Multi-KeyWord Spotting of Telephone Speech Using Orthogonal Transform-Based SBR and RNN Prosodic Model

Wern-Jun Wang, Chun-Jen Lee, Eng-Fong Huang and Sin-Horng Chen

1 Department of Communication Engineering, National Chiao Tung University, Taiwan, R.O.C.
2 Advanced Technology Research Laboratory, Chunghwa Telecommunication Laboratories, Taiwan, R.O.C.

Abstract

In this paper, orthogonal transform-based signal bias removal (OTSBR) approach and RNN prosodic model are proposed for multi-keyword spotting of telephone speech. OTSBR is employed in the pre-processing stage of acoustic decoding and aimed at channel bias estimation to eliminate the acoustic mismatch between training and testing environments. The RNN prosodic model is adopted in the post-processing stage of the acoustic decoding to detect word boundaries for reordering the keyword candidates from the keyword spotter. Simulations on the real speech database collected from the Phone Directory Assistant Service developed in Chunghwa Telecommunication Laboratories (CTL-PDAS) were performed to evaluate the proposed methods. Experimental results showed that 71.0% of keyword detection rate and 81.8% of top 5 keywords inclusion rate can be attained by incorporating OTSBR and RNN prosodic model into the system.

1. Introduction

Recently, speech recognition has been applied in many applications such as voice command, voice dictation, information inquiry, voice portal, etc. However, considering the systems to be realized, the robustness of speech recognition for adverse environments is an important issue to overcome. Due to the acoustic mismatch between training and testing environments, the performance always degrades when most speech recognizers are applied to the telephone network. Background noise and channel bias are the two major interference factors that seriously degrade the performances of speech recognizers operating in adverse environments. In this paper, an orthogonal transform-based signal bias removal (OTSBR) method is proposed to achieve a better channel bias estimation to minimize the acoustic mismatch between training and testing environments.

In addition, the application of prosody in keyword spotting task is also discussed in this paper. Prosody is an inherent supra-segmental feature of human’s speech. It controls the stress, intonation pattern, and timing structure of continuous speech, which in turn decide the naturalness of the speech. Many approaches based on statistical methods have been proposed for detecting prosodic phrasal boundaries and word prominence [1][2]. Generally speaking, for the utterances with one or more keywords embedded in non-keyword speech, most people tend to emphasize the keywords than the non-keywords speech. The prominence levels of keywords are hence higher than the ones of non-keywords. Therefore, the keywords are clearer to hear and the related boundaries are easy to found. In this paper, an RNN-based prosodic modeling approach is proposed for multi-keyword spotting. It is performed in the post-processing stage of the acoustic decoding aiming at detecting word boundaries for assisting in the following keyword candidates reordering. An additional word boundary confidence test is expected to further increase the keyword detection rate.

Voice Portal is an on going project developed in Chunghwa Telecommunication Laboratories. One of the services provided by the portal is the Phone Directory Assistant Service (CTL-PDAS) [3]. This is a large vocabulary keyword spotting system using a HMM based keyword-filler network. In this system, 1252 employee names and 63 department names are selected as the primary keywords and the secondary keywords, respectively. In order to improve the system performance, OTSBR and RNN prosodic model have been incorporated with the multi-keyword spotter. The details of these two methods will be described in the following sections. The organization of the rest of this paper is as follows. Section 2 briefly introduces the overview of our proposed system. The descriptions of OTSBR and RNN prosodic model are described in Section 3 and 4, respectively. Section 5 will present the experimental results. Finally, some conclusions will be given in Section 6.

2. Overview of the proposed system

The task of CTL-PDAS system is to spot employee names and department names in CTL. It allows people to inquire the phone directories over the public switched telephone network (PSTN). The query sentences can consist of one primary keyword and one optional secondary keyword embedded in non-keyword speech. For example, a query sentence with primary keyword and secondary keyword and its response could be: “Would you please tell me the extension phone number of Wern-Jun Wang, thanks.” and “Please call Mr. Chun-Jen Lee in Advanced Technology Research Lab.”, where “Wern-Jun Wang” and “Chun-Jen Lee” are primary keywords and “Advanced Technology Research Lab” is the secondary keyword that we need to spot. The word-spotting task in this system is to detect a large vocabulary of 1252 employee names and 63 department names from the inquiries.

