A STUDY OF SPEAKER ADAPTATION FOR SPEAKER INDEPENDENT SPEECH RECOGNITION METHOD USING PHONEME SIMILARITY VECTOR

Yoshio ONO¹, Maki YAMADA¹, Masakatsu HOSHIMI²

¹Matsushita Research Institute Tokyo, Inc.
3-10-1 Higashimita, Tama-ku, Kawasaki, 214-8501, JAPAN
²Tohoku University
04 Aramaki Aza Aoba, Aoba-ku, Sendai, 980-8579, JAPAN

ABSTRACT

In this paper we introduce an effective speaker adaptation technique to our unique compact speech recognizer especially designed for consumer electronics products. The compact ASR method we have developed in our previous work employs phoneme similarities as feature parameters, which are extracted temporally successive matching between speech sample and 24 context-independent phoneme standard templates in forms of time-spectral pattern. With these unique feature parameters, the recognizer has been successfully embedded in a single DSP without any external memory chips. Speaker adaptation method we propose in this paper focuses on adaptation of the phoneme standard templates in the ASR method. When evaluated with 20 or less training words in 100 isolated words recognition test, the proposed method achieved reduction of error rate by 20 % approximately. The test results also show its effect for child subjects and validity of applying this technology in the recognizer which is used by various users in consumer electronics products.

1. INTRODUCTION

Recently portable information devices such as PDA are rapidly spreading and speech recognition is expected as a useful interfaces for such small and portable consumer electronics products. To meet the expectation, a reasonable cost and high performance for unspecified speakers are important for a speech recognizer.

Especially when it comes to a small portable device which works with a single DSP, hardware resources available for speech recognition process are quite limited and it’s difficult to embed traditional ASR algorithm into it. Therefore, in our previous work, we’ve developed a speaker-independent speech recognition method (MSM: Model Speech Method) that works without extra hardware resources other than a single DSP.[1][2]

In addition to a reasonable cost, high accuracy for unspecified speakers is another indispensable feature for a speech recognizer which is designed for consumer electronics application. Speaker adaptation is a popular solution to achieve the feature.[3] But traditional adaptation methods often require a lot of training data or large hardware resources and they are not suitable to a compact recognizer such as MSM.

In this paper, first we briefly describe outline of MSM, our speech recognition method which employs phoneme similarity as a feature parameter. Secondary, we propose a new speaker adaptation which is designed to bring great improvement to MSM with a small amount of training speech and report experimental results to verify its validity.

2. FEATURES OF OUR COMPACT SPEECH RECOGNITION METHOD

Our previous work proved that phoneme similarity vector, which we employed as a feature parameter for our recognition method MSM, has smaller individual difference than cepstrum coefficients. We also verified that phoneme similarity vector keeps high recognition accuracy even if it’s represented in low precision.

Hereafter we summarize our unique speech recognition method MSM using phoneme similarity vector.

Fig.1 shows recognition processes in our method. Speech signal is acquired at 10kHz sampling frequency and analyzed into 10th order of LPC cepstrum coefficients together with normalized residual energy and time derivatives of power energy in each 10msec frame interval. Then time-spectral pattern is segmented from the analyzed acoustic parameters, and phoneme similarity based on linear discriminant function is obtained by matching the segmented time-spectral pattern of acoustic parameters of test speech and phoneme standard templates which are trained in advance.
Here each phoneme template consists of a time-spectral pattern of acoustic parameters including cepstrum coefficients over successive 10 frames, and total dimension of the time-spectral pattern is fixed as 120 for all phoneme categories. Corresponding to our definition of Japanese phoneme (5 vowels and 19 consonants), unit number of phoneme templates is 24.

