PHONEME RECOGNITION EXPERT SYSTEM USING SPECTROGRAM READING KNOWLEDGE AND NEURAL NETWORKS

Yasuhiro KOMORI, Kaichiro HATAZAKI†, Takaharu TANAKA, Takeshi KAWABATA and Kiyohiro SHIKANO

ATR Interpreting Telephony Research Laboratories
Seika-cho, Soraku-gun, Kyoto JAPAN

ABSTRACT

We present a method for phoneme recognition using an expert system combining spectrogram reading knowledge and neural networks, and report its performance. The proposed expert system consists of two parts: (1) phoneme segmentation based on spectrogram reading knowledge used by human experts, and (2) phoneme identification using neural networks applied to the phoneme boundaries determined in phoneme segmentation. Highly accurate phoneme segmentation can be achieved by using human-like contextual spectrogram reading knowledge. Moreover, high performance phoneme identification can be achieved by applying neural networks to the accurate phoneme segmentation result. The system was tested on Japanese consonants, with 90.8% of the phonemes correctly segmented and 92.4% of the phonemes correctly identified within the correct segment. 83.9% of the phonemes were correctly recognized both in segmentation and identification.

1. INTRODUCTION

Since Zue and his colleagues showed that human experts are able to determine phoneme category from a spectrogram with high accuracy, using their spectrogram reading knowledge [1], several speech recognition systems based on spectrogram reading knowledge have been developed [2][3][4][5][6]. This research has proved the effectiveness of using human expert knowledge for phoneme identification, based on acoustic features manually pre-extracted or simple features easily extracted automatically. However, it is very difficult to fully formalize phoneme identification knowledge, and to automatically extract acoustic features for phoneme identification.

Previously, we have focused on utilizing such spectrogram reading knowledge for phoneme segmentation rather than phoneme identification, and showed the effectiveness of a phoneme segmentation expert system[7].

Neural networks, on the other hand, have an incredibly high performance in identifying pre-segmented phonemes in continuous speech. Unfortunately, when applied to un-segmented continuous speech, neural networks have various deficiencies such as misfiring in the untrained regions in the input utterance.

For these reasons, we propose an expert system consisting of two parts: (1) a phoneme segmentation step based on spectrogram reading knowledge as used by human experts and (2) a phoneme identification step with neural networks using the phoneme boundaries determined in the first step.

In this paper, we describe a phoneme recognition expert system using spectrogram reading knowledge and neural networks, and report an experiment using this system.

2. EXPERT SYSTEM ARCHITECTURE

The proposed expert system is a rule-based inference system, into which human expert knowledge is incorporated as if-then rules. The expert system is implemented on a Symbolics 3600 using ART[8], a commercial tool for building expert systems. More details of this expert system, in particular, phoneme segmentation and phoneme identification, are described as follows.

2.1 PHONEME SEGMENTATION

CHARACTERISTICS

The following characteristics are incorporated into the system, to express human expert knowledge naturally and easily.

Assumption-based inference

Human expert knowledge is conditional on phonetic contexts including phoneme acoustic variations. Assumption-based inference is adopted to formalize such contextual knowledge. The contextual rules are written for their proper phonetic contexts, which are hypothesized when the rules are applied.

Top-down feature extraction

Human experts are able to precisely extract various kinds of global and local acoustic features, because they predict the presence of the features under phonetic context hypotheses. In the same manner, the system extracts acoustic features under the phonetic context hypotheses, in which the features are referred to by rules. This makes it possible to supply proper top-down control parameters for feature extraction. As a result, the various and suitable acoustic features can be precisely extracted.

Hypothesis evaluation using certainty factor

The correctness of the hypotheses is evaluated by referring to the acoustic features, which may be positive or negative evidence of the hypotheses, by several different rules. To evaluate the validity of the hypotheses from such evidence, certainty factors are calculated by the extending model built into MYCIN[9] (AND, OR or COMBINE functions).

Certainty factor assignment to acoustic features

Human expert knowledge of the acoustic features is fuzzy or qualitative. The system represents such fuzzy knowledge by mapping the acoustic measurements to the certainty factors, which represent the suitability of the measurements in their phonetic contexts. As a result, the various kinds of acoustic evidence can be treated uniformly by the certainty factors, which are accumulated into those of the hypotheses as described above.

SEGMENTATION STRATEGY

Knowledge of phoneme segmentation is elaborated in the following steps, as shown in Figure 1.

Detecting phoneme candidates

Coarsely classified phonemes and their rough locations are hypothesized by referring to global acoustic features. At this stage, extra phonemes may be hypothesized and, in addition, more than one phoneme may be hypothesized at the same location in the input utterance.
Hypothesizing phonetic contexts

The phonetic contexts, which include left and right phoneme classes and phoneme acoustic variations, are hypothesized for each phoneme candidate hypothesized above.

