Speaker adaptation of a code book of vector quantization

PACS number: 43.72.Ne

Y. Niimi and Y. Kobayashi
Kyoto Institute of Technology, Matsugasaki, Sakyo-ku, Kyoto 606
(Received 16 November 1987)

1. Introduction

This paper describes a new method for the speaker adaptation of a code book of vector quantization and its application to an isolated word recognition system based on the dynamic time-warping (DTW) algorithm.1)

A goal of speaker-adaptive systems is to adapt word templates to a new speaker using a small set of his training utterances. In several attempts of speaker adaptation a word is described as a string of smaller linguistic units than a word—such as demisyllables2) and phonemes—and a word template for a new speaker is created by concatenation of such smaller templates extracted from his training utterances. In another approach a speaker-adapted template is obtained by modification of a speaker-independent template with speaker-specific feature vectors.3) The speaker adaptation through vector quantization has been studied in which the correspondence of code vectors of a reference code book to those of a code book of a new speaker is made using his training utterances.4)

This study is based on an assumption that a speech spectrum of a speaker could be modeled by a two-factor model in the multivariate statistical analysis. In other words, it is represented by a sum of two main effects of 'phoneme' and 'speaker' and the interaction between the two. However, in a narrow subspace, such as a neighborhood of a steady vowel, the interaction term can be considered constant and therefore omitted from the model. If the effect of 'speaker' in such a subspace can be estimated from training utterances of a new speaker, the speaker-dependent effect of 'phoneme' can be estimated. This study applies the above idea to the speaker adaptation of a VQ code book.

2. An Overview of the Word Recognition System

Figure 1 depicts an overview of the recognition system of isolated words which incorporates the proposed procedure for the speaker adaptation. An incoming signal is pre-emphasized, low-pass filtered and digitized at 10 kHz with an accuracy of 12 bits.
Then the energy and zero crossing rate of the signal are calculated every 10 ms. Following the automatic end-point detection using these parameters, the interval of the speech signal is subjected to the 14-th order LPC-analysis with Hamming window of 25.6 ms, producing 20 cepstral coefficients every 10 ms.

The system illustrated in Fig. 1 works in three phases. The first (preparation) phase prepares a common code book and a speaker-independent word template for each word class. Utterances of several speakers are analyzed and averaged to produce an averaged word pattern for each word class. The common code book is designed with the spectra of selected frames of these averaged word patterns and then used to vector-quantize the averaged word patterns, yielding the speaker-independent word templates. These templates are expressed with a sequence of the code symbols, each uniquely assigned to a code vector.

The second (adaptation) phase adapts the common code book to a new speaker using a small set of training data. The procedure will be described in the next section. The adaptation of word templates is indirectly made in the sense that without changing the symbolic notation of the word templates, the spectral content of each code symbol is adapted.

The third (recognition) phase recognizes unknown words of the speaker to whom the system has been tuned. An incoming speech signal is LPC-analyzed and vector-quantized with the speaker-adapted code book. The DTW component matches the input utterance against the word templates using the speaker-adapted distance matrix and chooses one of possible word classes based on the minimum distance criterion.

3. Speaker Adaptation of a Code Book

The idea underlying the method of the speaker-adaptation used in this study is as follows. A speech spectrum $f(p,s)$ produced when a speaker $s$ utters a phoneme $p$ is modeled as

$$f(p,s)=f(p)+f(s)+f(p*s)+\text{residual},$$

where $f(p)$, $f(s)$ and $f(p*s)$ are spectra specific to the phoneme $p$, the speaker $s$ and their interaction, respectively. In a narrow subspace in which the term $f(p*s)$ is considered constant, the above formula can be written as

$$f(p,s)=f(p)+f_{sn}(s)+\text{residual},$$

where $f_{sn}(s)$ is the spectrum specific to the speaker $s$ in the subspace. We consider as such a subspace the neighborhood of the representative spectral vector of some phone with stationary parts, and assume that a shift of $f_{sn}(s)$ due to a speaker be proportional to a shift of the centroid vector in the subspace. We have selected seven phones (five vowels, a nasal group and a fricative group) to determine the centroids of the subspaces.

