $\mathbf{s'_{ASR}} := (s'_{ASR}(c_1), \ldots, s'_{ASR}(c_{|\mathcal{C}|}))$ is the vector of all candidates' normalized original confidence scores and $\mathbf{e} = (1, \ldots, 1)$ is the $|\mathcal{C}|$-dimensional 1-vector. 6.7 follows from $\sum_{c_i \in \mathcal{C}} u(c_i) = 1$ (cf. Equation 6.5).

$$\mathbf{u} = (1 - \alpha) \cdot \mathbf{s'_{ASR}} + \alpha (A(\mathcal{C}))^\top \mathbf{u} \tag{6.6}$$

$$= \underbrace{\Big((1 - \alpha)\mathbf{s'_{ASR}}\mathbf{e}^\top + \alpha (A(\mathcal{C}))^\top\Big)}_{=: Q} \mathbf{u} \tag{6.7}$$

Equation 6.7 states an eigenvector problem that is similar to the PageRank problem [CCL+11]. It is shown in [HKC07] and [LM05] that, based on the structure of Q, $\mathbf{u}$ exists and is unique.

The update $u(c_i)$ can be interpreted as a distribution of the original candidate probability mass based on acoustic similarity as defined by d_{DTW} or $d_{SelfSim}$.

$$s_{Rescored}(c_i) = s_{ASR}(c_i) \cdot (u(c_i))^\delta \tag{6.8}$$

The new score $s_{Rescored}$ for each detection candidate is calculated as seen in 6.8 based on the score update and a smoothing parameter δ, typically set to 1. However, $s_{Rescored}$ cannot necessarily be interpreted as the probability of a correct detection any more.

Hence, experiments have been performed with different types of additional score processing, for example histogram equalization-like normalization on the whole probability range or the original probability range.

Results of the rescoring approaches are reported in 7.2.5.

6.2. Thresholds

The final step in the Spoken Term Detection process is the binary decision, whether a candidate should be reported by the system as a detection or not. This is done based on thresholds—a detection is reported if and only if the candidate score is greater or equal to a threshold. Both a global threshold that is identical for every term and term-specific thresholds have been implemented.

6.2.1. Global Threshold

A natural way for the decision process is the use of a global threshold (GT). Based on a held-out development dataset, the global threshold that maximizes the ATWV is chosen.

This simple approach is assumed by NIST and the evaluation and diagnostic tools are optimized for it. DET curves as described in 2.3.2 are based on the assumption of global thresholds and a Maximum Term-Weighted Value (MTWV) based on the ATWV, but using the best global threshold instead of the system decision.

6.2.2. Term-Specific Thresholds

Term-specific Thresholds (TSTs) are a decision strategy specifically optimized for the ATWV metric and have been first published by [MKK+07]. They are based on two ideas. First, every term contributes equally to the ATWV score. Hence, thresholds should also be optimized/adapted for every term. Second, the detection score, interpreted as the probability of a correct detection, should be used to adapt the threshold to the cost/benefit ratio of the evaluation metric.

Let $I_{corr}(c)$ be an indicator variable that is 1 if candidate c is correct detection and 0 if c is spurious detection. Then the definition of the ATWV for a term list $\mathcal{T}$ in 6.9 can be reformulated as a sum over all detection candidates $\mathcal{C}(t)$ of a term $t \in \mathcal{T}$ as in Equation 6.10.

$$ATWV\left(\mathcal{T}\right) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left(\frac{N_{correct}\left(t\right)}{N_{true}\left(t\right)} - \beta \cdot \frac{N_{spurious}\left(t\right)}{N_{NonTarget}\left(t\right)} \right) \tag{6.9}$$

$$= \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}(t)} \left(\frac{I_{corr}(c)}{N_{true}\left(t\right)} - \beta \cdot \frac{1 - I_{corr}(c)}{N_{NonTarget}\left(t\right)} \right) \tag{6.10}$$

It is evident that the expected value of each summand in the inner sum of Equation 6.10 should be maximized in order to maximize the ATWV. Furthermore, each summand should be greater than or equal to zero in order to improve the ATWV.

Assuming that a candidate score $s(c_i)$ represents the actual probability of a correct detection (and hence $1 - s(c_i)$ the probability of a spurious detection), a threshold θ should be chosen so that the expected value of each inner summand in 6.10 is greater or equal zero for $s(c_i) \geq \theta$. Assuming a general benefit B_{hit} for a correct detection and a cost C_{fa} for a false alert, this leads to Equation 6.11.

$$\theta(t) \cdot B_{hit} - (1 - \theta(t)) \cdot C_{fa} \stackrel{!}{=} 0 \iff \theta(t) = \frac{C_{fa}}{B_{hit} + C_{fa}} \tag{6.11}$$

Using the parameters of the ATWV evaluation metric for B_{hit} and C_{fa}, this leads to the term-specific threshold in Equation 6.13.

$$B_{hit} = \frac{1}{N_{true}(t)}, \quad C_{fa} = \frac{\beta}{N_{NonTarget}(t)}$$

$$\Longrightarrow \theta(t) = \frac{\beta \cdot N_{true}(t)}{N_{NonTarget}(t) + \beta \cdot N_{true}(t)} \tag{6.12}$$

$$= \frac{\beta \cdot N_{true}(t)}{T_{speech} + (\beta - 1) \cdot N_{true}(t)} \tag{6.13}$$

The number of true occurrences $N_{true}(t)$ from Equation 6.13 is obviously unknown during the decision process. It can, however, be approximated by the sum of the candidates' scores. This expected value holds only true if the scores are a reasonable approximation of the actual probability of a correct detection.

6.3. Summary

In the decision process, a rescoring based on acoustic similarity between detection candidates can be performed. Similar to pseudo-relevance feedback, scores of detections similar to the best search results are boosted.

Term-specific thresholds as an ATWV-oriented concept are motivated and introduced as the primary decision strategy of the system. These concepts have been implemented and are evaluated in chapter 7.

7. Experimental Evaluation

This chapter describes the experiments performed with the previously described configurations. The data and setup used for the experiments is described and ASR performance as a basis of STD is evaluated. Subsequently, the results of different approaches are described and contrasted. At the end of the chapter, findings are evaluated in context.

7.1. Experimental Setup

The experiments in this chapter have been performed on the NIST 2006 development and DryRun data. Due to the similarity of corpus and term lists, this enables a comparison with the final 2006 NIST STD competition and an evaluation in the context of the 2006 results.

The experiments are relying on an available speech-to-text system as a prerequisite. Therefore, the experiments have been limited to the English broadcast news and meeting data.

7.1.1. Evaluation Data

The audio data used in the experiments consists of approximately three hours of English broadcast news and about two hours of round-table meetings.

The broadcast news data consists of six programs recorded in February 2001.
The meeting data consists of ten single channel recordings based on distantly placed microphones and has been produced in 2001 and 2005.
More details on the used data can be found in the NIST STD 2006 Evaluation Plan [Nat06].

Two term lists provided by NIST have been used for the experiments. The DryRun term list contains 1108 terms and the Dev term lists contains 1100 terms. As listed in Table 7.1,

	# of words in query				
Term List	**1**	**2**	**3**	**4**	**Average**
DryRun	597	410	68	32	1.55
Dev	599	400	85	15	1.56

Table 7.1.: Length of Queries

System	Vocab size	WER BNews	WER Meeting
TUB 2006	47k	30.31%	60.25%
Quaero 1^{rst} pass	128k (104k ci)	22.99%	65.40%
Quaero adapted	128k (104k ci)	19.88%	64.69%

Table 7.2.: Vocabulary sizes and Word Error Rates of different recognizer systems

both lists contain primarily short queries (about 1.6 words on average) and the queries consists of four words at most.

Generally, all words in multi-word queries also occur as a single-word query in the 2006 term lists. Although this approach allows a separate view on each word in a multi-word query, it also results in questionable queries for terms such as "e.s" that are contained in multi-word queries like "g. r. e.s aren't".

7.1.2. Speech Recognition Systems

Experiments were performed with hypotheses produced by two different speech-to-text systems.

The first STT system used is identical to the one used by the TU Berlin submission in 2006. It will be referred to as the "TUB" system subsequently. It is based on the first pass setup of ISL RT2004 Meeting system [MFPW05].

