Lexicon Expansion Using Pronunciation Variations Extracted on the Basis of Speaker-related Deviation in Recognition Error Statistics

Yoshifumi Onishi
Common Platform Software Research Laboratories, NEC Corporation, Kawasaki, Japan
y-onishi@bp.jp.nec.com

Abstract

We propose a novel method for lexicon expansion using pronunciation variations extracted on the basis of speaker-related deviations in ASR error statistics. Two types of pronunciation variations were extracted: common pronunciation variations found with most speakers, and speaker-related pronunciation variations, identified on the basis of recognition error elements weighted by idf and tf-idf measures. Experimental results for CSJ show that entries added to the lexicon from speaker-related pronunciation variations were more effective than those generated on the basis of common pronunciation variations, some of which were superfluous.

Index Terms: speech recognition, pronunciation variation, tf-idf

1. Introduction

Pronunciation variation modeling has long been an issue in the field of automatic speech recognition (ASR) research [1], and recently, ASR targeting spontaneous speech has become the order of the day. Here, pronunciation variation is a particularly important problem since it is likely to be common and significant enough to degrade ASR.

As an approach to modeling pronunciation variation, expanding the canonical lexicon with data-derived pronunciation variants has been studied [1][2][3]. Specifically, pronunciation variations are first obtained from comparison between phonetic symbol sequences of standard and actual pronunciations in a training corpus. Statistically significant variations are then added to the canonical lexicon. ASR such as phone recognition is most commonly used to obtain phonetic symbol sequences of actual pronunciation [2], while in some cases, a manually transcribed corpus may be used [3]. While recognition error in ASR may be caused by a large number of different factors, among them, variation in pronunciation is a highly significant factor. Not all variations are statistically significant, however, and it is important to be able to extract those which are. Otherwise, superfluous entries into the lexicon would introduce new error. By way of contrast, while a manually transcribed corpus of actual pronunciation offers some number of “noise-free” variations, it is difficult for manually transcribers to catch many subtleties of pronunciation, and the number of such variations will be limited.

One important issue in pronunciation variation is speaker dependence. In [4], a speaker-dependent lexicon generation method is proposed, but it is based on the adaptation of a lexicon to a specific speaker’s pronunciation variation, and the problem of variation in pronunciation for a wide variety of speakers has not yet been significantly studied.

In this paper we propose a novel method for lexicon expansion using pronunciation variations extracted on the basis of speaker-related deviations in ASR error statistics. From a statistical comparison between phonetic symbol sequences of standard pronunciation and ASR results, as they are found in training corpuses for each individual speakers, the following two types of pronunciation variations were extracted: common pronunciation variations that occur with most speakers, and speaker-related pronunciation variation obtained by clustering on the basis of deviation among speakers. If the lexicon is extended regardless of speaker dependency, entries irrelevant to a given speaker will produce corruption. We try to avoid this problem by taking speaker-related pronunciation variations into consideration. By adding both common variations and speaker-related variations to a standard pronunciation lexicon, it is possible to create a lexicon that is more effective for application to new speakers.

2. Pronunciation variation extraction

In this section we consider recognition error elements weighted by two types of measures. The first is “term frequency-inverse document frequency (tf-idf),” which is an often used term-weighting scheme in information retrieval systems [5]. tf-idf is a measure of the importance of the concerned term in the concerned document (an information-theoretic perspective on tf-idf measures is presented in [6]). The second type is “inverse document frequency (idf),” which is a measure of the general importance of the concerned term. Recognition error elements are extracted from a comparison of phonetic symbol sequences of standard pronunciation with ASR results, as they are found in a training corpus for each individual speaker. In our implementation of idf and tf-idf, recognition error elements are considered to be terms and each individual-speaker corpus is considered to be a document. Speaker independent pronunciation variations can then be obtained by extracting error elements with low idf weights. Further, using clusters of vectors constructed of tf-idf-weighted error elements, speaker-related variations in pronunciation can be obtained by extracting error elements with high tf-idf weights from individual clusters.

