A HYBRID WORD / PHONEME-BASED APPROACH FOR IMPROVED VOCABULARY-INDEPENDENT SEARCH IN SPONTANEOUS SPEECH

Peng Yu and Frank Seide

Microsoft Research Asia, 5F Beijing Sigma Center, No. 49 Zhichun Rd., 100080 Beijing, P.R.C.
{t-roger,yfseide}@microsoft.com

ABSTRACT

For efficient organization of speech recordings – meetings, interviews, voice mails, and lectures – being able to search for spoken keywords is essential. Today, most spoken document retrieval systems use large-vocabulary recognition. For the above scenarios, such systems suffer from the unpredictable domain, out-of-vocabulary queries, and generally high word-error rate (WER). In [1], we presented a system for phonetic indexing and searching of spontaneous speech. It is vocabulary-independent and based on phoneme lattices. In the present paper, we propose to combine it with word-based search into a hybrid approach.

We explore two methods of combination: posterior combination (merging search results of a word-based and a phoneme-based system) and prior combination (combining word and phoneme language models and vocabularies to form a hybrid recognizer).

The search accuracy of our best purely phonetic baseline is 64% (Figure of Merit), and our purely word-based baselines are below 50%. The new hybrid approach achieves 73%, if the recognizer uses a language model that matches the test-set domain. With a mismatched language model, 71% is achieved.

Our results show that the proposed hybrid model benefits from the best of two worlds: Word-level language context and robustness of phonetic search to unknown words and domain mismatch.

1. INTRODUCTION

Audio search – finding spoken words in speech recordings – is an important topic. Several approaches are reported in literature, e.g. in the Text Retrieval Conference (TREC) Spoken-Document Retrieval (SDR) track [2]. Most TREC benchmarking systems use well-tuned domain-specific recognizers to generate approximate transcripts, and then do text-based information retrieval. With low word-error rates around 20% and redundancy in audio segments and queries, retrieval accuracies similar to using human reference transcripts are achieved. However, this approach is not suitable for our conversational-speech scenarios, where vocabulary and language domain are unpredictable, and queries are often single keywords.

In [1], we presented a phonetic approach to vocabulary-independent audio search. Unlike most SDR systems, we search lattices instead of top-1 recognizer output to improve recall. By using phonene lattices instead of word lattices, also out-of-vocabulary (OOV) keywords can be found with similar accuracy.

Closer inspection of our results in [1] had shown that word-lattice based search and phonemic search produce somewhat complementary search results: Phonetic search tends towards low miss rates with many false alarms, and word-level search towards few false alarms but high miss rates. This gives rise to a hybrid approach.

A word / phoneme hybrid model was explored in [3]. The aim was to handle OOV words in speech recognition. The author added a “OOV” word to the dictionary and defined its pronunciation by a phoneme language model. Recognition of speech containing OOV words was improved. Note that here the phonetic language model should only take effect where in-vocabulary words score low and an OOV word should be hypothesized. However, in our search task we want to benefit from the low miss rates of phonetic search for both in-vocabulary and OOV keywords. Thus, phoneme language-model information should be used for both.

In this paper, we propose to combine word- and phoneme-based word spotting into a single hybrid approach. We explore two ways: posterior combination and prior combination. Both outperform our individual word and phonetic baselines. We compare performance using word language models trained in-domain (as in [1]) and out-of-domain (more realistic in many usage scenarios).

The paper is organized as follows. In section 2 we will recapitulate lattice-based audio search [1]. Sections 3, 4, and 5 will introduce the hybrid approach and its two variants, posterior and prior combination. Section 6 gives the experimental results.

2. LATTICE-BASED WORD SPOTTING

Here we want to briefly recapitulate our phonetic approach to vocabulary-independent search in speech presented in [1]. In phonetic search, “indexing” consists of speech recognition to generate a phonetic representation of each audio file – a “lattice” of scored phoneme hypotheses. “Search” means rapidly locating all sub-paths in the lattice set that match the query string’s phonetic representation, and it is “found” where a match’s “confidence” is above a certain threshold.