The similarity for a phoneme p, \( L_p \), is obtained as Eq. (1)

\[
L_p = a_p \cdot c - b_p
\]

where

\[
a_p = 2 \sum_{i=1}^{120} \mu_p^i,
\]

\[
b_p = \mu_p \cdot \sum_{i=1}^{120} \mu_p^i
\]

Here \( c \) presents segmented pattern of test speech, \( \mu_p \) is a mean vector of the standard pattern of phoneme p, \( \sum \) is a covariance matrix which is common for all the phoneme categories. These phoneme standard templates are context-independent ones and they are trained from a lot of training subjects’ data so that it can be used for speaker-independent recognition. 24 phoneme similarities and their time-derivatives, which are obtained at every frame by matching test speech and phoneme standard templates, compose feature parameters for word recognition. Based on Japanese syllable “KANA” representation, word standard templates are composed by concatenating phoneme similarity pattern of CV/VC sub-word units.

The phoneme similarity vector of i-th frame, \( d_i \), is presented as Eq.(2). Similarly, time derivatives of phoneme similarity vector are presented as Eq.(3).

\[
d_i = (L_1, L_2, \ldots, L_{24})
\]

\[
\Delta d_i = (\Delta L_1, \Delta L_2, \ldots, \Delta L_{24})
\]

where \( L_j \) is similarity of phoneme j at the i-th frame and \( \Delta L_j \) its time derivative.

DTW is conducted in time-alignment of word matching. Partial score \( s(i,j) \) is calculated by correlation cosine distance as Eq.(4), where \( i \text{th} \) i-th frame similarity vector of test speech, \( j \text{th} \) j-th frame similarity pattern of a word model, and \( \Delta \epsilon_j \) time derivatives.

\[
w = \text{a weighting ratio and fixed as 0.5 in this study.}
\]

Correlation cosine is an inner product of normalized similarity vector.

\[
s(i,j) = w \cdot \frac{d_i \cdot \epsilon_j}{\|d_i\| \cdot \|\epsilon_j\|} + (1 - w) \frac{\Delta d_i \cdot \Delta \epsilon_j}{\|\Delta d_i\| \cdot \|\Delta \epsilon_j\|}
\]

Consequently, phoneme similarities and their time derivatives are feature parameter in word matching and a word model is presented as trajectories of phoneme similarities which are phonologically related to the word.

For example, Fig.2 shows similarities of several phonemes in a standard template of a Japanese word “ZAMA” which was made by concatenating CV/VC sub-word units. In the figure, x-axis shows time (frame) direction, y-axis phoneme category, and z-axis phoneme similarity amplitude, and remarkably high value of phoneme similarities is shown in gray.

In the beginning of the word, similarity of a consonant /z/ rises and then that of a vowel /a/ follows. Then those of nasals /N/ and /m/ rise and that of /a/ follows again in the end of the word utterance.

![Phoneme Similarity L](image)

Fig.2: Phoneme similarities in a word standard template “ZAMA”

Obviously phoneme similarity values of other phonemes are quite small and they can be omitted in presentation of the standard word template.

Therefor, dimension of each sub-word unit, which are originally 48 in total (24 phoneme similarities and 24 time-derivatives of them) can be reduced greatly without any adverse effect in recognition accuracy. Furthermore, due to its nature, phoneme similarity doesn’t request high precision in its presentation. These advantages in terms of compact memory size and small dimensionality enables us to implement this speaker-independent speech recognition method MSM in a single DSP chip without any external RAM and ROM.

### 3. SPEAKER ADAPTATION METHOD

#### 3.1 Adaptation of Phoneme Standard Templates

Though our speech recognition method MSM is designed as a speaker-independent recognizer, it’s still true that some speakers have worse recognition accuracy than others due to their speech variety. One of major speaker variation of speech signal exists in its spectral features, which are closely affected by their variation of articulatory organs in vocal tract.

In conventional ASR methods, spectral parameters such as cepstrum coefficients are widely used to present sub-word models. Generally sub-word models have many units. For example, our definition of CV/VC has 536 units. Therefore, once the spectral features are affected by a speaker’s characteristics, a lot of training data is necessary for speaker adaptation to train all the sub-word categories.

