Open-Vocabulary Retrieval of Spoken Content with Shorter/Longer Queries
Considering Word/Subword-based Acoustic Feature Similarity

Hung-yi Lee¹, Po-wei Chou², Lin-shan Lee¹

¹Graduate Institute of Communication Engineering, National Taiwan University
²Department of Electrical Engineering, National Taiwan University

{ tlkagkb93901106, doomhydra1098 }@gmail.com

Abstract

Acoustic feature similarity between utterances has been shown to be very helpful for spoken term detection using pseudo-relevance feedback (PRF) and graph-based re-ranking. Both cases are based on the concept that utterances similar to those utterances with higher relevance scores in acoustic features should have higher scores, while graph-based re-ranking further considers the similarity structure between utterances with a graph. In this paper, we extend these approaches to consider acoustic feature similarity between utterances over both word and subword lattices, and offer a complete formulation for the general problem of open vocabulary retrieval of spoken content with shorter or longer queries. All these are verified by significant improvements in preliminary experiments with both in-vocabulary (IV) and OOV queries.

Index Terms: Spoken Content Retrieval, Pseudo-relevance Feedback, Random Walk

1. Introduction

Spoken content retrieval is believed to be very important in the future when people try to access the multimedia content over the Internet based on the included audio signals. In most cases, two processing stages are needed for spoken content retrieval [1], that is, the audio content is first transcribed into lattices, and the retrieval engine then searches through the lattices based on the user query and returns a list of relevant spoken utterances. When OOV words are present in the query, word-based lattices will not be adequate, and in such cases subword-based approaches are very useful. For the task considered in this paper, the query is in text form but can be shorter or longer (consisting of one to several words), and the goal is to return a list of spoken utterances containing the query.

Since the same words may have similar pronunciation thus similar acoustic feature sequences, acoustic feature similarity between utterances can be useful for spoken content retrieval. One way to realize this concept is by pseudo-relevance feedback (PRF) [2, 3]. In this approach, given a user query, the retrieval engine first searches through the lattices to produce a first-pass returned list, ranked according to a relevance score. A pseudo-relevant utterance set is then defined from the first-pass returned list to produce a pseudo-relevant utterance set. The similarity between each first-pass retrieved utterance and this pseudo-relevant utterance set is computed based on acoustic features, and the first-pass returned list is re-ranked accordingly. The PRF approach can be moved one step forward with graph-based re-ranking [4, 5]. In this approach, a graph is constructed for the first-pass retrieved utterances in which each node represents an utterance and the edges represent acoustic feature similarity between utterances. Based on the concept that utterances strongly connected to many utterances with high scores on the graph should have higher scores, the relevance scores for the utterances propagate over the graph, and then the utterances are re-ranked accordingly. In this way the utterances in the first-pass returned list are considered globally, rather than assuming a pseudo-relevant utterance set in the PRF approach. However, in the prior works the above approaches of utilizing acoustic feature similarity were formulated based on a relatively limited task in which the query includes only a single word which is in vocabulary, and the acoustic feature similarity was based on word lattices.

In this paper, we generalize the above approaches utilizing acoustic feature similarity and formulate a more complete task: the query can be shorter/longer including one to several words in-vocabulary (IV) or OOV, while considering word/subword-based acoustic feature similarity for re-ranking. This is towards the goal of open-vocabulary retrieval of spoken content.

2. Proposed Approach

The framework for the proposed approach for the considered task is shown in Fig. 1. The utterances in the spoken archive are first transcribed into word or subword lattices by a speech recognizer. When the user enters a query Q, which can be shorter or longer including words in-vocabulary (IV) or OOV, the retrieval engine searches over the lattices and produces the first-pass returned list X ranked by the relevance score R(x) for each utterance x, as described in Subsection 2.1. The feature space similarity S(x, y) between retrieved utterances x and y is computed as presented in Subsection 2.2. Pseudo-

Figure 1: The framework for open-vocabulary retrieval considering word/subword-based acoustic feature similarity.
relevance feedback (PRF) is then performed as mentioned in Subsection 2.3, and the graph-based re-ranking finally applied as discussed in Subsection 2.4. The list finally re-ranked by the graph-based approach is displayed to the user.

2.1. First Pass

The relevance score \( R(x_i) \) used to produce the first-pass returned list can be derived from either word or subword lattices, depending on which kinds of lattices are indexed. Relevance scores from word lattices are usually more accurate than those from subword lattices, but we have to rely on the latter when the query \( Q \) consists of OOV words. Below we first present the way to obtain \( R(x_i) \) based on word lattices, and then show that those based on subword lattices can be obtained similarly.

