# Phrase-Based Query Degradation Modeling for Vocabulary-Independent Ranked Utterance Retrieval

**J. Scott Olsson**
HLT Center of Excellence
Johns Hopkins University
Baltimore, MD 21211, USA
solsson@jhu.edu

**Douglas W. Oard**
College of Information Studies
University of Maryland
College Park, MD 15213, USA
oard@umd.edu

## Abstract

This paper introduces a new approach to ranking speech utterances by a system's confidence that they contain a spoken word. Multiple alternate pronunciations, or degradations, of a query word's phoneme sequence are hypothesized and incorporated into the ranking function. We consider two methods for hypothesizing these degradations, the best of which is constructed using factored phrase-based statistical machine translation. We show that this approach is able to significantly improve upon a state-of-the-art baseline technique in an evaluation on held-out speech. We evaluate our systems using three different methods for indexing the speech utterances (using phoneme, phoneme multigram, and word recognition), and find that degradation modeling shows particular promise for locating out-of-vocabulary words when the underlying indexing system is constructed with standard word-based speech recognition.

## 1 Introduction

Our goal is to find short speech utterances which contain a query word. We accomplish this goal by ranking the set of utterances by our confidence that they contain the query word, a task known as Ranked Utterance Retrieval (RUR). In particular, we are interested in the case when the user's query word can not be anticipated by a Large Vocabulary Continuous Speech Recognizer's (LVCSR) decoding dictionary, so that the word is said to be Out-Of-Vocabulary (OOV).

Rare words tend to be the most informative, but are also most likely to be OOV. When words are OOV, we must use vocabulary-independent techniques to locate them. One popular approach is to search for the words in output from a phoneme recognizer (Ng and Zue, 2000), although this suffers from the low accuracy typical of phoneme recognition. We consider two methods for handling this inaccuracy. First, we compare an RUR indexing system using phonemes with two systems using longer recognition units: words or phoneme multigrams. Second, we consider several methods for handling the recognition inaccuracy in the utterance ranking function itself. Our baseline generative model handles errorful recognition by estimating term frequencies from smoothed language models trained on phoneme lattices. Our new approach, which we call query degradation, hypothesizes many alternate "pronunciations" for the query word and incorporates them into the ranking function. These degradations are *translations* of the lexical phoneme sequence into the errorful recognition language, which we hypothesize using a factored phrase-based statistical machine translation system.

Our speech collection is a set of oral history interviews from the MALACH collection (Byrne et al., 2004), which has previously been used for *ad hoc* speech retrieval evaluations using one-best word level transcripts (Pecina et al., 2007; Olsson, 2008a) and for vocabulary-independent RUR (Olsson, 2008b). The interviews were conducted with survivors and witnesses of the Holocaust, who discuss their experiences before, during, and after the Second World War. Their speech is predominately spontaneous and conversational. It is often also emotional and heavily accented. Because the speech contains many words unlikely to occur within a general purpose speech recognition lexicon, it repre-



*Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, pages 182–190,
Boulder, Colorado, June 2009. 

sents an excellent collection for RUR evaluation.

We were graciously permitted to use BBN Technology's speech recognition system *Byblos* (Prasad et al., 2005; Matsoukas et al., 2005) for our speech recognition experiments. We train on approximately 200 hours of transcribed audio excerpted from about 800 unique speakers in the MALACH collection. To provide a realistic set of OOV query words, we use an LVCSR dictionary previously constructed for a different topic domain (broadcast news and conversational telephone speech) and discard all utterances in our acoustic training data which are not covered by this dictionary. New acoustic and language models are trained for each of the phoneme, multigram and word recognition systems.

The output of LVCSR is a *lattice* of recognition hypotheses for each test speech utterance. A lattice is a directed acyclic graph that is used to compactly represent the search space for a speech recognition system. Each node represents a point in time and arcs between nodes indicates a word occurs between the connected nodes' times. Arcs are weighted by the probability of the word occurring, so that the so-called "one-best" path through the lattice (what a system might return as a transcription) is the path through the lattice having highest probability under the acoustic and language models. Each RUR model we consider is constructed using the expected counts of a query word's phoneme sequences in these recognition lattices. We consider three approaches to producing these phoneme lattices, using standard word-based LVCSR, phoneme recognition, and LVCSR using phoneme multigrams. Our word system's dictionary contains about $50{,}000$ entries, while the phoneme system contains 39 phonemes from the ARPABET set.