A functional block diagram of incorporating our proposed methods in CTL-PDAS system is shown in Fig. 1. The OTSBR is first applied for channel bias removal. The syllable
hypothesis can then be obtained from the HMM-based acoustic decoding. Based on the segmentation information provided by the base-syllable sequence, prosodic features are extracted for RNN prosodic model to detect the word-boundary information for current syllable. The detailed descriptions of word boundary output sequence as shown in Fig. 1 will be described in Section 4. Meanwhile, the generalized multi-keyword spotter with a modified finite state network (FSN) structure is used to generate the keyword hypotheses. The schematic diagram of the FSN used in this study is displayed in Fig. 2. The main characteristic of the FSN is that each node in the FSN can be either a keyword network or a filler network [3]. It can be explicitly integrated into the keyword spotter to governing which keywords can logically follow other keywords to form valid connections. In this approach, it is very suitable for simple grammar-constraint task, such as the CTL-PDAS one. Finally, the hypotheses reordering process is applied to reorder the keyword candidates based on the word boundary information. If the keyword positions of one candidate from the spotter conflict with the word boundary information from the prosodic model, it will be moved to the tail of the candidate list.

![Functional block diagram of incorporating the proposed methods in multi-keyword spotting system.](image)

Fig. 1. A functional block diagram of incorporating the proposed methods in multi-keyword spotting system.

![Schematic diagram of the keyword-filler network.](image)

Fig. 2. The schematic diagram of the keyword-filler network.

3. Orthogonal transform-based signal bias removal

Orthogonal transform technique has been widely used in waveform coding for data compression. It is employed to decompose an input data sequence into mutually orthogonal components in transform domain. The input data sequence can therefore be represented by a smooth curve formed by orthogonal expansion using some low order transform coefficients. Basis functions used in an orthogonal transform need to comply with the orthogonality property. In this study, the following four basis functions are used [4]:

\[
\Phi(i/N) = \begin{cases} 
1 & i = 1 \\
\frac{12 \times N}{(N + 2)} & i = 2 \\
\frac{180 \times N^3}{(N-1)(N+2)(N+3)} & i = 3 \\
\frac{2800}{(N-1)(N-2)(N+2)} \left(1 - \frac{3}{2} \left(\frac{i}{N}\right)ight) & i = 4 \\
\frac{2800}{(N-1)(N-2)(N+2)} \left(1 - \frac{3}{2} \left(\frac{i}{N}\right)ight) - \frac{2800}{10 \times N^2} & i = 5 \\
\frac{2800}{(N-1)(N-2)(N+2)} \left(1 - \frac{3}{2} \left(\frac{i}{N}\right)ight) - \frac{2800}{10 \times N^2} - \frac{2800}{20 \times N^2} & i = 6 \\
\frac{2800}{(N-1)(N-2)(N+2)} \left(1 - \frac{3}{2} \left(\frac{i}{N}\right)ight) - \frac{2800}{10 \times N^2} - \frac{2800}{20 \times N^2} & i = 7 \\
\vdots & \vdots \\
0 & 0 
\end{cases}
\]

for \(0 \leq i \leq N\), where \(N + 1\) is the length of the contour and \(N \geq 3\). These basis functions are, in fact, discrete Legendre polynomials. The contour of the \(k\)-th feature element, \(f_k(i)\), of a segment with length of \(N + 1\) frames can thus be approximated by

\[
f_k(i) = \frac{1}{N+1} \sum_{i=0}^{N} c_j(k) \times \Phi(i/N)
\]

for \(0 \leq i \leq N\), where

\[
c_j(k) = \frac{1}{N+1} \sum_{i=0}^{N} \Phi(i/N) \times f_k(i)
\]

is the \(j\)-th order orthogonal transform coefficient. It is noted that the zero-th order coefficient represents the mean of the contour, and the other three represent its shape.