Detecting phoneme boundaries

Phoneme boundaries are detected under each phonetic context hypothesis, by referring to precise and local acoustic features. The acoustic features currently used are: (a) spectral power in certain frequency ranges, (b) time when the spectral power increases or decreases across thresholds, (c) time and magnitude of spectral power change peaks in certain frequency ranges, (d) frequency and magnitude of spectrum peaks, and (e) cutoff frequency of frication power.

Evaluating phoneme boundaries

At the same time, the hypotheses are evaluated by verifying the acoustic evidence for the hypotheses and by assigning certainty factors which express how likely the acoustic measurements around the boundaries are. As a result, the more reliable phoneme boundary hypothesis is assigned a larger certainty factor.

Selecting more reliable boundaries

As the result of detecting boundaries above, more than one left and right boundaries may be detected for a phoneme, or more than one phoneme may be detected at the same location in the input utterance. The more reliable boundary is selected from these boundaries. By default, the boundaries assigned larger certainty factors are selected.

As a result, the coarsely classified phonemes and their left and right boundaries are obtained with certainty factors.

2.2 PHONEME IDENTIFICATION

Phoneme identification is realized using Time-Delayed Neural Networks (TDNN) [10] applied to the phoneme boundary determined in phoneme segmentation.

CHARACTERISTICS OF TDNN

The TDNN we use has the following features:

- Easy training using the Back-propagation Learning Procedure.
- High performance phoneme identification using both time and spectrum domain information.
- Easy extension to large phoneme identification task by scaling up small modular TDNNs.
- Time shift invariance thanks to a time-shifted tied-connected weight architecture.

STRUCTURE

Fig. 2 shows the TDNN architecture, used here for identification of 18 consonants (B,D,G,P,T,K,M,N,sN,S,Sh,H,Z,Ch,Ts,R,W and Y). This network was constructed modularly from consonant subcategory networks (for the BDG,PTK,MN,sN,SShHZ, TsCh, RWY tasks), and, in addition, an interclass discrimination network that distinguishes between the consonant subclasses. The modular network is made up of 4 layers (the input layer, two hidden layers, and the output layer). The input layer has 15 frames X 16 spectral coefficient units, the 1st hidden layer has 13 X 8 units in the subclass modules and 20 X 8 units in the interclass module. The 2nd hidden layer has 9 X N units in the subclass modules, where N is the number of phonemes in the subclass, and 9 X 6 units in the interclass module, where 6 is the number of subclasses.
CONNECTIONS

The TDNN is made up of four layers. The lowest layer corresponds to spectral input values, the two next layers are hidden units, the topmost layer corresponds to phoneme output. The inputs of each unit are connected to units of the previous layer, via weights that are determined at training time. Each unit in hidden layer 1 is connected to a set of spectral values of three consecutive frames. Likewise, each unit of hidden layer 2 is connected to the set of units of hidden layer 1 for five consecutive frames. Finally, the output units are connected to the set of units of hidden layer 2 for nine consecutive frames. In this way, each output decision is made on a window of fifteen frames. The outputs correspond to phonemes for the intraclass modules, and to phoneme classes for the interclass module. Time-shifted connections were tied, i.e. the weights of corresponding connections were constrained to be identical wherever their position within the window.

As described above, each elementary TDNN net is duplicated for each single frame shift in time across the window. This applies to all connections. In this way, the network is forced to discover useful acoustic-phonetic features in the input, regardless of when they actually occurred within the window. This is an important property, as it makes the neural network less prone to slight segmentation errors.

LEARNING AND RECOGNITION

All phonemes used for training were segmented by hand and aligned with the TDNN input layer. In practice, the end point of every training phoneme was also aligned with the 10th frame of the TDNN. All weights were initialized with small random weights and then trained by the Back-propagation Learning Procedure [11].

Similarly, in recognition, the end point of the phoneme segmentation result was also adjusted to the 10th input point in the input layer. The phoneme identification result is identified by the maximum firing output unit.

3. EXPERIMENTS

TASK and DATA

The TDNN was trained using half of the 5,240 words (even numbered words) in ATR Japanese database, spoken by one male. An evaluation experiment was done using consonants taken from the other half of the database (odd numbered words). The task given to the expert system was to find the consonants in the words and to recognize their phoneme categories. The phoneme segmentation rules had been tuned to a success rate of 96.0% using 216 phonetically balance words from another database of the same speaker.

INPUTS

For the input to phoneme segmentation, speech, sampled at 12kHz, was analyzed by FFT on a 5msec Hamming window every 2.5msec. Then the spectrum was smoothed and the spectral power normalized to lie between -20dB and -80dB.

For the input to phoneme identification, the speech, sampled at 12 kHz, was analyzed by FFT on a 21.3msec Hamming window every 5msec. Melscale coefficients were computed from the power spectrum and adjacent coefficients in time collapsed resulting in an overall 10msec frame rate. The coefficients of an input token were then normalized to lie between -1.0 and +1.0 with the average at 0.0.