Originally proposed by Deligne and Bimbot (1997) to model variable length regularities in streams of symbols (e.g., words, graphemes, or phonemes), phoneme multigrams are short sequences of one or more phonemes. We produce a set of "phoneme transcripts" by replacing transcript words with their lexical pronunciation. The set of multigrams is learned by then choosing a maximum-likelihood segmentation of these training phoneme transcripts, where the segmentation is viewed as hidden data in an Expectation-Maximization algorithm. The set of all continuous phonemes occurring between segment boundaries is then chosen as our multigram dictionary. This multigram recognition dictionary contains $16{,}409$ entries.

After we have obtained each recognition lattice, our indexing approach follows that of Olsson (2008b). Namely, for the word and multigram systems, we first expand lattice arcs containing multiple phones to produce a lattice having only single phonemes on its arcs. Then, we compute the expected count of all phoneme $n$-grams $n \leq 5$ in the lattice. These $n$-grams and their counts are inserted in our inverted index for retrieval.

This paper is organized as follows. In Section 2 we introduce our baseline RUR methods. In Section 3 we introduce our query degradation approach. We introduce our experimental validation in Section 4 and our results in Section 5. We find that using phrase-based query degradations can significantly improve upon a strong RUR baseline. Finally, in Section 6 we conclude and outline several directions for future work.

## 2 Generative Baseline

Each method we present in this paper ranks the utterances by the term's estimated frequency within the corresponding phoneme lattice. This general approach has previously been considered (Yu and Seide, 2005; Saraclar and Sproat, 2004), on the basis that it provides a minimum Bayes-risk ranking criterion (Yu et al., Sept 2005; Robertson, 1977) for the utterances. What differs for each method is the particular estimator of term frequency which is used. We first outline our baseline approach, a generative model for term frequency estimation.

Recall that our vocabulary-independent indices contain the expected counts of phoneme sequences from our recognition lattices. Yu and Seide (2005) used these expected phoneme sequence counts to estimate term frequency in the following way. For a query term $Q$ and lattice $\mathcal{L}$, term frequency $\widehat{tf}_G$ is estimated as $\widehat{tf}_G(Q, \mathcal{L}) = P(Q|\mathcal{L}) \cdot N_{\mathcal{L}}$, where $N_{\mathcal{L}}$ is an estimate for the number of words in the utterance. The conditional $P(Q|\mathcal{L})$ is modeled as an order $M$ phoneme level language model,

$$\hat{P}(Q|\mathcal{L}) = \prod_{i=1}^{l} \tilde{P}(q_i|q_{i-M+1}, \ldots, q_{i-1}, \mathcal{L}), \quad (1)$$

so that $\widehat{tf}_G(Q, \mathcal{L}) \approx \hat{P}(Q|\mathcal{L}) \cdot N_{\mathcal{L}}$. The probability of a query phoneme $q_j$ being generated, given that the phoneme sequence $q_{j-M+1}, \ldots, q_{j-1} = q_{j-M+1}^{j-1}$ was observed, is estimated as

$$\tilde{P}(q_j|q_{j-M+1}^{j-1}, \mathcal{L}) = \frac{E_{P_{\mathcal{L}}}[C(q_{j-M+1}^{j})]}{E_{P_{\mathcal{L}}}[C(q_{j-M+1}^{j-1})]}.$$

Here, $E_{P_{\mathcal{L}}}[C(q_{j-M+1}^{j-1})]$ denotes the expected count in lattice $\mathcal{L}$ of the phoneme sequence $q_{j-M+1}^{j-1}$. We compute these counts using a variant of the forward-backward algorithm, which is implemented by the SRI language modeling toolkit (Stolcke, 2002).

In practice, because of data sparsity, the language model in Equation 1 must be modified to include smoothing for unseen phoneme sequences. We use a backoff $M$-gram model with Witten-Bell discounting (Witten and Bell, 1991). We set the phoneme language model's order to $M = 5$, which gave good results in previous work (Yu and Seide, 2005).