We define the “confidence” of a match as its posterior probability \( P(W_t, t_o | O) \), i.e. the sum of the probabilities of all paths that contain the query string \( W \) from \( t_1 \) to \( t_c \):

\[
P(W, t, t_o | O) = \sum_{W_{O}} \sum_{t_{1}} \frac{p(O, t, W_{1} | W_{2} \ldots W_{i})}{p(O, W_{i} | W_{1} \ldots W_{i-1})} \frac{p(W_{i} | W_{i-1})}{p(W_{i})} \frac{p(O | W_{i})}{p(O)}
\]

with \( W_{-} \) and \( W_{+} \) denoting any word sequence before \( t_{c} \) and after \( t_{o} \), respectively; \( W \) being any word sequence. Eq. (1) can be efficiently approximated by forward-backward scoring of lattices. During phoneme-lattice generation, we do not use a phoneme M-gram language model, but a longer-span phonetic word-fragment language model, for better focused beam search and lattice pruning.

The set of phonetic word fragments is automatically generated [4, 3], examples are /-/k-ih-ng/ (the syllable /king/) and /ih-z/ (the word is). All phonetic results below use this fragment model.

In [1] we compared phonetic search against searching on word lattices. For in-vocabulary keywords (keywords in the vocabulary of the word-based system), the phonetic approach nearly reaches the search accuracy of the word-based system. For out-of-vocabulary words (which cannot be found at all by the word-based approach), it still maintains similar accuracy.
3. HYBRID WORD / PHONETIC SEARCH

We consider two possibilities for integrating word and phonetic-level search into a hybrid approach:

- **Posterior combination.** For each utterance, word spotting is run separately word and phoneme based. The two sets of matches are merged by combining posterior probabilities.

- **Prior combination.** The prior information, i.e. language models and vocabularies, is combined. With this, hybrid lattices are created that may contain both phonetic and graphemic versions of a word.

Prior combination can be done by utterance or on word level. This is illustrated in Figs. 1 and 2. Hybrid queries are formulated by adding the graphemes as possible pronunciations of keywords into the search grammar, for example:

"accept" = ( ACCEPT | eh k s eh p t )

4. POSTERIOR COMBINATION

The starting point are two separate lattices for each utterance, a word and a phoneme-based one. Since different language models were used, scores may not be directly comparable. For posterior probabilities, however, this effect cancels out to a certain degree. This gives rise to the posterior combination method – running two separate searches and determining a match’s combined confidence score as a linear combination of the two posterior probabilities:

\[
P(W_t, t_e | O) = P_{WD} \cdot \sum_{(W, W_e) \in L_{WD}} p(Ot, t_e W W_e | M_{WD}) + P_{PH} \cdot \sum_{(W, W_e) \in L_{PH}} p(Ot, t_e W W_e | M_{PH})
\]

Here, \( L_{WD} \) and \( L_{PH} \) represent the word and phoneme lattice, respectively, and \( M_{WD} \) and \( M_{PH} \) the corresponding dictionary and language models. \( P_{WD} \) and \( P_{PH} \) are prior probabilities used to balance the importance of each contribution.

5. PRIOR COMBINATION

Prior combination refers to combination of the prior probabilities, i.e. the language models (and associated dictionaries). While this can be realized by combining two lattices, the ultimate goal is to eliminate the two recognition passes, and to generate a hybrid lattice in a single hybrid recognition pass.

5.1. Utterance-level

The simplest form is utterance-level combination, which is very similar to posterior combination. Mid-utterance transitions between words and phonemes are not allowed here. Thus, it can be implemented easily by merging separately generated word and phoneme lattices by connecting the start and end nodes (Fig. 1). The disadvantage of this method is that it also requires two separate recognition passes.

5.2. Word-level

For single-pass hybrid recognition, we need a joint hybrid language model. It should be accurate for in-vocabulary words, and at the same time model OOV words.

5.2.1. Hybrid language model

We chose a simple linear interpolation between a word-level and a phoneme-level language model:

\[
P(w|h) = P_{WD} \cdot P(w|h, M_{WD}) + P_{PH} \cdot P(\text{tr}(w)|\text{tr}(h), M_{PH})
\]

where \( w \) is the word to be predicted (in- or out-of-vocabulary), \( h \) the history, \( M_{WD} \) and \( M_{PH} \) the word and phonetic language model, respectively, and \( \text{tr}() \) the phonetic transcription operator. In our system, the word-level language model is a word trigram (which assigns probability 0 to unknown words), while the phoneme-level one is a phonetic word-fragment bigram as introduced in [1]. Figure 3 illustrates the resulting structure.

Notice that by this interpolation, the phoneme level language model is activated for both in-vocabulary and out-of-vocabulary words. As discussed in section 1, this is to avoid suffering an accuracy loss from the higher miss rate of the word-level model.