In the preparation phase, a set of monosyllable utterances including the seven phones stated above is also collected from each speaker who has provided the designing word utterances. The centroids $p_j$ ($j=1,...,7$) corresponding to the seven phones are computed from these speech data.

In the adaptation phase, a new speaker utters the same set of monosyllables in isolation to provide training samples from which the seven new centroids $q_j$ ($j=1,...,7$) are calculated. Let $c_i$ and $s_i$ ($i=1,...,L$) denote the code vectors in the common code book and the assigned code symbols, respectively. A speaker-adapted code vector $c_i'$ for a code symbol $s_i$ is computed by the following formula:

$$c_i'=c_i+\sum_{j=1}^{7} w_{ij}(q_j-p_j)$$

where $wij$ is a weighting factor inverse to a distance between $c_i$ and $p_j$, determined as follows:

$$wij=1/[D+d(c_i,p_j)]$$

where $D=\sum_{k=1}^{7} 1/d(c_i,p_k)$.

$d(x,y)$ is Euclidean distance between two vectors $x$ and $y$. The weighting is made to smoothly change code vectors across a boundary of two subspaces. Thus a physical correlate of a word template with $s_1s_2...s_n$ as its symbolic notation changes from the vector sequence $c_1c_2...c_n$ to $c_1c_2'...c_n'$. A new distance matrix whose $(k,l)$ element is $d(c'_k,c_l')$ is also calculated in this phase.

4. Recognition Experiments

The proposed method for the speaker adaptation was evaluated through the recognition of the vocabulary of 52 Japanese city names. In order to provide training and test utterances, twenty male speakers (test speakers) spoke the words in the vocabulary twice and the monosyllables once in a sound booth. Thus we have two sets of the word utterances for each speaker. The first set of spoken words and the monosyllables were used to train (or to adapt) the word recognition system and the second set to test it. In addition, separate five speakers spoke the words in the vocabulary and the monosyllables once, which were used to design the common code book and word templates, and to determine the seven subspaces respectively.

We conducted three word recognition experiments: speaker-adaptive, speaker-independent and speaker-dependent ones. The last two were performed for comparison. In the first experiment the speaker-adapted code book was made for each of the test speakers according to the procedure stated in the previous section: the monosyllable utterances spoken by the speaker were used to determine the centroids of the seven stationary phones, with which the code vectors of the common code book were modified to produce the speaker-adapted code book. The word utterances in the test data set were then coded with this code book.

In the second experiment, the speaker-independent code book and word templates designed in the preparation phase were used for all the test speakers.
Table 1 Word recognition rates (%).

<table>
<thead>
<tr>
<th>Code book size</th>
<th>128</th>
<th>256</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker-adaptive</td>
<td>98.8</td>
<td>98.9</td>
</tr>
<tr>
<td>Speaker-independent</td>
<td>98.5</td>
<td>98.6</td>
</tr>
<tr>
<td>Speaker-dependent</td>
<td>99.2</td>
<td>99.2</td>
</tr>
</tbody>
</table>

In the third experiment, the speaker-dependent code book and word templates were newly designed for each of the test speakers from his first set of word utterances. Then the code book was used to code the word utterances in his test data set.

Through these three experiments, the size of the code book was selected to be 128 and 256. Table 1 shows the results of the experiments. For the code book size of 256, the word recognition rate was 98.9% in the speaker-adaptive experiment while it was 98.6% and 99.2% in the speaker-independent and speaker-dependent ones respectively. We tested statistically the hypothesis that there be no difference between the recognition rate in the speaker-adaptive experiment and that in the speaker-independent experiment. Under the assumption that the number of word classes improved in the speaker-adaptive experiment for each speaker be distributed approximately as normal, the $F$-test showed that the significance level to reject this hypothesis was 0.14. The results were almost the same for the code book size of 128. This concludes that the proposed method is effective for the speaker adaptation.

5. Conclusion

The method for the speaker adaptation of a code book of vector quantization has been described and evaluated through the application to the recognition of the vocabulary of 52 Japanese city names. The experiments performed based on this method have shown that it is effective for the speaker adaptation. Several factors are, however, left unexamined in detail: the number of the subspaces, where to locate the subspaces and how to modify code vectors taking account of coarticulation effects. These should be investigated in the future work.

References