## 3 Incorporating Query Degradations

One problem with the generative approach is that recognition error is not modeled (apart from the uncertainty captured in the phoneme lattice). The essential problem is that while the method hopes to model $P(Q|\mathcal{L})$, it is in fact only able to model the probability of one *degradation* $H$ in the lattice, that is $P(H|\mathcal{L})$. We define a query degradation as any phoneme sequence (including the lexical sequence) which may, with some estimated probability, occur in an errorful phonemic representation of the audio (either a one-best or lattice hypothesis). Because of speaker variation and because recognition is errorful, we ought to also consider non-lexical degradations of the query phoneme sequence. That is, we should incorporate $P(H|Q)$ in our ranking function.

It has previously been demonstrated that allowing for phoneme confusability can significantly increase spoken term detection performance on one-best phoneme transcripts (Chaudhari and Picheny, 2007; Schone et al., 2005) and in phonemic lattices (Foote et al., 1997). These methods work by allowing weighted substitution costs in minimum-edit-distance matching. Previously, these substitution costs have been maximum-likelihood estimates of $P(H|Q)$ for each phoneme, where $P(H|Q)$ is easily computed from a phoneme confusion matrix after aligning the reference and one-best hypothesis transcript under a minimum edit distance criterion. Similar methods have also been used in other language processing applications. For example, in (Kolak, 2005), one-for-one character substitutions, insertions and deletions were considered in a generative model of errors in OCR.

In this work, because we are focused on constructing inverted indices of audio files (for speed and to conserve space), we must generalize our method of incorporating query degradations in the ranking function. Given a degradation model $P(H|Q)$, we take as our ranking function the expectation of the generative baseline estimate $N_{\mathcal{L}} \cdot \hat{P}(H|\mathcal{L})$ with respect to $P(H|Q)$,

$$\widehat{tf}_G(Q, \mathcal{L}) = \sum_{H \in \mathcal{H}} \left[\hat{P}(H|\mathcal{L}) \cdot N_{\mathcal{L}}\right] \cdot P(H|Q), \quad (2)$$

where $\mathcal{H}$ is the set of degradations. Note that, while we consider the expected value of our baseline term frequency estimator with respect to $P(H|Q)$, this general approach could be used with any other term frequency estimator.

Our formulation is similar to approaches taken in OCR document retrieval, using degradations of character sequences (Darwish and Magdy, 2007; Darwish, 2003). For vocabulary-independent spoken term detection, perhaps the most closely related formulation is provided by (Mamou and Ramabhadran, 2008). In that work, they ranked utterances by the weighted average of their matching score, where the weights were confidences from a grapheme to phoneme system's first several hypotheses for a word's pronunciation. The matching scores were edit distances, where substitution costs were weighted using phoneme confusability. Accordingly, their formulation was not aimed at accounting for errors in recognition per se, but rather for errors in hypothesizing pronunciations. We expect this accounts for their lack of significant improvement using the method.

Since we don't want to sum over all possible recognition hypotheses $H$, we might instead sum over the smallest set $\mathcal{H}$ such that $\sum_{H \in \mathcal{H}} P(H|Q) \geq \gamma$. That is, we could take the most probable degradations until their cumulative probability exceeds some threshold $\gamma$. In practice, however, because

degradation probabilities can be poorly scaled, we instead take a fixed number of degradations and normalize their scores. When a query is issued, we apply a degradation model to learn the top few phoneme sequences $\mathcal{H}$ that are most likely to have been recognized, under the model. In the machine translation literature, this process is commonly referred to as *decoding*.

We now turn to the modeling of query degradations $H$ given a phoneme sequence $Q$, $P(H|Q)$. First, we consider a simple baseline approach in Section 3.1. Then, in Section 3.2, we propose a more powerful technique, using state-of-the-art machine translation methods to hypothesize our degradations.

### 3.1 Baseline Query Degradations

Schone et al. (2005) used phoneme confusion matrices created by aligning hypothesized and reference phoneme transcripts to weight edit costs for a minimum-edit distance based search in a one-best phoneme transcript. Foote et al. (1997) had previously used phoneme lattices, although with *ad hoc* edit costs and without efficient indexing. In this work, we do not want to linearly scan each phoneme lattice for our query's phoneme sequence, preferring instead to look up sequences in the inverted indices containing phoneme sequences.