The second STT system has been developed for the Quaero evaluations and is therefore referred to as the "Quaero" system. It is based on the MFCC system with CMU dictionary from [SKN11]. Hypotheses from both the first run and an adapted second run (referred to as "Quaero Adapted") have been used.

The vocabulary sizes differ significantly as can be seen in Table 7.2. The TUB system has a case insensitive vocabulary, whereas the Quaero vocabulary is case sensitive. Based on the number of (case insensitive) words, the Quaero vocabulary is more than twice as large as the TUB one.

Table 7.3 contains an overview of the number out-of-vocabulary words in corpus and queries. The OOV rate on the corpus is relatively low (around 3%). Although there are more OOVs in the query terms (between 4% and 10%), it is noteworthy that the relation between query OOV rate and corpus OOV rate is much lower than Logan et al. report in a scenario with actual user queries [LMTW00].

There is, however, a mismatch between the Quaero vocabulary and the query term structure concerning the handling of acronyms. The Quaero vocabulary does not contain any single letters and treats acronyms as single (capitalized) words ("TV"), whereas the TUB vocabulary and the term list split acronyms into single letters ("t. v."). While this mismatch can be normalized by splitting acronyms in the Quaero hypotheses, single letters lead to the surprisingly high OOV rate on the data despite the higher vocabulary size of

	# of OOV words in corpus		# of OOV queries	
System	BNews	Meeting	DryRun	Dev
TUB	341 (1.38%)	583 (2.59%)	66 (5.96%)	57 (5.18%)
Quaero	774 (3.14%)	666 (2.96%)	105 (9.48%)	47 (4.24%)

Table 7.3.: Statistics of Out-of-vocabulary words in corpus and term lists

the Quaero system. Additionally, they also affect the scores due to numerous single letter queries (as mentioned in 7.1.1).

7.1.3. Recognizer Performance

The different systems perform significantly different on the audio data used in the experiments. The word error rates on broadcast news, as listed in Table 7.2, range from 30.31% with the TUB system to 19.88% with the adapted Quaero system. On the meeting data, the TUB system performs best with a 60.25% word error rate, the adapted Quaero system achieves 65.40% with the first-pass result being slightly worse.

The results on meeting data are much worse than on the broadcast news, severely effecting the ability of term detection. This is, however, not a particular problem of the used recognition systems. As shown in Table 3.1, the participants of the 2006 evaluation also reported word error rates of over 40%.

The word error rate as a whole can only give a single impression of the quality of the recognizer in the STD task, since its components, deletions, insertions and substitutions, contribute differently to the detection task.

Deletions, i.e. words not detected at all, generally result in a lower hit rate since less candidates for a particular query are detected. Insertions, i.e. words in the hypotheseis which are not in the reference, do not affect the number of correct detections, but increase the false positive rate, therefore decreasing the ATWV score. Substitutions both lower the correct detection rate and increase the number of false positives.

If term-specific thresholds are used, additional elements need to be considered. Since the number of true occurences in formula 6.13 depends on the posterior sum of all candidates, an increase in the number of candidates results in a higher threshold, typically decreasing the number of correct detections.

The use of confusion networks reduces the influence of the previously mentioned errors in the single-best transcript and the word error rate can be only an indicator of recognition quality.

7.2. Results

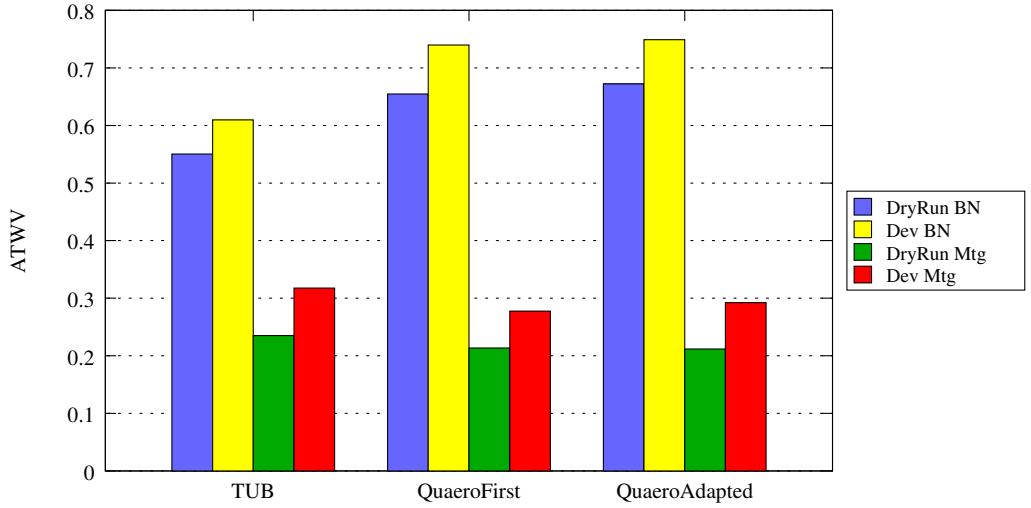
This section contains the results of the approaches described in previous chapters on the 2006 development data. Different approaches and configurations are primarily evaluated based on the ATWV as defined in 2.3.3. More detailed results with overall detection statistics can be found in the appendix.

7.2.1. Overall results

An overview over the ATWV results of the basic approaches can be found in Figure 7.1. It contains results of the same STD system configuration (phone-based expansion as described in 5.3) on different data. The different groups corresponds to the different STT systems generating the hypotheses.

First, it can be observed that within each of the three groups (each system) the order of ATWV scores on different term lists is identical. The Dev term list is “easier” in the way that all systems achieve higher scores on it than on the DryRun term list.

Second, all systems perform much better on Broadcast News than on Meeting data. Considering the WERs from Table 7.2, this is not very surprising. Similarly, the Quaero systems perform better on BN data and worse on Meeting data compared with the TUB



System	Broadcast News ATWVs		Meeting ATWVs	
	DryRun	Dev	DryRun	Dev
TUB	0.5503	0.6097	**0.2351**	**0.3175**
QuaeroFirst	0.6548	0.7398	0.2135	0.2776
QuaeroAdapted	**0.6722**	**0.7491**	0.2119	0.2924

Figure 7.1.: Overview of ATWV results of different STT systems and term lists

system. The ASR confidences of the meeting hypotheses are quite high compared with the recognition accuracy. Therefore, an empirically chosen confidence penalty of 0.15 has been applied to all term occurrences in meeting data.

More details on system performance on the DryRun and Dev term lists can be found in A and B, respectively.

7.2.2. Influence of Word Error Rate

As mentioned in the previous subsection, the word error rate plays a critical role in the achievable ATWV value. Figure 7.2 shows the (monotone) relationship between word error rate and ATWV. All experiments in the plot use an identical STD system configuration and vary only in the accuracy of the underlying confusion network hypotheses. Different accuracies based on the same ASR system can be produced by varying the master-beam setting in the Ibis decoder, i.e. adjusting the search-depth of the recognizer. The leftmost marker in the plot corresponds to the adapted Quaero system, the next two "+"-markers correspond to the Quaero first-pass system at different master-beam configurations and the "x"-markers correspond to the TUB system at different master-beam configurations.

The general observation is that a better WER results in a better ATWV score and e.g. the Quaero system with reduced master-beam at 26.5% WER still performs better than the TUB system. However, while the relative increase in WER from the Quaero first-pass system to the one with reduced master-beam is 15% and the increase to the TUB system is 32%, the ATWV values experience a relative drop of 12% and 16%, respectively. Hence, the reduce in search-depth of the Quaero system hurts the ATWV more than the WER compared with the TUB system. This could be attributed to the fact, that especially unlikely word hypotheses that do not affect the WER are not contained in the confusion networks.