According to the additive bias assumption in cepstral domain for SBR [5], the bias-corrupted feature \(f_k^b(i)\) can be modeled by

\[
f_k^b(i) = f_k(i) + b_k
\]

where \(b_k\) is a bias. The orthogonal transform coefficients \(c_j^b(k)\) of \(f_k^b(i)\) can then be expressed by

\[
c_j^b(k) = \frac{1}{N+1} \sum_{i=0}^{N} \Phi(i/N) \times f_k^b(i)
\]

From the characteristics of these four basis functions, it is easy to find that

\[
c_j^b(k) = \begin{cases} 
c_j(k) + b_k & \text{for } j = 0 \\
c_j(k) & \text{for } j \neq 0
\end{cases}
\]

From above analysis, the orthogonal transform coefficients of order greater than 0 of the bias-corrupted speech are the same as those of the clean speech. Such high order coefficients are
bias-free and therefore can be used to find out the optimum codeword without interfered by the corrupted bias. After finding out the best-matched codeword, the bias can then be obtained by subtracting the zero-th order component of the codeword from \( c^{k}_{0} \). The orthogonal transform operation is realized in a moving window process with consecutive windows being overlapped by several frames. In the training phase, all orthogonal transform coefficients of each feature element in the training set are collected and used to train a codebook via the vector quantization process. In the testing phase, the orthogonal transform coefficients of the testing utterance are calculated and compared with these pre-trained codebooks in the above-mentioned way to find the bias estimates. By subtracting the corresponding bias estimates from the features of every frame, we obtain the bias-removed speech features for recognition. It is worthy to note that the bias estimation process of the proposed method is non-iterative so that it is computationally efficient.

The above-mentioned OTSBR method has been conducted in the simulations of our previous study [6]. In the study, the speech database was constructed artificially by passing each utterance of the clean-speech testing set through a filter, which simulated a telephone channel. A set of 32 simulated filters generated from a large telephone-speech database was used in the study. Experimental results reveal that significant gains on the recognition accuracy and computation time were achieved as compared with the conventional SBR method. To further examine the validity of the proposed OTSBR method, it is worthy doing more tests on the telephone speech collected from the real environment. The task of CTL-PDAS can therefore provide the suitable database for testing the effectiveness of OTSBR.

4. The RNN prosodic model

The proposed prosodic model used in this paper has shown its effectiveness on speech-to-text conversion task [7]. It used an RNN to detect information related to word boundaries from the input prosodic features extracted from the vicinity of the current syllable. The RNN is a three-layer network with all outputs of the hidden layer being fed back as additional inputs. It is a dynamic system suitable for realizing a complex mapping between the context of the current input prosodic features and the output word-boundary information. The RNN can be trained using a large speech database by the back propagation through time algorithm. Input prosodic features are extracted from each training utterance with all syllable boundaries being given by the preceding acoustic decoding. Output targets are word-boundary information extracted from the word sequence of the associated text tagged by a statistical model-based method. After properly training, the RNN can then be used in the testing phase to generate word-boundary information using the prosodic features extracted from the testing utterance.

Different from our previous study on speech-to-text conversion [7], the prosodic features adopted in this study have been simplified to seven acoustic features extracted from the vicinity of the current syllable. They include normalized syllable final duration, syllable pitch mean and its differences with the two nearest neighboring syllables, and syllable log-energy maximum and its differences with the two nearest neighboring syllables. The output linguistic features are 4 flags indicating whether the current syllable is a monosyllabic word or is the beginning syllable, the intermediate syllable, or the ending syllable of a polysyllabic word. These 4 output features are denoted as MSW, BPSW, IPSW, and EPSW, respectively. To prepare output targets for training the RNN, a large corpus with all texts are tagged into word sequences in advance is necessary. An automatic statistical model-based tagging algorithm based on the long-word-first criterion is first applied. Then, several simple word-merging rules are used to modify the resulting word sequences. Lastly, tagging errors are corrected manually. All output targets to train the RNN are extracted from these well-tagged word sequences.

A finite state machine (FSM) with some thresholds determined empirically is used to examine whether the responses of the RNN are good enough to make reliable classifications for the word-tag set [7]. The topology of the FSM is shown in Fig. 3, where M, B, I, E and U stand for the MSW, BPSW, IPSW, EPSW and uncertain state, respectively. Note that if the output responses of processing syllables are not reliable enough, they will remain in uncertain states. It is also noted that the prosodic model trained from a single speaker database [7] is used in this study owing to there are not database with prosodic features for sufficient speakers and well-tagged word sequences available for our needs. However, in the testing phase of this study, we have considered to adapt the prosodic model by the pitch mean and log-energy maximum for each speaker. Therefore, the utterance-level syllable pitch mean and syllable log-energy maximum must be calculated in advance for each testing utterance.