RESULTS

Table 1 shows the experiment results for both phoneme segmentation and phoneme identification. In Segmentation, segmentation correct is the percentage of phonemes which were detected within 50msec of the hand-labeled boundaries, alignment error is the average of the boundary alignment errors compared with hand-labeled boundaries, and false alarms is the ratio of extra segments to the number of consonants. In Identification, expert system ability is the percentage of phonemes both correctly segmented and identified, correct segmentation is the percentage of correctly identified phonemes given correct segmentation, and TDNN ability is the percentage of identification tested on the hand-labeled pre-segmented phonemes.

The segmentation score, 90.8% in total, was worse than the score for the tuning set 216 words, 96.0%, because of the acoustic variation or allophones not seen in the 216 words used for training. The boundary alignment errors compared with hand-labeled boundaries were 5.8msec on the average. The boundaries detected by the system are as accurate as the hand-labeled boundaries, which also have errors averaging less than 8msec [13].

The number of false alarms, or extra segments, were 32.8% of
the total number of consonants, respectively.

TDNN correctly identified 92.4% (within correct segmentation) of the phonemes which were correctly segmented by the system. This score was as good as the 93.3% score (TDNN ability) obtained on the hand-labeled pre-segmented phonemes. As a whole, the expert system correctly recognized 83.9% of the total number of phonemes, both in segmentation and identification. Each phoneme recognition ratio is over 75%, except for the phoneme /gi/. This is because we did not describe the rules to detect typical /ng/ segments, and because the identification ability of the TDNN for voiced consonants was slightly worse than that for other phonemes.

4. CONCLUSION

This paper presented a phoneme recognition expert system using spectrogram reading knowledge and neural networks, and an experiment using this system for speaker dependent phoneme recognition.

The expert system proposed here has the following characteristics:

- Highly accurate phoneme segmentation can be achieved by hypothesizing the coarse classified phoneme and its left and right contexts simultaneously when determining phoneme boundaries.
- High performance phoneme identification can be achieved by applying neural networks to the accurate result from the phoneme segmentation.

The experiment showed that the system achieved good performance in spite of acoustic variation in phoneme and phonetic contexts. In particular, boundary alignment was as good as that done by human experts, and phoneme identification using these boundaries was as good as when using hand-labeled ones.

In this experiment, TDNN which is the identification method, is only applied to the segmentation result. The best way to combine the segmentation and identification methods, so as to make use of their respective merits, remains to be studied.

ACKNOWLEDGMENTS

Thanks are due to Dr. Hidetoshi Sawai and Mr. Masanori Miyatake for their support on the neural networks. We also acknowledge Dr. Akira Kuematsu, President of ATR Interpreting Telephony Research Laboratories for his continuous support of this work.

REFERENCES


<table>
<thead>
<tr>
<th>phoneme category</th>
<th>number of samples</th>
<th>segmentation correct [%]</th>
<th>alignment error [msec]</th>
<th>false alarms [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>28</td>
<td>96.4</td>
<td>4.2</td>
<td>20.8</td>
</tr>
<tr>
<td>t</td>
<td>461</td>
<td>98.0</td>
<td>4.3</td>
<td></td>
</tr>
<tr>
<td>k</td>
<td>1300</td>
<td>97.8</td>
<td>5.7</td>
<td></td>
</tr>
<tr>
<td>ch</td>
<td>141</td>
<td>95.0</td>
<td>5.5</td>
<td></td>
</tr>
<tr>
<td>ts</td>
<td>220</td>
<td>93.6</td>
<td>5.4</td>
<td></td>
</tr>
<tr>
<td>s</td>
<td>572</td>
<td>89.0</td>
<td>3.8</td>
<td>39.6</td>
</tr>
<tr>
<td>sh</td>
<td>387</td>
<td>93.8</td>
<td>4.3</td>
<td></td>
</tr>
<tr>
<td>h</td>
<td>313</td>
<td>88.5</td>
<td>8.3</td>
<td>17.6</td>
</tr>
<tr>
<td>z</td>
<td>315</td>
<td>87.0</td>
<td>9.3</td>
<td>9.5</td>
</tr>
<tr>
<td>b</td>
<td>230</td>
<td>92.2</td>
<td>5.2</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>177</td>
<td>94.4</td>
<td>3.7</td>
<td>21.8</td>
</tr>
<tr>
<td>q</td>
<td>263</td>
<td>70.3</td>
<td>8.7</td>
<td></td>
</tr>
<tr>
<td>m</td>
<td>485</td>
<td>84.9</td>
<td>6.8</td>
<td>40.8</td>
</tr>
<tr>
<td>n</td>
<td>273</td>
<td>90.8</td>
<td>7.2</td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>760</td>
<td>84.5</td>
<td>6.2</td>
<td>120.7</td>
</tr>
<tr>
<td>total</td>
<td>5925</td>
<td>90.8</td>
<td>5.8</td>
<td>32.8</td>
</tr>
</tbody>
</table>

Table 1: Performance of phoneme recognition expert system.