On the other hand, our unique recognition method MSM employs phoneme similarity as feature parameters to present sub-word templates. This parameter is proved to have less speaker difference than spectral parameters such as cepstrum coefficients. The cepstrum coefficients are used to compose the context-independent phoneme standard templates. That means we only need to adapt the phoneme standard templates and we do not train that the sub-word templates at all. Furthermore, the fact that phoneme standard templates have just 24 units, which is much smaller than 536, the number of sub-word templates, reminds us that effective adaptation must be possible with less training speech data.

Based on above idea, we propose a speaker adaptation method for MSM with following four steps.

...
1. Search frames (Fref) in a word standard template where phoneme similarity has a value larger than a threshold for one or more phonemes. For example, in Fig.2, gray remarked frames should be searched as Fref.

2. By DP time-alignment between the word template and a training data sample, search training sample frames (Ftrain) which correspond to the frames Fref which are selected in step 1.

3. Extract time-spectral pattern of spectral parameters (TSPtrain) from training sample near Ftrain.

4. Repeating step 1 to step 3 for all the training samples, average TSPtrain for each phoneme category respectively to acquire speaker-dependent phoneme template.

5. Finally, according to Eq.(5), produce a speaker-adapted phoneme standard template \( B_{\text{adapt}} \) from an original phoneme standard template and the speaker-dependent phoneme template \( B_{\max} \) with a proper weighting ratio.

\[
\mu_{\text{adapt}} = \alpha \cdot \mu_{\text{train}} + (1 - \alpha) \cdot \mu_{\text{orig}} \tag{5}
\]

### 3.2 Observation of Phoneme Similarity with Speaker-Adapted Phoneme Standard Templates

Fig.3 and Fig.4 show effect of the proposed speaker adaptation method in terms of phoneme similarity trajectory. Fig.3 is an original phoneme similarity trajectory of a sample speech “ZAMA” by a speaker which is generated by original phoneme standard templates. This subject’s spectral features are far different from the standard ones and large similarity of vowel /o/ is observed instead of that of /a/.

Fig.4 is phoneme similarities of the same speech data with adapted phoneme standard templates. In Fig.4, phoneme similarities of vowel /a/, nasals /m/ and fricative /z/ compose a trajectory fairly close to that of the word standard template in Fig.2. This observation conceptually proves the validity of the proposed adaptation method.

### 4. Experimental Results of Recognition Tests

To prove validity of the proposed speaker adaptation method, word recognition tests were conducted. Table 1 shows specifications of acoustic analysis, time-spectral pattern and feature parameter in word matching. Table 2 shows Japanese phoneme category definition in our experiments. The original 24 phoneme standard templates were trained from about 200 word set spoken by 40 speakers. Total 536 units of sub-word patterns are trained from about 500 word set spoken by 16 speakers.

Two test sets were evaluated. One is adult subjects’ speech and the other is child subjects’ one. All the training was carried out under a supervised condition.

---

**Table 1: Specifications of Feature Parameter Extraction**

<table>
<thead>
<tr>
<th>Acoustic Analysis</th>
<th>Phoneme Standard Template (Time-Spectral Pattern)</th>
<th>Feature Parameter in Word Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling Frequency</td>
<td>10kHz</td>
<td>Spectral Dimension – 12</td>
</tr>
<tr>
<td>Frame Length</td>
<td>10ms</td>
<td>Cepstrum Coefficients;C1 C10</td>
</tr>
<tr>
<td>Analysis Window Length</td>
<td>20ms</td>
<td>Normalized Residual Energy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Time Derivatives of Power Energy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Time Dimension – 10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Successive 10 Frames</td>
</tr>
<tr>
<td></td>
<td></td>
<td>near Phoneme Discriminative Frame</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total 120 dimensions</td>
</tr>
</tbody>
</table>

**Table 2: 24 Japanese Phoneme Categories**

<table>
<thead>
<tr>
<th>Vowel</th>
<th>{ a, i, u, e, o }</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consonant</td>
<td>Fricative{z,s,hv,hu} Plosive{p,t,k,c,b,d} Glide/Semivowel{w,y,yu,r} Nasal{m,n,ng, N}</td>
</tr>
</tbody>
</table>

---

### 4.1 Experimental Results on Adult Subjects

Firstly adult subjects’ test set was evaluated. Total 50 subjects spoke 100 Japanese city names. When all the 24 phonemes were adapted with 20 training word samples, nearly 18 % reduction in error rate was attained on the average over 50 subjects as Table 3 shows.