Our baseline degradation approach is related to the edit-cost approach taken by (Schone et al., 2005), although we generalize it so that it may be applied within Equation 2 and we consider speech recognition hypotheses beyond the one-best hypothesis. First, we randomly generate $N$ traversals of each phonemic recognition lattice. These traversals are random paths through the lattice (i.e., we start at the beginning of the lattice and move to the next node, where our choice is weighted by the outgoing arcs' probabilities). Then, we align each of these traversals with its reference transcript using a minimum-edit distance criterion. Phone confusion matrices are then tabulated from the aggregated insertion, substitution, and deletion counts across all traversals of all lattices. From these confusion matrices, we compute unsmoothed estimates of $P(h|r)$, the probability of a phoneme $h$ being hypothesized given a reference phoneme $r$.

Making an independence assumption, our baseline degradation model for a query with $m$ phonemes is then $P(H|Q) = \prod_{i=1}^{m} P(h_i|r_i)$. We efficiently compute the most probable degradations for a query $Q$ using a lattice of possible degradations and the forward backward algorithm. We call this baseline degradation approach CMQD (Confusion Matrix based Query Degradation).

| AY | K | M | AA | N |
|---|---|---|---|---|
| Vowel | Consonant | Semi-vowel | Vowel | Semi-vowel |
| Dipthong | Voiceless plosive | Nasal | Back vowel | Nasal |

Figure 1: Three levels of annotation used by the factored phrase-based query degradation model.

### 3.2 Phrase-Based Query Degradation

One problem with CMQD is that we only allow insertions, deletions, and one-for-one substitutions. It may be, however, that certain pairs of phonemes are commonly hypothesized for a particular reference phoneme (in the language of statistical machine translation, we might say that we should allow some non-zero *fertility*). Second, there is nothing to discourage query degradations which are unlikely under an (errorful) language model—that is, degradations that are not observed in the speech hypotheses. Finally, CMQD doesn't account for similarities between phoneme classes. While some of these deficiencies could be addressed with an extension to CMQD (e.g., by expanding the degradation lattices to include language model scores), we can do better using a more powerful modeling framework. In particular, we adopt the approach of phrase-based statistical machine translation (Koehn et al., 2003; Koehn and Hoang, 2007). This approach allows for multiple-phoneme to multiple-phoneme substitutions, as well as the soft incorporation of additional linguistic knowledge (e.g., phoneme classes). This is related to previous work allowing higher order phoneme confusions in bigram or trigram contexts (Chaudhari and Picheny, 2007), although they used a fuzzy edit distance measure and did not incorporate other evidence in their model (e.g., the phoneme language model score). The reader is referred to (Koehn and Hoang, 2007; Koehn et al., 2007) for detailed information about phrase-based statistical machine translation. We give a brief outline here, sufficient only to provide background for our query degradation application.

Statistical machine translation systems work by

converting a source-language sentence into the most probable target-language sentence, under a model whose parameters are estimated using example sentence pairs. Phrase-based machine translation is one variant of this statistical approach, wherein multiple-word *phrases* rather than isolated words are the basic translation unit. These phrases are generally not linguistically motivated, but rather learned from co-occurrences in the paired example translation sentences. We apply the same machinery to hypothesize our pronunciation degradations, where we now translate from the "source-language" reference phoneme sequence $Q$ to the hypothesized "target-language" phoneme sequence $H$.

Phrase-based translation is based on the noisy channel model, where Bayes rule is used to reformulate the translation probability for translating a reference query $Q$ into a hypothesized phoneme sequence $H$ as

$$\arg\max_H P(H|Q) = \arg\max_H P(Q|H)P(H).$$

Here, for example, $P(H)$ is the language model probability of a degradation $H$ and $P(Q|H)$ is the conditional probability of the reference sequence $Q$ given $H$. More generally however, we can incorporate other *feature functions* of $H$ and $Q$, $h_i(H, Q)$, and with varying weights. This is implemented using a log-linear model for $P(H|Q)$, where the model covariates are the functions $h_i(H, Q)$, so that

$$P(H|Q) = \frac{1}{Z} \exp \sum_{i=1}^{n} \lambda_i h_i(H, Q)$$

The parameters $\lambda_i$ are estimated by MLE and the normalizing $Z$ need not be computed (because we will take the argmax). Example feature functions include the language model probability of the hypothesis and a hypothesis length penalty.