Since the STD system is based on the confusion networks whereas the WER only reflects the quality of the 1-best transcript, Figure 7.2 certainly has shortcomings. Nevertheless,

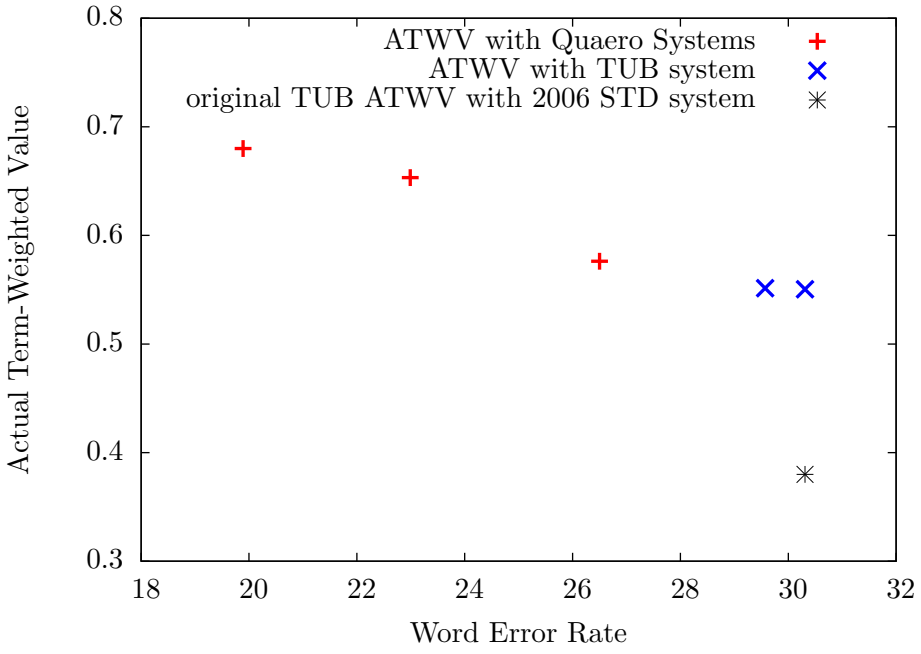


Figure 7.2.: Broadcast News performance of STD system with phone-based expansion on different ASR hypotheses with different Word Error Rates

it gives a good impression of the influence of transcription quality on the ASR value and shows that word error rate on a single-best transcript can be used as an indicator even though confusion networks are searched.

For comparison, the ATWV of the original TUB STD system is also plotted in Figure 7.2. Core differences between the different STD systems operating at identical WER are a more robust term retrieval, query normalizations/expansion and the use of term-specific thresholds.

7.2.3. Influence of Thresholds

The type of thresholds used for the binary classification of candidate occurrences as described in 6.2 makes a substantial difference. Figure 7.3 gives an overview of the ATWV performance of different global threshold values compared with the performance of term-specific thresholds. Whereas 0.4424 is the highest ATWV with global thresholds (at a threshold of 0.571), term-specific thresholds yield an ATWV of 0.5502.

The large "plateau" in the center of the curve indicates that most detection candidates have either a relatively high or a relatively low score and therefore varying thresholds between 0.3 and 0.6 make few differences.

7.2.4. Results on Out-Of-Vocabulary terms

The different query expansion approaches described in chapter 5 enable the detection of out-of-vocabulary words. As seen in the Broadcast News results in Figure 7.4, the use of the different query expansion techniques do not affect the overall ATWV very much, whereas they make a significant difference on the actual OOV words. All experiments in this subsection are based on TUB ASR hypotheses with the DryRun term list.

First, basic term-normalization as described in 5.1 already gives the ability to detect a substantial amount of term occurrences. The ATWV value of 0.2712 is largely based on

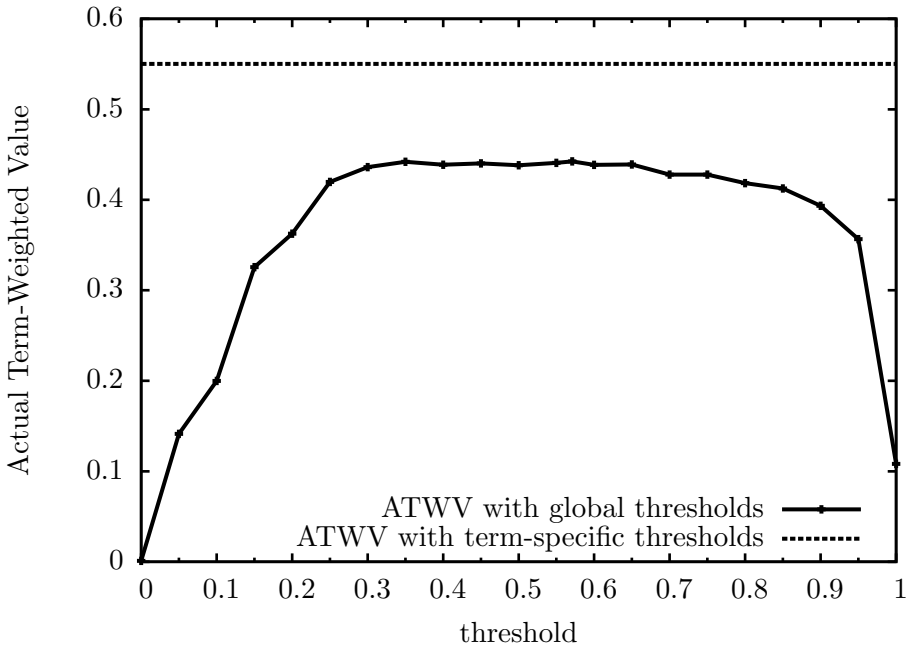


Figure 7.3.: Different Thresholds on TUB Broadcast News Data

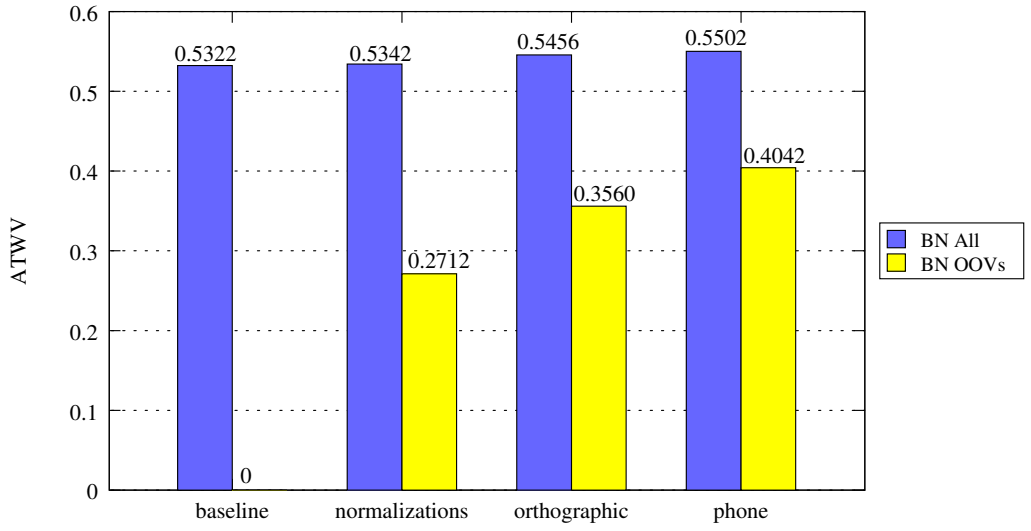


Figure 7.4.: ATWV results on OOVs of DryRun term list with different query expansion strategies without rescoring on TUB hypotheses