Given a query \( Q \) consisting of one to several words, \( Q = \{w_j, j = 1, 2, \ldots, N\} \), \( w_j \) being the \( j \)-th word and \( N \) the number of words in \( Q \). To compute \( R(x_i) \) for an utterance \( x_i \) from the word lattice, we calculate the expected count for each \( n \)-gram \( \{w_i, \ldots, w_{i+n-1}\} \), \( i = 1, \ldots, N - n + 1 \), in the query from the lattice of \( x_i \) as in (1), and then aggregate the results for all such \( n \)-grams to produce a score \( R_{n\text{-gram}}(x_i, Q) \) for each order of \( n \) in (2).

\[
E[w_i, \ldots, w_{i+n-1}|x_i] = \sum_{u \in W(x_i)} P(x_i|u)P(u|\{w_i, \ldots, w_{i+n-1}\})
\]

\[
R_{n\text{-gram}}(x_i, Q) = \sum_{i=1}^{N-n+1} E[w_i, \ldots, w_{i+n-1}|x_i].
\]

The different proximity types, one for each \( n \)-gram order \( n \) allowed by the query length, are finally integrated in a weighted sum to give the relevance score \( R(x_i) \) for word lattices as in (3),

\[
R(x_i) = \sum_{n=1}^{N} a_n R_{n\text{-gram}}(x_i, Q),
\]

where \( a_n \) is a weight parameter.

The relevance score \( R(x_i) \) based on subword lattices is exactly the same as those in (1) - (3), except that the query is represented as a sequence of subword units, \( \{s_j, j = 1, 2, \ldots, M\} \), where \( s_j \) is the \( j \)-th subword unit and \( M \) the number of subword units in \( Q \), and \( E[s_i, \ldots, s_{i+n-1}|x_i] \) is computed on a subword lattice.

2.2. Acoustic Feature Similarity

Here the acoustic feature similarity \( S(x_i, x_j) \) between retrieved utterances \( x_i \) and \( x_j \) is computed, which will be used in PRF and graph-based re-ranking in the next two subsections. \( S(x_i, x_j) \) can be obtained again based on either words or subword units, and here we show the word-based version for demonstration.

Given a query \( Q \) consisting of a sequence of words \( \{w_j, j = 1, 2, \ldots, N\} \), for each \( n \)-gram \( \{w_i, \ldots, w_{i+n-1}\} \) in \( Q \), dynamic time warping (DTW) distance [6] is first performed between the acoustic feature sequences corresponding to the subpaths in the lattices of \( x_i \) and \( x_j \) for word hypotheses \( \{w_i, \ldots, w_{i+n-1}\} \). An example is shown in Fig. 2. This gives \( d(x_i, x_j; \{w_i, \ldots, w_{i+n-1}\}) \), the DTW distance between \( x_i \) and \( x_j \) considering the \( n \)-gram \( \{w_i, \ldots, w_{i+n-1}\} \) in the query. The similarity \( S(x_i, x_j; \{w_i, \ldots, w_{i+n-1}\}) \) between \( x_i \) and \( x_j \) considering \( \{w_i, \ldots, w_{i+n-1}\} \) is then obtained in (4),

\[
S(x_i, x_j; \{w_i, \ldots, w_{i+n-1}\}) = 1 - \frac{d(x_i, x_j; \{w_i, \ldots, w_{i+n-1}\}) - d_{\text{min}}}{d_{\text{max}} - d_{\text{min}}},
\]

where \( d_{\text{max}} \) and \( d_{\text{min}} \) are the largest and smallest values of \( d(x_i, x_j; \{w_i, \ldots, w_{i+n-1}\}) \) for all pairs of utterances in the first-pass returned list. Equation (4) simply normalizes the DTW distance and transforms it into a similarity score between 0 and 1. Below in the experiments MFCC features were used as the acoustic features. If the \( n \)-gram \( \{w_i, \ldots, w_{i+n-1}\} \) does not exist in the lattice of either \( x_i \) or \( x_j \), \( S(x_i, x_j; \{w_i, \ldots, w_{i+n-1}\}) \) is set to 0. Then we aggregate the similarities considering all such \( n \)-grams to produce a score \( S_{n\text{-gram}}(x_i, x_j) \) for each order of \( n \) in (5).

\[
S_{n\text{-gram}}(x_i, x_j) = \sum_{i=1}^{N-n+1} S(x_i, x_j; \{w_i, \ldots, w_{i+n-1}\}).
\]

The different proximity types are finally integrated in a weighted sum to give the similarity between \( x_i \) and \( x_j \) as in (6),

\[
S(x_i, x_j) = \sum_{n=1}^{N} b_n S_{n\text{-gram}}(x_i, x_j),
\]

where \( b_n \) is another weight parameter. The computation of \( S(x_i, x_j) \) based on subword units is exactly the same as those in (4) - (6), except replacing each word \( w_i \) by a subword unit \( s_j \).