When looking into the worst 10 subjects’ improvement, error reduction rate is 34 % approximately. This means that the proposed method is effective especially for speakers with worse recognition performance with the original phoneme standard templates.

To estimate how much training data is necessary, 20 or less word training words were also tested. Fig 5 shows the proposed adaptation can work with 10 or less training words. To see which phoneme category contributes most to accuracy improvement in speaker adaptation, different three cases were compared as follows.

**Case I** Only 19 consonants were adapted.
Case 2: Only 5 vowels were adapted.

Case 3: Both of 19 consonants and 5 vowels were adapted.

Fig. 6 shows that most improvement was brought by adaptation of five vowels.

4.2 Experimental Results on Child Subjects

Secondary child subjects’ test set was evaluated, where original phoneme standard templates were trained with adult speech. Total 19 child subjects spoke 95 Japanese command words. When all the 24 phonemes were adapted with 20 training word samples, error rate reduction was 53% approximately. This error rate reduction is similar to that of adult speakers who were categorized as worse subjects with original phoneme standard templates.

### Table 3: Improvement of Adult Subject Test Set

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Phoneme Adaptation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition</td>
<td>96.82%</td>
<td>97.40%</td>
</tr>
</tbody>
</table>

Fig. 5: Recognition Rate versus Training Word Number

The result of child subject test set is also encouraging to apply this method to our recognizer which will be used by various users including children. By employing the adaptation method, we can easily apply our speech recognizer to child users in future.

4.3 Discussions on Experimental Results and Future Work

The average error rate reduction by 18% in adult subject test shows that the proposed adaptation method can achieve a reasonable improvement. From a viewpoint of ASR application in consumer electronics products in which a certain level of accuracy must be guaranteed for every user, error rate reduction by 34% in the worst 10 adult subjects is much more preferable result.

Test results on amount of training data also shows that the proposed method is preferable in consumer electronics application where a demanding task of so many training words’ utterance cannot be expected for users. It shows that the proposed method can still maintain a high improvement with 10 or less training words.

The comparison test of phoneme category shows reasonable result that vowel adaptation is much more important in phoneme adaptation than consonant adaptation. Since individual variation of speakers’ articulatory organs in vocal tract primarily affects spectral feature parameters which characterize vowel patterns, this result sounds quite reasonable. The fact that significant effect can be generated with adaptation of only five vowels, not all the 24 phonemes, will be advantages in practical application in terms of training speech amount and hardware resources necessary for implementation of the adaptation algorithm.

The result of child subject test set is also encouraging to apply this method to our recognizer which will be used by various users including children. By employing the adaptation method, we can easily apply our speech recognizer to child users in future.

As future work, we will evaluate the proposed speaker adaptation method under unsupervised conditions. Furthermore, in order to attain further improvement by speaker adaptation, we need to consider speakers’ individuality which speaking speed difference causes such as variation in duration time and co-articulation.

5. CONCLUSION

In this paper we proposed an effective speaker adaptation technique to our unique speech recognition method MSM especially designed for consumer electronics products. When evaluated with 20 or less training words in 100 isolated words recognition test, the proposed adaptation method achieved reduction of average error rate by 20% approximately. The test results were quite promising in terms of requirements of ASR application in consumer electronics products, and we believe that the method will be useful to realize a compact recognizer which can be adapted to each user with a small mount of training data.

6. REFERENCES