2.1. Recognition error elements weighted by idf or tf-idf

Viterbi alignment is performed to obtain phonetic symbol sequences of standard pronunciations that are transcribed in a training corpus. ASR results are assumed to be sequences of actual pronunciation. In this study, we used the same acoustic models for both Viterbi alignment and ASR. We used a triphone-level-sequences based acoustic model. Obtained sequences were compared with ASR results frame by frame. Here we refer to any context-dependent phoneme-confusion pair obtained in this comparison as a recognition error element. We
2.2. Speaker-independent pronunciation variation

Recognition error elements with small idf weights indicate a commonality in pronunciation variation among speakers in the training corpora, and speaker-independent pronunciation variations, \( \{ X \}^{K}_{idf} \), can be extracted from all recognition error elements by choosing \( K \) elements with small idf weights.

2.3. Speaker-related pronunciation variation

Recognition error elements with large tf-idf weights indicate speaker-specific pronunciation variations. Using this property, we first consider the vectors for individual speakers, which are constructed of tf-idf weighted recognition error elements and are just like the vectors in vector space models in information retrieval systems. We then perform speaker clustering to deal with variations in pronunciation for a wide variety of speakers. Next, we extract speaker-related pronunciation variations from vector elements having large tf-idf weights on the representative points of individual clusters. In this study, we employ a hierarchical clustering method using cosine distance measures until a desired number of clusters is obtained, and a simple mean vector is used as a representative point for each cluster. Specifically, speaker-related variations are obtained, each element is corresponding to the rule 4), 5), 6), and 10) in Table 1 in sequence, a word with standard pronunciation “oNsee” is expanded to

\[
\begin{align*}
\text{oNsee} & \rightarrow \text{s} - \text{e} + \# \\
\text{e} - \text{e} + \# & \rightarrow \text{s} - \text{e} + \# \\
\text{e} - \text{e} + \# & \rightarrow \text{e} - \text{e} + \text{e} \\
\text{s} - \text{e} + \# & \rightarrow \text{s} - \text{u} + \# \\
\text{e} - \text{e} + \# & \rightarrow \text{e} - \text{N} + \#
\end{align*}
\]

and the language model (LM), \( P(v|w) \) is the probability of the variant given the words, and \( P(x|v) \) is the acoustic model. N-grams (\( N = 3 \)) are used for the LM, but for the sake of simplicity, we use a unigram model as \( P(v|w) \). Finally, we assign a uniform probabilistic value, \( u \), to each expanded variation, and the probabilistic value of each entry is normalized within a word (see Eq. (5)).

Let us consider an expansion example. Here, recognition error elements,

\[
\{ X \} = \{ e - e + \# \rightarrow s - e + \#; e - e + \# \rightarrow e - e + e; s - e + \# \rightarrow s - u + e; e - e + \# \rightarrow e - N + \# \}
\]

are obtained, each element is corresponding to the rule 4), 5), 6), and 10) in Table 1 in sequence, a word with standard pronunciation "oNsee" is expanded to

\[
\begin{align*}
\text{o N S e e} & \rightarrow \text{1/(1+4u)} \\
\text{o N S e} & \rightarrow \text{u/(1+4u)} \\
\text{o N S e e e} & \rightarrow \text{u/(1+4u)} \\
\text{o N S u e} & \rightarrow \text{u/(1+4u)} \\
\text{o N S e e N} & \rightarrow \text{u/(1+4u)}
\end{align*}
\]

where \( u/(1 + 4u) \) is the probabilistic value for each pronunciation within the word. As may be seen, four variations have been added.