In addition to feature functions being defined on the surface level of the phonemes, they may also be defined on non-surface annotation levels, called *factors*. In a word translation setting, the intuition is that statistics from morphological variants of a lexical form ought to contribute to statistics for other variants. For example, if we have never seen the word *houses* in language model training, but have examples of *house*, we still can expect *houses are* to be more probable than *houses fly*. In other words, factors allow us to collect improved statistics on sparse data. While sparsity might appear to be less of a problem for phoneme degradation modeling (because the token inventory is comparatively very small), we nevertheless may benefit from this approach, particularly because we expect to rely on higher order language models and because we have rather little training data: only 22,810 transcribed utterances (about 600k reference phonemes).

In our case, we use two additional annotation layers, based on a simple grouping of phonemes into broad classes. We consider the phoneme itself, the broad distinction of vowel and consonant, and a finer grained set of classes (e.g., front vowels, central vowels, voiceless and voiced fricatives). Figure 1 shows the three annotation layers we consider for an example reference phoneme sequence. After mapping the reference and hypothesized phonemes to each of these additional factor levels, we train language models on each of the three factor levels of the hypothesized phonemes. The language models for each of these factor levels are then incorporated as features in the translation model.

We use the open source toolkit `Moses` (Koehn et al., 2007) as our phrase-based machine translation system. We used the SRI language modeling toolkit to estimate interpolated 5-gram language models (for each factor level), and smoothed our estimates with Witten-Bell discounting (Witten and Bell, 1991). We used the default parameter settings for `Moses`'s training, with the exception of modifying `GIZA++`'s default maximum fertility from 10 to 4 (since we don't expect one reference phoneme to align to 10 degraded phonemes). We used default decoding settings, apart from setting the distortion penalty to prevent any reorderings (since alignments are logically constrained to never cross). For the rest of this chapter, we refer to our phrase-based query degradation model as PBQD. We denote the phrase-based model using factors as PBQD-Fac.

Figure 2 shows an example alignment learned for a reference and one-best phonemic transcript. The reference utterance "snow white and the seven dwarves" is recognized (approximately) as "no white a the second walks". Note that the phrase-based system is learning not only acoustically plausible confusions, but critically, also confusions aris-

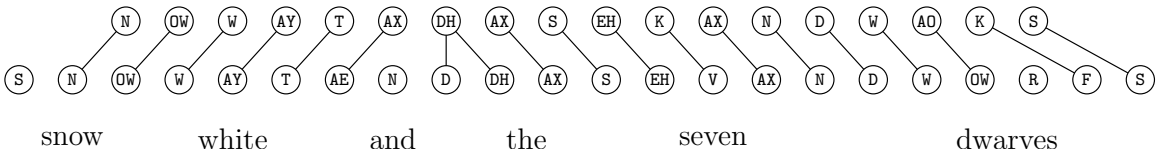


Figure 2: An alignment of hypothesized and reference phoneme transcripts from the multigram phoneme recognizer, for the phrase-based query degradation model.

ing from the phonemic recognition system's peculiar construction. For example, while V and K may not be acoustically similar, they are still confusable—within the context of S EH—because multigram language model data has many examples of the word *second*. Moreover, while the word *dwarves* (D-W-OW-R-F-S) is not present in the dictionary, the words *dwarf* (D-W-AO-R-F) and *dwarfed* (D-W-AO-R-F-T) are present (*N.B.*, the change of vowel from AO to OW between the OOV and in vocabulary pronunciations). While CMQD would have to allow a deletion and two substitutions (without any context) to obtain the correct degradation, the phrase-based system can align the complete phrase pair from training and exploit context. Here, for example, it is highly probable that the errorfully hypothesized phonemes W AO will be followed by *K*, because of the prevalence of *walk* in language model data.

## 4 Experiments

An appropriate and commonly used measure for RUR is Mean Average Precision (MAP). Given a ranked list of utterances being searched through, we define the *precision* at position $i$ in the list as the proportion of the top $i$ utterances which actually contain the corresponding query word. Average Precision (AP) is the average of the precision values computed for each position containing a relevant utterance. To assess the effectiveness of a system across multiple queries, Mean Average Precision is defined as the arithmetic mean of per-query average precision, $\mathrm{MAP} = \frac{1}{n}\sum_n \mathrm{AP}_n$. Throughout this paper, when we report statistically significant improvements in MAP, we are comparing AP for paired queries using a Wilcoxon signed rank test at $\alpha = 0.05$.