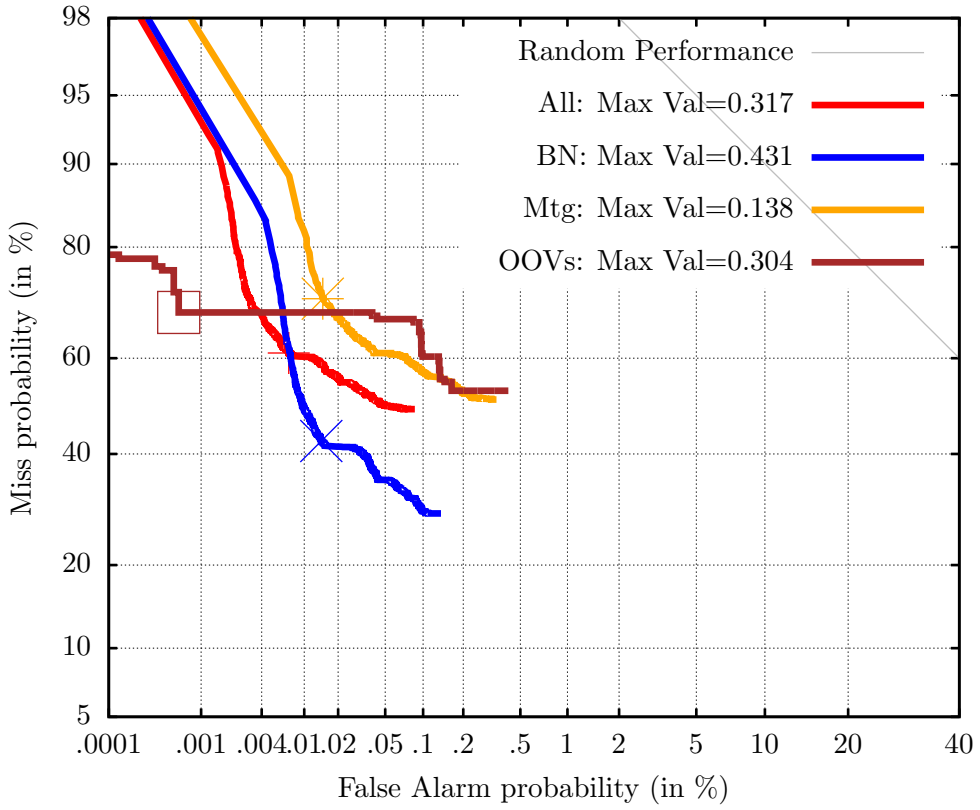


Figure 7.5.: Detection-Error-Tradeoff curves for OOVs compared with all terms; phone-based expansion on TUB hypotheses

the detection of expanded hyphenated terms (e.g. "brother-in-law") or normalization of apostrophes (e.g. "form's").

Using orthographic expansion as a way to generate "similar" words for OOVs, the ATWV value of 0.3560 shows the ability to improve the results without any language-specific knowledge by using string similarity metrics. For this experiment, the Levenshtein distance (5.2.1) has been used. The Dice coefficient performs slightly worse on the DryRun term list, but generally very similar.

The phone-based expansion scheme to generate similar words based on the presumed pronunciation of an OOV query yields further improvements. Using Festival pronunciation rules as described in 5.3 improves the ATWV on DryRun OOVs to 0.4042. This makes it the best performing system on OOV queries.

The typical relation of precision and recall of the OOV results can be seen in Figure 7.5. Although this DET curve is based on the use of global thresholds, it gives some insight into the structure of OOV query results. Whereas the curves for all terms look relatively similar with MTWVs at a false alarm probability of about 0.01, the OOV curve has more of a "ladder" form with the MTWV at a false alarm probability of less than 0.001.

The two step-like decreases in miss probability can be interpreted as the recognition of different types of OOV queries. Whereas normalizable OOV words are detected at a relatively low false alarm probability, several other typical OOV terms can only be detected at a much higher false alarm probability. Considering the relatively low number of OOV terms, the long "horizontal stretch" of the OOV curve is probably rather an effect of the term data than of general OOV considerations.

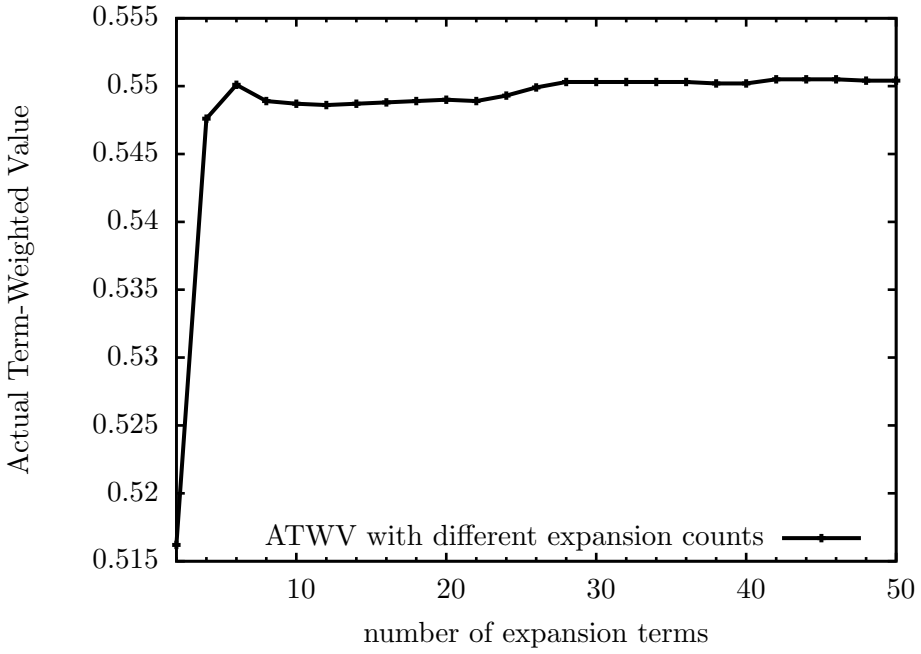


Figure 7.6.: Effect of query expansion item count on DryRun ATWV score with phone-based expansion

The previously reported experiments use 50 query expansion terms generated based on orthographic or acoustic similarity to dictionary words. Figure 7.6 shows the influence of the number of expansion terms on the ATWV value based on the phone-based expansion technique. Naturally, most improvements over the original ATWV value come from the first, i.e. most similar, query expansion words. However, the score does not deteriorate with a higher item count. This can be attributed both to the decreasing similarity factors that reduce the probability of a "YES" decision for an occurrence and to the use of term-specific thresholds which take into account the number of detection candidates by estimating N_{true} (cf. subsection 6.2.2).

7.2.5. Influence of Acoustic Similarity-Based Rescoring

Using the rescoring approach discussed in 6.1 with term-specific thresholds does not yield major improvements as shown in Table 7.4. On the TUB system with DryRun term list and phone-based query expansion, the ATWV on OOV terms rises slightly from 0.3838 to 0.3857.

Different weighing parameters or normalization strategies as mentioned in 6.1.3 do not improve the ATWV either and the rescoring decreases the ATWV if orthographic query expansion is used. For more details, see A.1.2 and A.1.3.

Expansion type and Rescoring	**OOVs ATWV**	**OOVs MTWV**
phone-based without Rescoring	0.3838	0.3044
phone-based with Rescoring	**0.3857**	0.3883
phone-based with Rescoring & Self-Sim	**0.3857**	**0.3890**
orthographic without Rescoring	0.3706	0.2667
orthographic with Rescoring	0.3270	0.3874

Table 7.4.: Influence of Rescoring on different system configurations on OOVs with TUB hypotheses