For speaker-independent lexicon expansion, \( \{ X \}^{K}_{idf} \) is used. For speaker-related lexicon expansion, \( \{ X \}^{K,L}_{C_n} \) is used for individual clusters. Specifically, speaker-related variations are added to speaker-independent variations, as in Eq. (3). It will not be known which cluster lexicon should be used for any given speaker in the recognition process. We apply the ROVER method [8] to the recognition output of the extended lexicon for each individual cluster in order to select likely candidates.
Table 1: Rules of lexicon expansion.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Deletion of a consonant, X [A/N-X+B \rightarrow \star-A/N-B, A/N-B++] #-X+A \rightarrow #-A++ ]</td>
</tr>
<tr>
<td>2</td>
<td>Insertion of a consonant, X [\star-A/N+B, A/N-B++ \rightarrow A/N-X+B #-A++ \rightarrow #-X+A ]</td>
</tr>
<tr>
<td>3</td>
<td>Substitution of a consonant, X [A/N-X+B/N \rightarrow A/N-Y+B/N #-X+A/N \rightarrow #-Y+A/N ]</td>
</tr>
<tr>
<td>4</td>
<td>Deletion of a vowel, B [A/N+B% \rightarrow \star-A/N++, A/N-%++ ]</td>
</tr>
<tr>
<td>5</td>
<td>Insertion of a vowel, B [\star-A/N+%, A/N/X-%++ \rightarrow A/N/X-B% #-B% \rightarrow #-B++ #-A+# \rightarrow #-A++ ]</td>
</tr>
<tr>
<td>6</td>
<td>Substitution of a vowel, A [C/X+N-A% \rightarrow C/X-N-B% #-A+% \rightarrow #-B++ %+#-A+# \rightarrow %+B++ ]</td>
</tr>
<tr>
<td>7</td>
<td>Deletion of a double consonant, Q [A/N-Q+X \rightarrow \star-A/N+X, A/N-X++ #-Q+X \rightarrow #-X++ ]</td>
</tr>
<tr>
<td>8</td>
<td>Insertion of a double consonant, Q [\star-A/N+X, A/N-X++ \rightarrow A/N-Q+X #-X++ \rightarrow #-Q+X ]</td>
</tr>
<tr>
<td>9</td>
<td>Deletion of a syllabic nasal, N [A/N-N+% \rightarrow \star-A/N++, A/N-N% #-N+% \rightarrow #-N++ #A/N-N+# \rightarrow #-A/N++ ]</td>
</tr>
<tr>
<td>10</td>
<td>Insertion of a syllabic nasal, N [\star-A/N++, A/N-%++ \rightarrow A/N-N% \star-X+A, X-A++ \rightarrow X+N+A #-%++ \rightarrow #--N+% \star-A/N++ \rightarrow A/N-N++ ]</td>
</tr>
</tbody>
</table>

4. Experimental evaluation

4.1. Experimental conditions

We evaluated the proposed method on “the Corpus of Spontaneous Japanese (CSJ)” [7]. In the extraction of pronunciation variations, we used 1382 speakers and 2671 spontaneous speech items. Each speech item was a monologue of roughly 10 to 25 minutes. Although some speakers produced more than one speech item, we specified each item separately as a corpus of a speaker, D, i.e., M = 2671 in Eq. (1). Recognition error elements not conforming to the rules in Table 1 were discarded before speaker clustering to avoid data sparseness problem. Or the dimensionality of the vectors constructed of tf-idf-weighted error elements become too large.

For a base line system, we used a lexicon with standard pronunciation that employed 68k entries. Word accuracy (WA) was 67.0% for 10 test-set speeches. We used the same test-set throughout our evaluations.

4.2. Results

We first evaluated lexicons that has been extended using speaker-independent pronunciation variations. Table 2 shows WA with a lexicon extended by \( \{X\}_{C_{n}}^{K,L} \). WA was improved to 67.9% at \( K = 100 \) and \( u = 0.05 \), where the extended lexicon included 164k entries. When more variation entries were added (i.e. with increasing \( K \)), or with a larger probability of variation (i.e. an increasing \( u \)), WA was degraded.