Note, RUR is different than spoken term detection in two ways, and thus warrants an evaluation measure (e.g., MAP) different than standard spoken term detection measures (such as NIST's *actual term weighted value* (Fiscus et al., 2006)). First, STD measures require locating a term with granularity finer than that of an utterance. Second, STD measures are computed using a fixed detection threshold. This latter requirement will be unnecessary in many applications (e.g., where a user might prefer to decide themselves when to stop reading down the ranked list of retrieved utterances) and unlikely to be helpful for downstream evidence combination (where we may prefer to keep all putative hits and weight them by some measure of confidence).

For our evaluation, we consider retrieving short utterances from seventeen fully transcribed MALACH interviews. Our query set contains all single words occurring in these interviews that are OOV with respect to the word dictionary. This gives us a total of 261 query terms for evaluation. Note, query words are also not present in the multigram training transcripts, in any language model training data, or in any transcripts used for degradation modeling. Some example query words include BUCHENWALD, KINDERTRANSPORT, and SONDERKOMMANDO.

To train our degradation models, we used a held out set of 22,810 manually transcribed utterances. We run each recognition system (phoneme, multigram, and word) on these utterances and, for each, train separate degradation models using the aligned reference and hypothesis transcripts. For CMQD, we computed 100 random traversals on each lattice, giving us a total of 2,281,000 hypothesis and reference pairs to align for our confusion matrices.

## 5 Results

We first consider an intrinsic measure of the three speech recognition systems we consider, namely Phoneme Error Rate (PER). Phoneme Error Rate is calculated by first producing an alignment of

the hypothesis and reference phoneme transcripts. The counts of each error type are used to compute $PER = 100 \cdot \frac{S+D+I}{N}$, where $S, D, I$ are the number of substitutions, insertions, and deletions respectively, while $N$ is the phoneme length of the reference. Results are shown in Table 1. First, we see that the PER for the multigram system is roughly half that of the phoneme-only system. Second, we find that the word system achieves a considerably lower PER than the multigram system. We note, however, that since these are not true phonemes (but rather phonemes copied over from pronunciation dictionaries and word transcripts), we must cautiously interpret these results. In particular, it seems reasonable that this framework will overestimate the strength of the word based system. For comparison, on the same train/test partition, our word-level system had a *word* error rate of 31.63. Note, however, that automatic word transcripts can not contain our OOV query words, so word error rate is reported only to give a sense of the difficulty of the recognition task.

Table 1 shows our baseline RUR evaluation results. First, we find that the generative model yields statistically significantly higher MAP using words or multigrams than phonemes. This is almost certainly due to the considerably improved phoneme recognition afforded by longer recognition units. Second, many more unique phoneme sequences typically occur in phoneme lattices than in their word or multigram counterparts. We expect this will increase the false alarm rate for the phoneme system, thus decreasing MAP.

Surprisingly, while the word-based recognition system achieved considerably lower phoneme error rates than the multigram system (see Table 1), the word-based generative model was in fact indistinguishable from the same model using multigrams. We speculate that this is because the method, as it is essentially a language modeling approach, is sensitive to data sparsity and requires appropriate smoothing. Because multigram lattices incorporate smaller recognition units, which are not constrained to be English words, they naturally produce smoother phoneme language models than a word-based system. On the other hand, the multigram system is also not statistically significantly better than the word-based generative model, suggesting this may be a promising area for future work.

Table 1 shows results using our degradation models. Query degradation appears to help all systems with respect to the generative baseline. This agrees with our intuition that, for RUR, low MAP on OOV terms is predominately driven by low recall.[1] Note that, at one degradation, CMQD has the same MAP as the generative model, since the most probable degradation under CMQD is almost always the reference phoneme sequence. Because the CMQD model can easily hypothesize implausible degradations, we see the MAP increases modestly with a few degradations, but then MAP decreases. In contrast, the MAP of the phrase-based system (PBQD-Fac) increases through to 500 query degradations using multigrams. The phonemic system appears to achieve its peak MAP with fewer degradations, but also has a considerably lower best value.

The non-factored phrase-based system PBQD achieves a peak MAP considerably larger than the peak CMQD approach. And, likewise, using additional factor levels (PBQD-Fac) also considerably improves performance. Note especially that, using multiple factor levels, we not only achieve a higher MAP, but also a higher MAP when only a few degradations are possible.