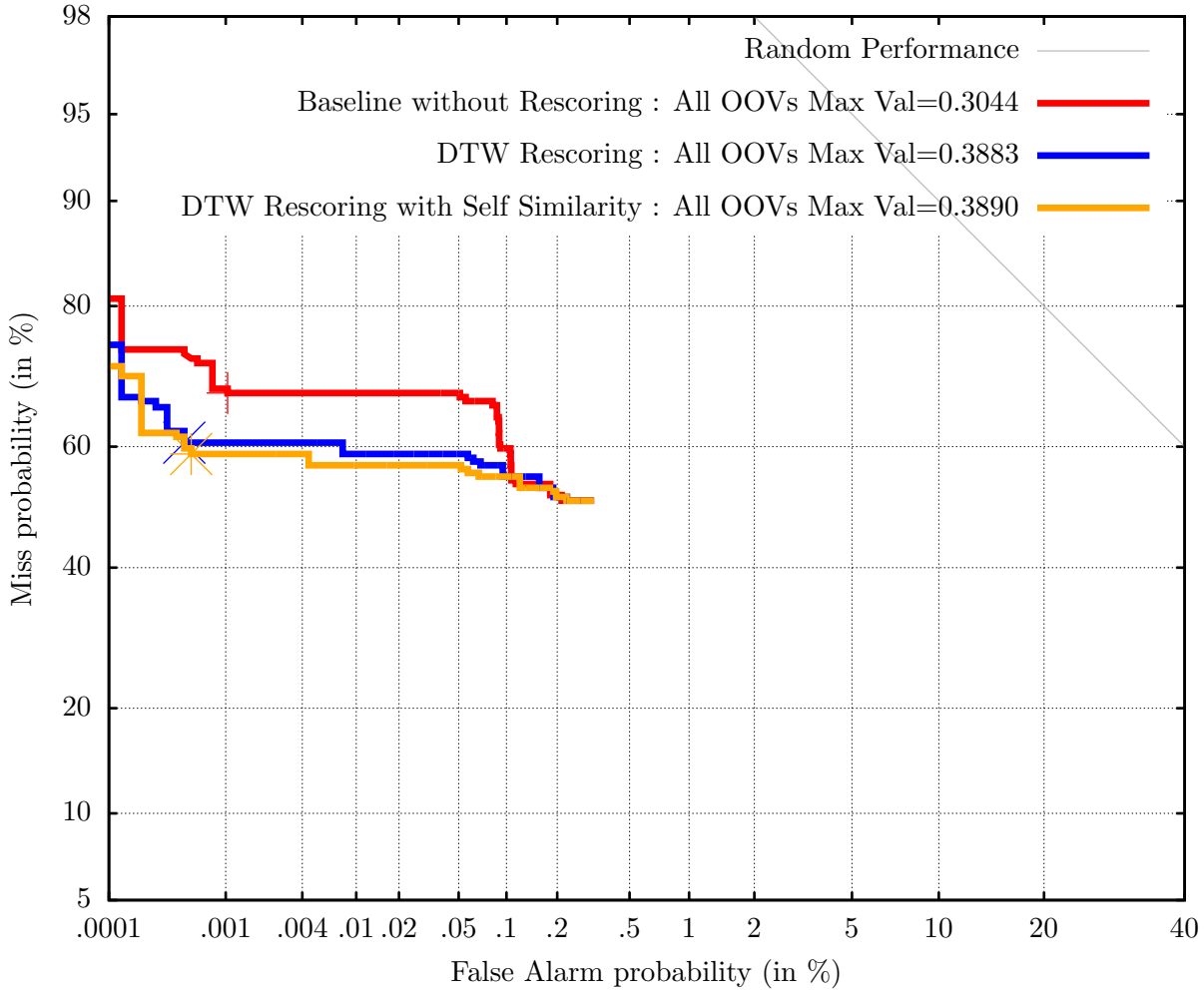


Figure 7.7.: Detection-Error-Tradeoff curves for different rescoring strategies (Global Thresholds) with TUB system and phone-based query expansion on OOV words

System	Broadcast News		Meeting		All OOVs	
	$ATWV$	$ATWV_o$	$ATWV$	$ATWV_o$	$ATWV$	$ATWV_o$
no expansion	0.5322	0.6888	0.2350	0.4930	0.0000	0.0000
normalization	0.5342	0.6891	0.2278	0.4673	0.2507	0.2546
orthographic expansion	0.5456	0.7143	**0.2364**	**0.4923**	0.3706	0.4321
phone-based expansion	**0.5502**	**0.7152**	0.2327	0.4862	**0.3838**	**0.4676**
QuaeroAdapt (phones)	0.6722	0.7890	0.2119	0.4444	0.3463	0.3534

Table 7.5.: Maximal DryRun oracle scores $ATWV_o$ achievable by perfect decisions regardless of detection score. Candidates are based on TUB hypotheses for the first four system

If, however, global thresholds are used instead, improvements can be seen both in the overall DET curve and in the Maximum Term-Weighted Value (see Figure 7.7 and Table 7.4). While the phone-based expansion approach without rescoring results in a MTWV of 0.3044, using a DTW-based distance measure for similarity calculations (6.1.1.1) yields a MTWV of 0.3883. The use of a self-similarity-based distance measure as described in 6.1.1.2 improves the MTWV further to 0.3890.

When using rescoring, the MTWV of 0.3883 actually tops the ATWV of 0.3857 with term-specific thresholds without rescoring. As depicted in the DET plot in 7.7, the rescoring improves recall at fixed precisions compared with the baseline. However, since the underlying detection candidates are identical, the best recall cannot be improved (curves are almost identical at high false alarm probabilities).

Additional experiments have shown that the parameters values of $\alpha = 0.9$ and $\delta = 1$ work well in praxis for the rescoring described in 6.1.3. Rescoring has only been used for terms with at least one OOV word. If used for all words, it deteriorates the results.

7.2.6. Influence and Potential of System Decision

Ignoring confidence scores and decision strategies, it is interesting to evaluate the maximally possible results based on the ASR hypotheses. Table 7.5 shows the corresponding $ATWV_O$ results using oracle decisions from the references for each candidate.

Usually, the oracle scores are substantially higher than the actual ones, with the performance on OOVs using normalizations or the Quaero system being exceptions. Whereas optimal decisions improve the ATWV on Broadcast News data by about 30%, the value more than double on meeting data. This is likely a symptom of the less accurate ASR confidences on the meeting data. Contrarily, the Quaero system seems to perform better with actual decisions compared with the maximum ATWV. This can similarly be attributed to an improved confidence estimation in the hypotheses.

Whereas the calculated scores combined with term-specific thresholds seem to work well on OOVs for normalizations and in the phone-based expansions of the Quaero system, there are substantial shortfalls in the actual decision with orthographic and phone-based expansion. Although this suggest improvable similarity estimations for query expansion, the shortfalls are still smaller than on all terms.

7.3. Evaluation

Based on the results reported in the previous section, different conclusions can be drawn and the influence of different techniques on the STD task can be evaluated.

Generally, the word error rate and improvements in the underlying ASR technology result in large differences in STD performance. Hence, using the hypotheses from the Quaero system instead of the TUB system on the BN data increases the ATWV score by over 20 percent. Nevertheless, also the generalized infrastructure and system design with better detection performance contribute to the score improvement over the existing 2006 system.

At large, the biggest single improvement is the use of term-specific thresholds as a decision strategy. They have both strong analytical motivation and offer a substantial increase in ATWV score. Furthermore, an advantage of TSTs is the nonnecessity of threshold tuning on held-out data.

It is generally helpful to avoid out-of-vocabulary terms resulting from a major mismatch between the vocabulary or word structure of the ASR system and the task evaluation term lists or tools, e.g. the handling of acronyms with the Quaero system. Although normalizations are possible, they typically still result in score deterioration and require additional (ASR configuration-specific) adaptation of the STD system.

7.3.1. OOV Strategies

Considering the overall result, the handling of OOV terms makes only a small difference in the ATWV, with a deterioration of less than 5 percent if OOVs aren't treated at all. Nevertheless, this is a characteristic of the existing term lists and there is strong motivation for OOV handling from the practical application of a STD system.

Normalizable OOV terms can be seen as "low hanging fruits" in the OOV detection task. With some basic adjustments to query terms such as hyphenation normalization, the ATWV improves significantly. However, none of these normalizable terms would be judged as "real" OOV terms by a human whereas the two main query expansion approaches offer detections of the latter kind.

The phone-based query expansion approach performs better than the orthographic one. This can be mainly attributed to the lower false alarm rate of the first technique. Considering the rather poor grapheme-phoneme relation in the English language, the orthographic expansion performs fairly well.

Although the false alarm rate of the expansion techniques is relatively high, this does not represent a major obstacle for their application. In the STD task, it is generally beneficial to find occurrences instead of decreasing false positives. For example, finding one out of two term occurrences with two false alerts still results in a positive ATWV. Hence, the ATWV metric encourages extensive query expansion.

7.3.2. Rescoring

Despite promising results in [CCL+11], the rescoring based on acoustic similarity does not yield significant improvements over the best ATWV. The approach is primarily a re-ranking which is quite helpful when using rank-based evaluation metrics such as Mean Average Precision as in [CCL+11] whereas it makes less of a difference in the binary decision task of STD.