We next evaluated the effect of the resulting speaker-related pronunciation variations. We set the number of clusters for extraction of speaker-related variations to 5. Table 3 shows the resulting number of speech items in each cluster. Here, four speech items have already been discarded because they composed isolated clusters containing one member only. For each cluster, the lexicon was extended by \( \{X\}_{C_{n}}^{K,L} \) and recognition was proceeded. And their outputs were integrated by the ROVER method. Evaluation results are shown in Table 4. WA was improved to 68.4% at \( (K,L) = (50,100) \) and \( u = 0.10 \), where the extended lexicons of individual clusters included from 170k to 196k entries.

5. Discussion

There have been proposed some speaker clustering based on acoustic features [9][10]. In such approach, there will be a difference in clustering between males and females to begin with. The relative balance between male and female clustering shown
Table 4: $WA(\%)$ with lexicon extended by $\{X\}_K^{K,L}$.

<table>
<thead>
<tr>
<th>$K,L$</th>
<th>$\alpha$</th>
<th>0.03</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
</tr>
</thead>
<tbody>
<tr>
<td>50, 50</td>
<td></td>
<td>67.9</td>
<td>68.0</td>
<td>68.2</td>
<td>68.2</td>
</tr>
<tr>
<td>50, 100</td>
<td></td>
<td>68.1</td>
<td>68.3</td>
<td>68.4</td>
<td>68.4</td>
</tr>
<tr>
<td>50, 150</td>
<td></td>
<td>68.0</td>
<td>68.2</td>
<td>68.2</td>
<td>68.2</td>
</tr>
<tr>
<td>100, 50</td>
<td></td>
<td>68.0</td>
<td>68.1</td>
<td>68.3</td>
<td>68.2</td>
</tr>
</tbody>
</table>

in Table 3 then confirms, at least, that our method is not trivial.

When we used a lexicon extended only with speaker-independent pronunciation variations, WA was maximized at $K = 100$ and $u = 0.05$ (see Table 2). That is, above these levels, superfluous entries were added to the lexicon. We have also confirmed that using a lexicon also extended with speaker-related pronunciation variations resulted in further improvement at $K + L = 150$ (see Table 4). At the fixed value of $u = 0.10$, in Table 2 $K = 100$ is the most effective value, while in Table 4 WA was improved at $(K, L) = (50, 100)$. This means that those entries added to the lexicon from the speaker-related pronunciation variations worked more effectively than those generated by $\{X\}_l^{R=50}$. This indicates the validity of the proposed method.

Here, the extended lexicon of each cluster is used with the ROVER method. Since there was a possibility that the improvement in Table 4 merely resulted from use of the ROVER method, we performed a verification experiment. We created pronunciation variations, $\{X\}_l^{R=100}$, 100 randomly selected variations, where $n = 1$ to 5, i.e. same number of clusters mentioned in Sec. 4.2. Combination with $\{X\}_l^{R=50}$ in the same manner as in Eq. (3) results in $\{X\}_l^{R=50}$. Lexicons extended by $\{X\}_l^{R=50}$ were used in conjunction with the ROVER method, and the results were evaluated. We conducted 5 tests with different random seeds, and the resulting average WA was 67.97%($\pm0.06$), where $u = 0.01$. This is better than WA=67.5%, at $K = 150$ in Table 2, however, significantly inferior to WA=68.4%, at $(K, L) = (50, 100)$ in Table 4. Thus the efficacy of the proposed method was confirmed.

6. Summary

In this paper we have proposed a novel method for lexicon expansion using pronunciation variations extracted on the basis of speaker-related deviations in ASR error statistics. Two types of pronunciation variations were extracted: common pronunciation variations found with most speakers, and speaker-related pronunciation variations, identified on the basis of recognition error elements weighted by idf and tf-idf measures. Experimental results for CSJ show that entries added to the lexicon from speaker-related pronunciation variations were more effective than those generated on the basis of common pronunciation variations, some of which were superfluous. In an investigation of the possible influence of use of the ROVER method, pronunciation variations added to the lexicon randomly were not effective, which indicates that the proposed method was effective.

7. Acknowledgements

The author would like to thank Prof. K. Shinoda, Mr. T. Koshinaka, and Mr. T. Emori for useful discussions and comments.

8. References