Although we can obtain the relevance score \( R(x_i) \) and similarity \( S(x_i, x_j) \) based on different units, for example, it is possible to derive \( R(x_i) \) from word lattices but compute \( S(x_i, x_j) \) on subword lattices, for simplicity in the experiments below we always use \( R(x_i) \) and \( S(x_i, x_j) \) obtained from the same types (word or subword) of lattices.

1If there are multiple subpaths whose word hypotheses are \( \{w_i, \ldots, w_{i+n-1}\} \) in a lattice, only the one with the highest posterior probability is considered.
2.3. Pseudo-relevance Feedback (PRF)

In PRF, a pseudo-relevant utterance set $Y$ is defined as the top $K$ utterances in the first-pass returned list $X$ with the highest relevance scores $R(x_i)$. The similarity between each utterance $x_i$ in $X$ and the set $Y$ is then defined as

$$SIM(x_i, Y) = \frac{1}{K} \sum_{x_j \in Y} S(x_i, x_j).$$  \hspace{1cm} (7)

The relevance score $R(x_i)$ for each utterance $x_i$ is then updated into a new relevance score $R_p(x_i)$,

$$R_p(x_i) = R(x_i)SIM(x_i, Y)^{\delta_1},$$  \hspace{1cm} (8)

where $\delta_1$ is a weight parameter. The utterances in $X$ are then re-ranked accordingly.

2.4. Graph-based Re-ranking

Here a graph is constructed from the first-pass returned list $X$, in which each node represents an utterance. The utterances $x_i$ and $x_j$ with acoustic feature similarity $S(x_i, x_j)$ in (6) exceeding a threshold is connected by an edge whose weight is the similarity $S(x_i, x_j)$. Then a set of new graph-based relevance scores $R_g(x_i)$ for all $x_i$ in the first-pass returned list $X$ is obtained via score propagation on the graph:

$$R_g(x_i) = (1 - \alpha) R_p(x_i) + \alpha \sum_{x_j \in \mathcal{A}_i} \tilde{S}(x_j, x_i) R_g(x_j),$$  \hspace{1cm} (9)

where $R_p(x_i)$ is the initial score from PRF as in (8), $\alpha$ is an interpolation weight between 0 and 1, $\mathcal{A}_i$ the set of all utterances connected to utterance $x_i$, $\tilde{S}(x_j, x_i)$ the normalized edge weight $S(x_j, x_i)$ over the edges connected to the node for $x_j$ on the graph:

$$\tilde{S}(x_j, x_i) = \frac{S(x_j, x_i)}{\sum_{x_k \in \mathcal{A}_j} S(x_j, x_k)},$$  \hspace{1cm} (10)

where $\mathcal{A}_j$ is the set of utterances connected with $x_j$. In (9) the graph-based score $R_g(x_i)$ of an utterance $x_i$ depends on two factors interpolated by $\alpha$, the relevance score from PRF (the first term on the right side of (9)) and the score propagation over the graph based on the normalized edge weights $S(x_j, x_i)$ (the second term on the right side). The normalization in (10) normalizes $S(x_j, x_i)$ as a random walk problem on the graph, and the theory of random walk guarantees a set of unique solution of $R_g(x_i)$ can be found efficiently by the power method [7]. $R_g(x_i)$ is finally integrated with $R_p(x_i)$ for re-ranking as

$$R'(x_i) = R_g(x_i)(R_p(x_i))^{\delta_2},$$  \hspace{1cm} (11)

where $\delta_2$ is a parameter. The final retrieval results ranked according to $R'(x_i)$ in (11) are then displayed to the user.

3. Experimental Setup

The testing spoken archive is a corpus of 45 hours of recorded lectures for a course offered at National Taiwan University produced by a single instructor, which is quite noisy and spontaneous [8]. The lectures were given in Mandarin Chinese but with some terms and phrases produced in English and embedded within the Mandarin utterances. We split the corpus into two parts: 12 hours for acoustic and language model training and 33 hours for retrieval testing. Mean Average Precision (MAP) was used as the retrieval performance measure.

In order to evaluate the retrieval performance with respect to acoustic models of different matched conditions, we used three sets of acoustic models:

- Speaker-independent models (SI) trained on a Mandarin corpus of 24.6 hours of read speech, produced by 100 male and 100 female speakers, plus the Sinica L2 Taiwanese English corpus with 59.7 hours of English read speech, produced by 229 male and 256 female Taiwanese speakers.

- Speaker-adaptive models (SA) adapted by MLLR with 256 classes cascaded with the maximum a posterior estimation from the above SI model based on 500 utterances taken from the training set of the lecture corpus.

- Speaker-dependent models (SD) trained on the 12 hours training set of the lecture corpus.