As seen in Figure 7.7, the rescoring makes more of a difference in the detection-error trade-off when comparing the maximum term-weighted values with global thresholds. Over all OOV queries, correct and incorrect detections become better separable, e.g. showing an

MTWV increase from 0.3044 to 0.3883 due to rescoring. Although this shows that rescoring might conceptually help in the task, it is primarily of diagnostic interest. The MTWVs are still close to the ATWVs (0.3838 in the previous example) and can only be realized when the perfect threshold is known which is unrealistic for actual tasks.

A major drawback of general rescoring approaches is the debatable meaning of the scores after application. Whereas the original confidence values can be seen as the probability of a correct detection, this typically cannot be said after rescoring. Since the calculation of term-specific thresholds in 6.2.2 assumes such a probability, rescoring should be used rather reluctantly in the STD task.

Considering that the rescoring uses direct acoustic information whereas the rest of the system is essentially text-based, the additional amount of information and query complexity is not rewarded by an appropriately higher system result. Consequently, the technique as described in 6.1 is not very beneficial and possible improvements are negligible compared with other adjustments.

7.3.3. Comparison to Published Results

Although the evaluation term lists and references for the 2006 NIST task are not available and hence a direct comparison with the evaluation results is not possible, the term structure and audio difficulty was very similar to the development data and competitors reported similar scores.

Looking at the ATWV scores in Table 3.1, both the IBM and the SRI system achieve better results on broadcast news. However, the word error rate of the IBM and (presumably) the SRI system are much lower than the one of the Quaero system. Taking into account the influence of word error rate on ATWV result as depicted in Figure 7.2, the performance of the STD system itself can be seen as comparable although the overall result is not on same level as the competitors'.

The results on meeting data are generally worse than the ones on broadcast news, both with the implemented system and the competitors in the 2006 evaluation which achieve a lower word error rate. However, SRI's evaluation ATWV of 0.26 corresponds to an ATWV of 0.24 of the implemented system on the DryRun term list. On the development data using term-specific thresholds, SRI reports an ATWV of about 0.45 with a WER of 37% whereas the implemented system on TUB data achieves 0.32 with a WER of 60%. Although the ATWV scores differ significantly, the large difference in word error rate can be seen as the major influence factor.

Aside from the worse recognizer hypotheses, the system performance seems to be comparable to the results in literature.

7.4. Summary

This chapter presented the experimental evaluation of the implemented Spoken Term Detection system. The best results on the 2006 development term list are an ATWV of 0.7491 on Broadcast News and an ATWV of 0.3175 on Meeting data. The use of term-specific thresholds contributes most to the large improvement over a previously existing system (from 0.38 to 0.55 on same data).

The accuracy of the underlying ASR system is crucial for the STD task. Using a current Quaero ASR system instead of the 2006 TUB system improves the Broadcast News ATWV values by more than 20 percent.

Using orthographic or phone-based expansion, the detection of out-of-vocabulary words is possible and the corresponding ATWV has been increased to 0.3560 and 0.4042, respectively.

Using rescoring based on acoustic similarity does not yield substantial improvements. Although gains can be seen when using global thresholds, the rescoring does not significantly benefit the overall result when using term-specific thresholds.

8. Conclusion

In this work, a system for the Spoken Term Detection task has been implemented and evaluated with prevalent data and evaluation metrics from the NIST 2006 English Spoken Term Detection task.

Based on confusion network output of the Janus ASR system, the hypothesized textual content of the audio is indexed and stored in a database including metadata such as recognizer confidences. Given a term list, this index is searched and all occurrences are reported along with a binary decision whether the system expects this detection to be correct.

Using this common implementation framework, different techniques for search and retrieval have been evaluated. Generally, the quality of the underlying transcription is crucial for the STD task. The requirement of reporting the individual detection timestamps and disregard of result ranking puts more emphasis on the fine-grained transcription quality than other tasks like Spoken Document retrieval which focus more on information retrieval.

At large, the results using the primary evaluation metric, the ATWV, have been improved to 0.5503 compared with 0.38 of an existing STD system from 2006. The best overall results of 0.7491 is comparable to other published scores and deficits can be primarily attributed to the underlying ASR accuracy.

The detection of terms containing out-of-vocabulary words has been a focus of this work. Whereas none of these query words can be detected in a baseline system, two different approaches for detection based on query expansion have been proposed, implemented and evaluated.

Using orthographically similar words accompanied by some query normalizations yields an ATWV of 0.3560 on out-of-vocabulary terms. This largely language-agnostic approach yields an acceptable ATWV and finds about 40 percent of term occurrences that would otherwise not have been detected.

The phone-based query expansion approach, using a text-to-speech engine to predict OOV pronunciations and calculating acoustically similar words, finds slightly more OOV occurrences, but primarily results in a better precision, yielding an ATWV of 0.4042 on the DryRun OOV words.

These two implemented approaches enables a fairly good detection of OOV terms. Although the influence on the overall results is rather small on the 2006 material, this ability is highly desirable, especially considering practical detection applications.

Given a set of detection candidates, two methods of enhancing the decision process and ATWV results have been evaluated. Term-specific thresholds as an optimized decision strategy for the ATWV metric vary the detection threshold based on the probabilities of detected candidates. They largely avoid the necessity to optimize the thresholds and offer major gains in the overall results.

Using a rescoring approach based on acoustic similarity and pseudo-relevance feedback, the decision process takes into account both the recognizer confidence for a candidate and the acoustic similarity to other (especially probable) candidates. Although this approach improves the global separability of correct and incorrect candidates based on score, it does not work very well with term-specific thresholds and has only a small positive impact on the best ATWV results.

Aside from the phone-based expansion and some rather language-specific normalizations, the STD system does not rely on specifics of the English language. Assuming a STT system able to produce confusion network output, the STD system should be easily usable with different languages.

Overall, the resulting system represents a thorough implementation for the Spoken Term Detection task. It offers an extensible framework for the detection of both in-vocabulary and out-of-vocabulary terms that reaches competitive evaluation scores.

This work has been focused on the detection component of the STD task, i.e. the underlying ASR systems have only been varied for diagnostic purposes and not been optimized for the STD task and evaluation metrics. Consequently, such an optimization would be a natural component of future work. Starting from adjustments to the direct recognizer output such as varying the probabilities of word or filler detections, an abundance of adaptations to the task on various ASR stages could be envisioned. Additionally, the potential benefit of combining the output of various recognizers could be evaluated.

Albeit the text-based evaluation metric and discouraging results in the rescoring part of this work and in literature, it might be worthwhile to use additional information and features aside from the transcription to enhance the detection performance or to make results more robust.

Whereas all results in this thesis have been produced on English data to allow comparison with other works, it would be helpful to evaluate performance on other languages, especially ones with fewer resources, less optimized speech recognizers or non-existing letter-to-sound rules. Besides the general sensitivity to ASR accuracy, varying languages with varying grapheme-phoneme relations will pose different challenges to the query expansion strategies used to recognize out-of-vocabulary words.

Bibliography

[AVS08] M. Akbacak, D. Vergyri, and A. Stolcke, "Open-vocabulary spoken term detection using graphone-based hybrid recognition systems," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 5240–5243.

[BN05] M. Bisani and H. Ney, "Open vocabulary speech recognition with flat hybrid models," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[CA05] C. Chelba and A. Acero, "Position specific posterior lattices for indexing speech," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ser. ACL '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 443–450.

[CCL+11] Y.-N. Chen, C.-P. Chen, H.-Y. Lee, C.-A. Chan, and L.-S. Lee, "Improved spoken term detection with graph-based re-ranking in feature space," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 5644–5647.

[Che03] S. Chen, "Conditional and joint models for grapheme-to-phoneme conversion," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[CHS08] C. Chelba, T. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *Signal Processing Magazine, IEEE*, vol. 25, no. 3, pp. 39–49, May 2008.

[CL10] C. Chan and L. Lee, "Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[CLYL10] C. Chen, H. Lee, C. Yeh, and L. Lee, "Improved spoken term detection by feature space pseudo-relevance feedback," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[CP07] U. Chaudhari and M. Picheny, "Improvements in phone based audio search via constrained match with high order confusion estimates," in *Automatic Speech Recognition Understanding, 2007. ASRU. IEEE Workshop on*, December 2007, pp. 665–670.