5.2.2. Implementation by word-level hybrid lattices

The criterion and model defined by Eqs. (1) and (2) can be efficiently implemented by a word-level hybrid lattice, which contains both word and phoneme edges like the one shown in Fig. 2. The summation in Eq. (1) is over all paths that match the query string. In the hybrid lattice, this includes both grapheme and phonetic matches. If we assign the first summand in Eq. (2) to word edges (including the weight) and the second summand to phonetic edges, Eq. (1) will evaluate the correct interpolated language model probabilities.

5.2.3. Within-word phonetic language model

The phonetic component of the hybrid language model should model phoneme sequences of single OOV words. Cross-OOV-word transitions should be realized by the word-level part. Thus, we trained the phonetic part on a special version of the training set.
Table 1. Test-set perplexities (PP) and top-1 word error rates (WER) for the word-based in-domain and out-of-domain setups.

<table>
<thead>
<tr>
<th></th>
<th>in domain (Voicemail)</th>
<th>out of domain (Switchboard)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>word LM:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PP</strong></td>
<td>102</td>
<td>442</td>
</tr>
<tr>
<td><strong>WER</strong></td>
<td>43.9%</td>
<td>60.6%</td>
</tr>
</tbody>
</table>

The word-based language models, there are two versions for all word-level setups – one in-domain and one out-of-domain. The in-domain setup was trained on the VM-I training transcriptions (about 160,000 words, as in [1]), and the out-of-domain setup on Switchboard transcriptions. Recognition vocabularies are the training-set word lists (7469 and 27463 words, respectively). Table 1 shows the test-set perplexities and top-1 word error rates. The keywords list is same as used in [1], consisting of 2049 entries. For the in-domain dictionary (Voicemail), 620 entries are OOV (30.3%). For the out-of-domain dictionary (Switchboard), 530 entries are OOV (25.9%).

We measure word-spotting accuracy by the common “Figure Of Merit” (FOM) defined by NIST (National Institute of Standards & Technology) as the average of the detection/false-alarm curve over the range [0..10] false alarms per hour per keyword.

6.2. Word lattice and phoneme lattice baseline

Table 2 shows search accuracies (FOM) for the two baseline setups: The first row (“word”) for word spotting on word lattices, the second line (“phonetic”) on phonetic lattices. Let us first look at the left half (“in domain” section). The word-based setup achieves a FOM of 68.3% for the in-vocabulary (“INV”) keywords – keep in mind here that the language model used by the recognizer well matches the test set (both Voicemail).

For the same keyword set, the phoneme-lattice based search (“phonetic”) achieves 63.0%. This is not as good as the word-based setup, but consider that the phoneme-based system’s language model is out of domain. Insightful are the upper-bound recall rates here, i.e. the percentage of keywords detected in the lattice. The word-based approach achieves its FOM of 68.3% at an upper-bound recall of 75.7%, while phonetic approach reaches FOM 63.0% with an upper-bound recall of 96.2%! This confirms that the information in both lattices is complementary, which is the main motivation for combining the two as proposed in this paper.

Keywords that are out-of-vocabulary (“OOV”) can obviously not be found by the word-based setup, reducing the overall accuracy (“all”) to 49.3%, while the phonetic search maintains similar accuracy for the OOV words, and achieves an overall FOM of 64.0%.

Looking at the right half (“out of domain”), the situation for the INV keywords is reversed: Due to the domain mismatch, the word-based setup achieves a FOM of 68.3% for the in-vocabulary (“INV”) keywords – keep in mind here that the language model used by the recognizer well matches the test set (both Voicemail). The phonetic search maintains similar accuracy for the OOV keywords, and achieves an overall FOM of 64.0%.

Table 2. Search accuracy (Figure of Merit, FOM) for word spotting on word lattices vs. phonetic lattices. We compare two different word-level setups, “in domain” (word LM trained on Voicemail) and “out of domain” (on Switchboard). The phonetic setup always uses an out-of-domain LM, but we report results for two different INV / OOV splits matching the two word-level setups.