To account for errors in phonemic recognition, we have taken two steps. First, we used longer recognition units which we found significantly improved MAP while using our baseline RUR technique. As a second method for handling recognition errors, we also considered variants of our ranking function. In particular, we incorporated query degradations hypothesized using factored phrase-based machine translation. Comparing the MAP for PBQD-Fac with MAP using the generative baseline for the most improved indexing system (the word system), we find that this degradation approach again statistically significantly improved MAP. That is, these two strategies for handling recognition errors in RUR appear to work well in combination.

Although we focused on vocabulary-independent RUR, downstream tasks such as *ad hoc* speech retrieval will also want to incorporate evidence from in-vocabulary query words. This makes

[1] We note however that the preferred operating point in the tradeoff between precision and recall will be task specific. For example, it is known that precision errors become increasingly important as collection size grows (Shao et al., 2008).

| Method | Phone Source | PER | QD Model | Baseline | Query Degradations | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | 1 | 5 | 50 | 500 |
| Degraded Model | Phonemes | 64.4 | PBQD-Fac | 0.0387 | 0.0479 | 0.0581 | **0.0614** | 0.0612 |
| | Multigrams | 32.1 | CMQD | 0.1258 | 0.1258 | 0.1272 | 0.1158 | 0.0991 |
| | Multigrams | 32.1 | PBQD | 0.1258 | 0.1160 | 0.1283 | 0.1347 | 0.1317 |
| | Multigrams | 32.1 | PBQD-Fac | 0.1258 | 0.1238 | 0.1399 | 0.1510 | **0.1527** |
| | Words | 20.5 | PBQD-Fac | 0.1255 | 0.1162 | 0.1509 | **0.1787** | 0.1753 |

Table 1: PER and MAP results for baseline and degradation models. The best result for each indexing approach is shown in bold.

our query degradation approach which indexed phonemes from word-based LVCSR particularly attractive. Not only did it achieve the best MAP in our evaluation, but this approach also allows us to construct recognition lattices for both in and out-of-vocabulary query words without running a second, costly, recognition step.

## 6 Conclusion

Our goal in this work was to rank utterances by our confidence that they contained a previously unseen query word. We proposed a new approach to this task using hypothesized degradations of the query word's phoneme sequence, which we produced using a factored phrase-based machine translation model. This approach was principally motivated by the mismatch between the query's phonemes and the recognition phoneme sequences due to errorful speech indexing. Our approach was constructed and evaluated using phoneme-, multigram-, and word-based indexing, and significant improvements in MAP using each indexing system were achieved. Critically, these significant improvements were in addition to the significant gains we achieved by constructing our index with longer recognition units.

While PBQD-Fac outperformed CMQD averaging over all queries in our evaluation, as expected, there may be particular query words for which this is not the case. Table 2 shows example degradations using both the CMQD and PBQD-Fac degradation models for multigrams. The query word is *Mengele*. We see that CMQD degradations are near (in an edit distance sense) to the reference pronunciation (`M-EH-NX-EY-L-EH`), while the phrase-based degradations tend to sound like commonly occurring words (*mental, meant a lot, men they. . . , mentally*). In this case, the lexical phoneme sequence does not occur in the PBQD-Fac degradations until degradation nineteen. Because deleting `EH` has the same cost irrespective of context for CMQD, both CMQD degradations 2 and 3 are given the same pronunciation weight. Here, CMQD performs considerably better, achieving an average precision of 0.1707, while PBQD-Fac obtains only 0.0300. This suggests that occasionally the phrase-based language model may exert too much influence on the degradations, which is likely to increase the incidence of false alarms. One solution, for future work, might be to incorporate a false alarm model (e.g., down-weighting putative occurrences which look suspiciously like non-query words). Second, we might consider training the degradation model in a discriminative framework (e.g., training to optimize a measure that will penalize degradations which cause false alarms, even if they are good candidates from the perspective of MLE). We hope that the ideas presented in this paper will provide a solid foundation for this future work.

| CMQD | Phrase-based |
|---|---|
| `M-EH-NX-EY-L-EH` | `M-EH-N-T-AX-L` |
| `M-EH-NX-EY-L` | `M-EH-N-T-AX-L-AA-T` |
| `M-NX-EY-L-EH` | `AH-AH-AH-AH-M-EH-N-T-AX-L` |
| `M-EH-NX-EY-EH` | `M-EH-N-DH-EY-L-EH` |
| `M-EH-NX-L-EH` | `M-EH-N-T-AX-L-IY` |

Table 2: The top five degradations and associated probabilities using the CMQD and PBQD-Fac models, for the term *Mengele* using multigram indexing.