[CR76] R. Christiansen and C. Rushforth, "Word spotting in continuous speech using linear predictive coding," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '76.*, vol. 1, April 1976, pp. 557–560.

[CS09] D. Can and M. Saraclar, "Score distribution based term specific thresholding for spoken term detection," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the*

Association for Computational Linguistics, Companion Volume: Short Papers. Boulder, Colorado: Association for Computational Linguistics, June 2009, pp. 269–272.

[CSA07] 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, 2007.

[CWSC04] A. Cutler, A. Weber, R. Smits, and N. Cooper, "Patterns of english phoneme confusions by native and non-native listeners," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3668–3678, 2004.

[CYF+97] J. G. Carbonell, Y. Yang, R. E. Frederking, R. D. Brown, Y. Geng, and D. Lee, "Translingual information retrieval: A comparative evaluation," in *International Joint Conference on Artificial Intelligence*, 1997, pp. 708–715.

[DT05] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, june 2005, pp. 886–893.

[FAGD07] J. Fiscus, J. Ajot, J. Garofolo, and G. Doddingtion, "Results of the 2006 spoken term detection evaluation," in *Proceedings of the ACM SIGIR Workshop "Searching Spntaneous Conversational Speech"*, 2007, p. 51.

[GAV00] J. Garofolo, C. Auzanne, and E. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proc. Recherche d'Informations Assiste par Ordinateur: Content Based Multimedia Information Access Conf.*, 2000.

[GSO+02] S. Gustman, D. Soergel, D. Oard, W. Byrne, M. Picheny, B. Ramabhadran, and D. Greenberg, "Supporting access to large digital oral history archives," in *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, ser. JCDL '02. New York, NY, USA: ACM, 2002, pp. 18–27.

[HKC07] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking through random walk over document-level context graph," in *Proceedings of the 15th international conference on Multimedia*, ser. MULTIMEDIA '07. New York, NY, USA: ACM, 2007, pp. 971–980.

[KSS03] M. Killer, S. Stüker, and T. Schultz, "Grapheme based speech recognition," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[LM05] A. N. Langville and C. D. Meyer, "A survey of eigenvector methods for web information retrieval," *SIAM Rev.*, vol. 47, pp. 135–161, January 2005.

[LMTW00] B. Logan, P. Moreno, J. Thong, and E. Whittaker, "An experimental study of an audio indexing system for the web," in *Sixth International Conference on Spoken Language Processing*, 2000.

[LVTM05] B. Logan, J.-M. Van Thong, and P. Moreno, "Approaches to reduce the effects of oov queries on indexed spoken audio," *Multimedia, IEEE Transactions on*, vol. 7, no. 5, pp. 899–906, October 2005.

[MBS00] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 373–400, 2000.

[MCH06] J. Mamou, D. Carmel, and R. Hoory, "Spoken document retrieval from call-center conversations," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '06. New York, NY, USA: ACM, 2006, pp. 51–58.

[MFPW05] F. Metze, C. Fugen, Y. Pan, and A. Waibel, "Automatically transcribing meetings using distant microphones," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, vol. 1, March 2005, pp. 989–992.

[MGB11] A. Muscariello, G. Gravier, and F. Bimbot, "Towards robust word discovery by self-similarity matrix comparison," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 5640–5643.

[MKK+07] D. Miller, M. Kleber, C. Kao, O. Kimball, T. Colthurst, S. Lowe, R. Schwartz, and H. Gish, "Rapid and accurate spoken term detection," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[MKL+00] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava, "Speech and language technologies for audio indexing and retrieval," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1338–1353, August 2000.

[MRS07] J. Mamou, B. Ramabhadran, and O. Siohan, "Vocabulary independent spoken term detection," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '07. New York, NY, USA: ACM, 2007, pp. 615–622.

[Nat06] *The 2006 Spoken Term Detection Evaluation Plan*, 10th ed., National Institute of Standards and Technology, September 2006, last accessed on September 14, 2011. [Online]. Available: http://www.itl.nist.gov/iad/mig//tests/std/2006/docs/std06-evalplan-v10.pdf

[NZ00] K. Ng and V. W. Zue, "Subword-based approaches for spoken document retrieval," *Speech Communication*, vol. 32, no. 3, pp. 157–186, 2000.

[PSR09] C. Parada, A. Sethy, and B. Ramabhadran, "Query-by-example spoken term detection for oov terms," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on.* IEEE, 2009, pp. 404–409.

[RSM+09] B. Ramabhadran, A. Sethy, J. Mamou, B. Kingsbury, and U. Chaudhari, "Fast decoding for open vocabulary spoken term detection," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, ser. NAACL-Short '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 277–280.

[SB90] G. Salton and C. Buckley, "Improving retrieval performance by relevance feedback," *Journal of the American Society for Information Science*, vol. 41, no. 4, pp. 288–297, 1990.

[SC78] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, pp. 43–49, 1978.

[SKN11] S. Stüker, K. Kilgour, and J. Niehues, "Quaero speech-to-text and text translation evaluation systems," in *High Performance Computing in Science and Engineering '10*, W. E. Nagel, D. B. Kröner, and M. M. Resch, Eds. Springer Berlin Heidelberg, 2011, pp. 529–542.

[SMFW01] H. Soltau, F. Metze, C. Fugen, and A. Waibel, "A one-pass decoder based on polymorphic linguistic context assignment," in *Automatic Speech Recognition and Understanding, 2001. ASRU '01. IEEE Workshop on*, 2001, pp. 214–217.

[SRM06] O. Siohan, B. Ramabhadran, and J. Mamou, "The IBM 2006 spoken term detection system," NIST Spoken Term Detection Workshop, December 2006, last accessed on September 14, 2011. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/std/2006/pubdata/pres/IBM-STD-system.ppt

[TBC98] P. Taylor, A. Black, and R. Caley, "The architecture of the festival speech synthesis system," in *The Third ESCA Workshop in Speech Synthesis*, 1998, pp. 147–151.

[VSGW06] D. Vergyri, A. Stolcke, R. R. Gadde, and W. Wang, "The SRI 2006 spoken term detection system," NIST Spoken Term Detection Workshop, December 2006, last accessed on September 14, 2011. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/std/2006/pubdata/pres/SRI-std06.ppt

[VSS+07] D. Vergyri, I. Shafran, A. Stolcke, R. Gadde, M. Akbacak, B. Roark, and W. Wang, "The SRI/OGI 2006 spoken term detection system," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[WFTK08] D. Wang, J. Frankel, J. Tejedor, and S. King, "A comparison of phone and grapheme-based spoken term detection," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, April 2008, pp. 4969–4972.

[WKF09] D. Wang, S. King, and J. Frankel, "Stochastic pronunciation modelling for spoken term detection," in *Tenth Annual Conference of the International Speech Communication Association*, 2009, pp. 2135–2138.

[WKFB09] D. Wang, S. King, J. Frankel, and P. Bell, "Term-dependent confidence for out-of-vocabulary term detection," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[You96] S. Young, "A review of large-vocabulary continuous-speech," *Signal Processing Magazine, IEEE*, vol. 13, no. 5, p. 45, September 1996.

Nomenclature

ASR	Automatic Speech Recognition
ATWV	Actual Term-Weighted Value
BN	Broadcast News
CONFMTG	Conference Meetings
CTS	Conversational Telephone Speech
DB	Database
DET	Detection-Error-Tradeoff
DTW	Dynamic Time Warping
FA	False Alert
GT	Global Threshold
HOG	Histogram of Oriented Gradients
HTTP	Hypertext Transfer Protocol
IV	In-vocabulary [word]
MAP	Mean Average Precision
MFCC	Mel-Frequency Cepstral Coefficients
MTWV	Maximum Term-Weighted Value
NIST	National Institute for Standards and Technology
OOV	Out-of-vocabulary [word]
PRF	Pseudo-Relevance Feedback
PSPL	Position Specific Posterior Lattice
SDR	Spoken Document Retrieval
SSM	Self-Similarity Matrix
STD	Spoken Term Detection

STT	Speech-to-text [system]
TREC	Text REtrieval Conference
TST	Term-Specific Threshold
WER	Word Error Rate
XML	eXtensible Markup Language

Appendix

This appendix contains detailed results of the experiments performed. Each table with with detection statistics is preceded by a short definition list of the systems described. The "System ID" corresponds to the name used for archiving the detection files.