<table>
<thead>
<tr>
<th></th>
<th>in domain</th>
<th>out of domain</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>word LM:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>word</strong></td>
<td>49.3</td>
<td>38.3</td>
</tr>
<tr>
<td><strong>phonetic</strong></td>
<td>64.0</td>
<td>64.4</td>
</tr>
</tbody>
</table>

In real-life scenarios, a well-matched language model is often not available. To assess the effect of domain dependence of

![Fig. 3. Topology of the hybrid language model. For simplicity, the word-level section does not show the bigram back-off branches, and the phonetic section disregards the complex expansion into different phonetic fragment histories (just denoted by * instead).](image)
based search only achieves a FOM of 50.4%, while the phonetic match yields a FOM of 64.4% (and again similar accuracy for OOV words). Notice that the “phonetic” results on the right half are the same as on the left half, but split differently into INV / OOV keywords w.r.t. the respective word-based vocabularies.

Our previous work [1] only considered the in-domain scenario, and let us conclude that phonetic search is beneficial for out-of-vocabulary scenarios. Our new results show that it also is for in-vocabulary keywords in the case of a domain mismatch.

### 6.3. Utterance-level posterior and prior combination

Table 3 shows the results for combining the lattices of the baselines of Table 2, the first row for posterior and the second for utterance-level prior combination.

The power of posterior combination is seen from the INV results. 1 In-domain, FOMs are improved from 68.3% (word baseline) and 63.0% (phonetic baseline) to 76.7% (hybrid). It still works well for the out-of-domain setup, where the word-based baseline is much worse (50.4%), an improvement to 73.0% is achieved.

The prior combination on utterance level, on the other hand, does not work. We observe that the difference in total log likelihood between word and phonetic lattices is approximately proportional to the utterance duration. This introduces a great deal of randomness to the weighting of the word vs. phonetic contribution in posterior computation. FOMs are as low as 36.9%.

### 6.4. Word-level prior combination

Table 4 shows search accuracies with word-level prior combination. For the first row (“prior (lattice)”), the lattices from Table 2 were combined to form a word-level hybrid lattice as shown in Fig. 2. Search accuracies similar to posterior combination (Table 3) are achieved. Word-level prior combination is principally working. For the second row (“prior (1-pass)”), the same type of word-level hybrid lattices were generated by a single-pass recognizer that uses the joint hybrid language model described in section 5.2.1. Our previous work [1] only considered the in-domain scenario, and let us conclude that phonetic search is beneficial for out-of-vocabulary scenarios. Our new results show that it also is for in-vocabulary keywords in the case of a domain mismatch.

We have extended our phonetic-lattice based approach to spontaneous speech search with a hybrid model. First, compared to searching word lattices, our phonetic search leads to significant accuracy improvements for both out-of-vocabulary and mismatching domain situations. In the out-of-domain scenario, phonetic search not only allows to find out-of-vocabulary words, but also outperforms word-based search for in-vocabulary words.

Second, we have presented two hybrid word / phoneme-based approaches to search: posterior and prior combination. When implemented by combining individually generated lattices, FOMs for in-vocabulary keywords are improved to 72-76% compared to the pure phonetic baseline (64%). However, when implemented by single-pass recognition using a hybrid language model, only about half the gain is realized, due to pruning errors caused by the mismatching score levels of the word and phonetic language model. Our results have shown that the proposed hybrid model successfully integrates the best of two worlds: High accuracy from the word-level language context and robustness to unknown words and domain mismatch from phonetic search.

### 7. CONCLUSION

We have extended our phonetic-lattice based approach to spontaneous speech search with a hybrid model. First, compared to searching word lattices, our phonetic search leads to significant accuracy improvements for both out-of-vocabulary and mismatching domain situations. In the out-of-domain scenario, phonetic search not only allows to find out-of-vocabulary words, but also outperforms word-based search for in-vocabulary words.

Second, we have presented two hybrid word / phoneme-based approaches to search: posterior and prior combination. When implemented by combining individually generated lattices, FOMs for in-vocabulary keywords are improved to 72-76% compared to the pure phonetic baseline (64%). However, when implemented by single-pass recognition using a hybrid language model, only about half the gain is realized, due to pruning errors caused by the mismatching score levels of the word and phonetic language model. Our results have shown that the proposed hybrid model successfully integrates the best of two worlds: High accuracy from the word-level language context and robustness to unknown words and domain mismatch from phonetic search.

### 8. ACKNOWLEDGEMENTS

The authors wish to thank Asela Gunawardana and Issam Bazzi for sharing their acoustic models for Switchboard, Alex Acero and Ciprian Chelba for fruitful discussions, and Mark Gales of Cambridge University for sharing their \texttt{wtest} set definition.

### 9. REFERENCES