For all the three versions of acoustic models, we trained a set of state-tied triphones spanned from 37 Mandarin monophones and 35 English monophones based on the recently developed state mapping and recovery techniques [9].

Two sets of experiments respectively with in-vocabulary (IV) and OOV queries were performed as below.

3.1. IV query experiments

The IV query set included 275 Chinese queries composed of 1 to 3 words, or 2 to 7 Chinese characters. In the experiments here, the language model was trained with the manual transcriptions of the training set of the lecture corpus. A close-to-oracle lexicon was used which included 11K Chinese words plus 2K English words covering all words in the testing archive. Each utterance was transcribed into a bilingual word lattice. The recognition accuracies were 49.7%, 80.8%, and 88.0% respectively for the SI, SA, and SD models. Then we transformed each Chinese word arc into a sequence of concatenated Chinese character or Mandarin syllable arcs to respectively form character or syllable lattices. Therefore, each utterance had three lattices, word-based, character-based and syllable-based.

3.2. OOV query experiments

110 English queries were used as the OOV query set, each consisting of a single word. For the experiments for OOV queries, we used a lexicon composed of 11K Chinese words and 10K English syllables for recognition. In this way none of the English queries were in the lexicon. A Chinese word-based trigram language model was trained with the training set of the lecture corpus, and we also trained an English syllable-based trigram language model with the 20,000 English documents of 20Newsroups\(^2\). These two language models were interpolated for producing the hybrid lattices composed of a mixture of Chinese word and English syllable arcs. The Chinese character accuracy was 79.0% for SD models, but the English syllable accuracy was only 41.5%. We substituted the Chinese word arcs in the hybrid-unit lattices with its corresponding Mandarin syllables to obtain a set of syllable-based lattices, and a set of phone-based lattices was generated by replacing each syllable arc in the syllable lattices by a sequence of phone arcs. The syllable- and phone-based lattices of each utterance were then used in the experiments for OOV queries.

We trained a 6-grams joint-sequence model from the CMU dictionary with 130K words as the grapheme-to-phoneme con-

\(^2\)http://people.csail.mit.edu/jrennie/20Newsroups/
Table 1: MAP results (%) of first pass, PRF, and cascade of PRF and graph-based re-ranking (PRF+Graph) for the IV experiments under different acoustic models (SI, SA, SD) with word- (column (1)), character- (column (2)) or syllable-based (column (3)) lattices, and the integrated results (column (4)).

<table>
<thead>
<tr>
<th>Approach</th>
<th>(1) Word</th>
<th>(2) Char</th>
<th>(3) Syl</th>
<th>(4) Integrated</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) SI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>first pass</td>
<td>56.18</td>
<td>44.47</td>
<td>43.17</td>
<td>55.98</td>
</tr>
<tr>
<td>PRF</td>
<td>60.57</td>
<td>55.69</td>
<td>51.64</td>
<td>61.73</td>
</tr>
<tr>
<td>PRF+Graph</td>
<td>63.81</td>
<td>63.15</td>
<td>57.56</td>
<td>65.54</td>
</tr>
<tr>
<td>(b) SA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>first pass</td>
<td>79.68</td>
<td>70.95</td>
<td>67.28</td>
<td>80.50</td>
</tr>
<tr>
<td>PRF</td>
<td>81.71</td>
<td>80.55</td>
<td>76.76</td>
<td>83.39</td>
</tr>
<tr>
<td>PRF+Graph</td>
<td>82.61</td>
<td>82.71</td>
<td>78.82</td>
<td>83.52</td>
</tr>
<tr>
<td>(c) SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>first pass</td>
<td>84.26</td>
<td>72.27</td>
<td>69.58</td>
<td>85.45</td>
</tr>
<tr>
<td>PRF</td>
<td>86.17</td>
<td>84.45</td>
<td>80.68</td>
<td>87.98</td>
</tr>
<tr>
<td>PRF+Graph</td>
<td>86.65</td>
<td>84.59</td>
<td>82.16</td>
<td>87.79</td>
</tr>
</tbody>
</table>

lower performance, but the results were reasonable if the acoustic models were good. We also found that PRF offered improvements in all cases (PRF vs first pass), and graph-based re-ranking further improved the results (PRF+graph vs PRF). Since phone offered limited and confusing information, the performance was poor when used alone. However, phone and syllable- and word-based versions contained complementary information, so the integration always helped.

5. Conclusion

In this paper, PRF and graph-based re-ranking method considering word/subword-based acoustic feature similarity are applied on open vocabulary retrieval of spoken content with shorter/longer queries. We found that these approaches can improve the performance of subword-based lattices, and thereby meliorate the results for both IV and OOV queries.

6. References


[3]The terms in the OOV queries were excluded from the CMU dictionary when training.