5. Experimental results

5.1. Database description

Effectiveness of the proposed method was examined by simulations on CTL-PDAS. Two separate databases were set up for training and testing. The first speech database (SDB1), used for training, consists of 14,400 utterances uttered by 120 speakers. The second speech database (SDB2), used for testing, consists of 500 spontaneous inquiries for the CTL-PDAS collected from 7 male speakers. All the utterances were collected directly from the inquiry calls through the PSTN. For the reason that each utterance may contain more than one keyword, the spotting accuracy is defined as the number of correctly recognized keywords relative to the total keywords. Three types of utterances are considered in the
simulation. They are a single keyword (WORD), a single keyword embedded in non-keyword speech (S_KW), and two keywords embedded in non-keyword speech (M_KW).

A set of 25 features, including 12 MFCCs, 12 delta MFCCs, and a delta log-energy was extracted for each frame from SDB1. A sub-syllable-based HMM recognizer was constructed by the maximum likelihood training algorithm. The OTSBR method used three separate codebooks also trained from SDB1 for the three feature sets containing 12 MFCCs, 12 delta MFCCs, and a delta log-energy, respectively. The orthogonal transform coefficients of these 25 features were calculated for all utterances in the training set and used to create 25 codebooks. The size of all the codebooks is set to be 256.

As described in Section 4, a single speaker database is used for training the prosodic model. All speech signals of this database were manually segmented into syllable sequences. The pitch contour was detected by the SIFT algorithm and corrected by hand. 7 prosodic features mentioned previously were extracted for each syllable. All texts were also tagged into word sequences. Four output linguistic features were then extracted for each character. An RNN taking seven prosodic features of the current syllable as inputs was then trained using the BPTT algorithm. The number of nodes in the hidden layers was determined empirically and set to be 30.

Table 1: The performance comparisons based on different approaches for different data types (unit: %).

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Baseline</th>
<th>With OTSBR</th>
<th>With prosodic model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top 1</td>
<td>Top 5</td>
<td>Top 1</td>
</tr>
<tr>
<td>WORD</td>
<td>87.7</td>
<td>93.8</td>
<td>90.6</td>
</tr>
<tr>
<td>S_KW</td>
<td>45.0</td>
<td>63.8</td>
<td>48.9</td>
</tr>
<tr>
<td>M_KW</td>
<td>73.9</td>
<td>80.9</td>
<td>78.3</td>
</tr>
<tr>
<td>Average</td>
<td>67.1</td>
<td>78.0</td>
<td>70.2</td>
</tr>
</tbody>
</table>

5.2. Results of multi-keyword spotter

Word spotting test results are given in Table 1. The first major column shows the results by baseline method [3]. The results by incorporating the OTSBR into the system are shown in the second major column. The results obtained by further incorporation of the prosodic model are demonstrated in the third major column. The keyword detection rates (top 1) and top 5 keyword inclusion rates shown in this table are calculated by considering top 10 candidates provide by the multi-keyword spotter. From this table, it is noted that the WORD type inquiries get the highest keyword detection rate and the rates of the M_KW type of utterances are higher than that of S_KW ones. The reason is due to the support from the constraints exist between the primary keywords and the secondary keywords. In addition, it is clear to find that the performance improvement can be achieved with the help both from OTSBR and prosodic model. The improvement afforded by OTSBR also confirms our previous study on the speech database with simulated channel bias [6]. The reasons for no gain has been found for the WORD type inquiries are that there are usually only one keyword in such utterances. Prosodic model cannot be helpful for such utterances.

6. Conclusion

A new method for robust multi-keyword spotting has been proposed in this paper. The OTSBR is applied to remove the effect of mismatch between testing utterance and the trained models in telephone environment. Besides, according to the keyword position predicted by the RNN prosodic model, the keyword hypotheses generated from the multi-keyword spotter can be verified and reordered. Therefore, the keyword detection rate can be effectively increased by including fewer candidates. Experimental results show that OTSBR together with the RNN prosodic model can achieve an encouraging improvement. The future work will be how to efficiently combine the prosodic model with the search processes of keyword spotter to further improve the performance.

7. References