## References


W. Byrne et al. 2004. Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives. *IEEE Transactions on Speech and Audio Processing, Special Issue on Spontaneous Speech Processing*, 12(4):420–435, July.

U.V. Chaudhari and M. Picheny. 2007. Improvements in phone based audio search via constrained match with high order confusion estimates. *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, pages 665–670, Dec.

Kareem Darwish and Walid Magdy. 2007. Error correction vs. query garbling for Arabic OCR document retrieval. *ACM Trans. Inf. Syst.*, 26(1):5.

Kareem M. Darwish. 2003. *Probabilistic Methods for Searching OCR-Degraded Arabic Text*. Ph.D. thesis, University of Maryland, College Park, MD, USA. Directed by Bruce Jacob and Douglas W. Oard.

S. Deligne and F. Bimbot. 1997. Inference of Variable-length Acoustic Units for Continuous Speech Recognition. In *ICASSP '97: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1731–1734, Munich, Germany.

Jonathan Fiscus et al. 2006. English Spoken Term Detection 2006 Results. In *Presentation at NIST's 2006 STD Eval Workshop*.

J.T. Foote et al. 1997. Unconstrained keyword spotting using phone lattices with application to spoken document retrieval. *Computer Speech and Language*, 11:207—224.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *EMNLP '07: Conference on Empirical Methods in Natural Language Processing*, June.

Philipp Koehn et al. 2003. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.

Philipp Koehn et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL '07: Proceedings of the 2007 Conference of the Association for Computational Linguistics, demonstration session*, June.

Okan Kolak. 2005. *Rapid Resource Transfer for Multilingual Natural Language Processing*. Ph.D. thesis, University of Maryland, College Park, MD, USA. Directed by Philip Resnik.

Jonathan Mamou and Bhuvana Ramabhadran. 2008. Phonetic Query Expansion for Spoken Document Retrieval. In *Interspeech '08: Conference of the International Speech Communication Association*.

Spyros Matsoukas et al. 2005. The 2004 BBN 1xRT Recognition Systems for English Broadcast News and Conversational Telephone Speech. In *Interspeech '05: Conference of the International Speech Communication Association*, pages 1641–1644.

K. Ng and V.W. Zue. 2000. Subword-based approaches for spoken document retrieval. *Speech Commun.*, 32(3):157–186.

J. Scott Olsson. 2008a. Combining Speech Retrieval Results with Generalized Additive Models. In *ACL '08: Proceedings of the 2008 Conference of the Association for Computational Linguistics*.

J. Scott Olsson. 2008b. Vocabulary Independent Discriminative Term Frequency Estimation. In *Interspeech '08: Conference of the International Speech Communication Association*.

Pavel Pecina, Petra Hoffmannova, Gareth J.F. Jones, Jianqiang Wang, and Douglas W. Oard. 2007. Overview of the CLEF-2007 Cross-Language Speech Retrieval Track. In *Proceedings of the CLEF 2007 Workshop on Cross-Language Information Retrieval and Evaluation*, September.

R. Prasad et al. 2005. The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech Recognition System. In *Interspeech '05: Conference of the International Speech Communication Association*.

S.E. Robertson. 1977. The Probability Ranking Principle in IR. *Journal of Documentation*, pages 281–286.

M. Saraclar and R. Sproat. 2004. Lattice-Based Search for Spoken Utterance Retrieval. In *NAACL '04: Proceedings of the 2004 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*.

P. Schone et al. 2005. Searching Conversational Telephone Speech in Any of the World's Languages.

Jian Shao et al. 2008. Towards Vocabulary-Independent Speech Indexing for Large-Scale Repositories. In *Interspeech '08: Conference of the International Speech Communication Association*.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *ICSLP '02: Proceedings of 2002 International Conference on Spoken Language Processing*.

I. H. Witten and T. C. Bell. 1991. The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression. *IEEE Trans. Information Theory*, 37(4):1085–1094.

Peng Yu and Frank Seide. 2005. Fast Two-Stage Vocabulary-Independent Search In Spontaneous Speech. In *ICASSP '05: Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing*.

P. Yu et al. Sept. 2005. Vocabulary-Independent Indexing of Spontaneous Speech. *IEEE Transactions on Speech and Audio Processing*, 13(5):635–643.