The BN and MTG statistics correspond to the results on broadcast news and meeting data for all terms. The OOV statistics correspond to the result on out-of-vocabulary terms from both audio types. Note, that the set of OOV terms differs between the TUB and the Quaero STT system.

For each term subset, ATWV (using term-specific thresholds), MTWV (assuming an optimal global threshold) along with correct detections, false alerts (FAs) and missed occurrences of the actual system decision (i.e. the basis of ATWV calculation) are reported.

A. Results on DryRun term list

A.1. TUB STT system

A.1.1. Basic Configurations

couch3_noNorms
: Basic STD system without any normalizations or query expansions or rescoring

couch3_noSubs_ignoreAp
: system with normalizations including ignoring apostrophes but without GLM substitutions or query expansions or rescoring

System ID	couch3_noNorms	couch3_noSubs_ignoreAp
BN ATWV	0.5322	0.5342
BN MTWV	0.4305	0.4319
BN Correct	3954	3939
BN FAs	1158	1175
BN Misses	1713	1728
MTG ATWV	0.2350	0.2278
MTG MTWV	0.1558	0.1337
MTG Correct	707	661
MTG FAs	335	310
MTG Misses	3289	3335
OOV ATWV	0.0000	0.2507
OOV MTWV	0.0000	0.2520
OOV Correct	0	22
OOV FAs	0	6
OOV Misses	0	60

A.1.2. Orthographic Expansion

fuzzy3_noSubs_ignoreAp
system using orthographic expansion (Levenshtein) with normalizations including ignoring apostrophes but without GLM substitutions or rescoring

fuzzy3_dtw_noSubs_ignoreAp
system using orthographic expansion and DTW rescoring with normalizations including ignoring apostrophes but without GLM substitutions

System ID	fuzzy3_noSubs_ignoreAp	fuzzy3_dtw_noSubs_ignoreAp
BN ATWV	0.5456	0.5442
BN MTWV	0.4235	0.4157
BN Correct	3958	3950
BN FAs	1249	1218
BN Misses	1709	1717
MTG ATWV	0.2364	0.2368
MTG MTWV	0.1337	0.1337
MTG Correct	665	665
MTG FAs	324	311
MTG Misses	3331	3331
OOV ATWV	0.3706	0.3270
OOV MTWV	0.2667	0.3874
OOV Correct	33	25
OOV FAs	73	29
OOV Misses	49	57

A.1.3. Phone-based Expansion

phones3_noSubs_ignoreAps
system with phone-based query expansion

phones3_dtw_noSubs_ignoreAps
system with phone-based query expansion and rescoring based on Dynamic Time Warping (rescoring and normalization for OOV queries only)

phones3_dtw_selfSim_noSubs_ignoreAps
system with phone-based query expansion and rescoring based on Self Similarity Matrices (rescoring and normalization for OOV queries only)

phones3_dtw_probNorm_noSubs_ignoreAps
system with phone-based query expansion and rescoring based on Dynamic Time Warping and subsequent confidence normalization (rescoring and normalization for OOV queries only)

System ID	phones3 noSubs ignoreAps	phones3 dtw noSubs ignoreAps	phones3 dtw selfSim noSubs ignoreAps	phones3 dtw probNorm noSubs ignoreAps
BN ATWV	0.5502	0.5503	0.5503	0.5489
BN MTWV	0.4313	0.4281	0.4287	0.4291
BN Correct	3964	3964	3964	3960
BN FAs	1195	1194	1194	1193
BN Misses	1703	1703	1703	1709
MTG ATWV	0.2327	0.2331	0.2331	0.2331
MTG MTWV	0.1383	0.1383	0.1383	0.1383
MTG Correct	663	663	663	663
MTG FAs	312	310	310	310
MTG Misses	3333	3333	3333	3331
OOV ATWV	0.3838	0.3857	0.3857	0.3608
OOV MTWV	0.3044	0.3883	0.3890	0.2810
OOV Correct	34	34	34	30
OOV FAs	15	12	12	11
OOV Misses	48	48	48	52

A.2. Quaero STT system

quaero_phones3_noPeriodExpansion
system based on Quaero hypotheses and phone-based query expansion

quaero_fuzzy3_noPeriodExpansion
system based on Quaero hypotheses and orthographic query expansion

quaeroAdapt_phones3_noSubs_noAcronymExpansion
system based on hypotheses of the adapted Quaero system using phone-based query expansion

quaeroAdapt_fuzzy3_levenshtein_noAcronymExpansion
system based on hypotheses of the adapted Quaero system using orthographic expansion

System ID	quaero phones3 noPeriodExpansion	quaero fuzzy3 noPeriodExpansion	quaeroAdapt phones3 noSubs noAcronymExpansion	quaeroAdapt fuzzy3 levenshtein noAcronymExpansion
BN ATWV	0.6548	0.6612	0.6722	0.6766
BN MTWV	0.5788	0.5697	0.6129	0.5885
BN Correct	3787	3803	3896	3910
BN FAs	876	881	825	870
BN Misses	1880	1864	1771	1757
MTG ATWV	0.2135	0.2138	0.2119	0.2102
MTG MTWV	0.2135	0.1621	0.1379	0.1375
MTG Correct	610	618	684	685
MTG FAs	277	287	371	386
MTG Misses	3386	3378	3312	3311
OOV ATWV	0.1223	0.3760	0.3463	0.3643
OOV MTWV	0.1053	0.3206	0.3382	0.3388
OOV Correct	21	34	31	32
OOV FAs	31	37	13	39
OOV Misses	104	48	51	50

Comment The number of OOV words for the Quaero systems is unusually high due to the mismatch in acronym handling (cf. 7.1.2. Although the expansion of acronyms into single letters is possible and greatly increases the number of detected OOV words, the number of false alerts also increases significantly, making the ATWV score worse than the one without acronym expansion.

B. Results on Development term list

Results in this section are based on the development term list. Details can be found in 7.1.1.

B.1. TUB STT system

The following two systems use TUB hypotheses

dev_phones3_noSubs_ignoreAp
system using phone-based expansion

dev_fuzzy3_levenshtein_noSubs_ignoreAp
system using orthographic expansion with Levenshtein distance

System ID	dev phones3 noSubs ignoreAp	dev fuzzy3 levenshtein noSubs ignoreAp
BN ATWV	0.6097	0.5951
BN MTWV	0.5189	0.5027
BN Correct	3502	3499
BN FAs	882	990
BN Misses	1386	1389
MTG ATWV	0.3175	0.3175
MTG MTWV	0.1746	0.1746
MTG Correct	530	530
MTG FAs	179	188
MTG Misses	2640	2640
OOV ATWV	0.1604	-0.0033
OOV MTWV	0.0222	0.0222
OOV Correct	8	3
OOV FAs	8	99
OOV Misses	49	54

B.2. Quaero STT system

dev_quaero_phones3
system based on Quaero hypotheses using phone-based expansion

dev_quaeroAdapt_phones3_noAcronymExpansion
system based on hypotheses from the adapted Quaero system using phone-based expansion

dev_quaeroAdapt_fuzzy3_levenshtein_noAcronymExpansion
system based on hypotheses from the adapted Quaero system using orthographic expansion with Levenshtein distance

System ID	dev quaero phones3	dev quaeroAdapt phones3 noAcronymExpansion	dev quaeroAdapt fuzzy3 levenshtein noAcronymExpansion
BN ATWV	0.7398	0.7491	0.7425
BN MTWV	0.6667	0.7087	0.6513
BN Correct	3507	3581	3575
BN FAs	679	642	673
BN Misses	1381	1307	1313
MTG ATWV	0.2776	0.2924	0.2918
MTG MTWV	0.2106	0.2258	0.2258
MTG Correct	473	520	520
MTG FAs	149	203	211
MTG Misses	2697	2650	2650
OOV ATWV	0.1817	0.1404	0.0093
OOV MTWV	0.0658	0.0658	0.0000
OOV Correct	9	8	2
OOV FAs	30	21	60
OOV Misses	448	449	455