RECOGNITION/UNDERSTANDING OF SPONTANEOUS SPEECH

Seiichi NAKAGAWA
Department of Information and Computer Sciences,
Toyohashi University of Technology
1-1, Hibarigaoka Tempaku-chou Toyohashi-shi Aichi-ken 441, JAPAN
E-mail: nakagawa@slp.tutics.tut.ac.jp

1. INTRODUCTION

In the conventional researches, a speech recognition system was mainly designed so that it could deal with read speech, not spontaneous speech. However, an ideal man machine interface can never be constructed without any techniques for recognizing spontaneous speech correctly. To achieve this goal, robustness of speech recognizers must be highly improved. With a view to realizing it, several studies have been made mainly from the following viewpoints.

1) Refinement of acoustic models to robustly characterize input speech in HMM paradigm. Segmental statistics were taken into considerations in this study and comparative experiments among a great number of forms of HMMs were carried out.

2) Robust treatment of illegal and unexpected linguistic phenomena occurred in spontaneous utterances. This study enabled the recognizer to deal with unknown words, interjections, ellipses, inversions, repairs in an adequate manner.

3) Development of a robust interpreter which can accept sentences including misrecognized words as well as the illegal linguistic phenomena as in 2). This study was conducted to apply the outputs from the improved recognizer by 2) in higher-level processings.

By integrating the above techniques into the system, its robustness was greatly improved, which consequently resulted in highly increase of recognition/understanding rate of spontaneous speech

2. REFINEMENT OF ACOUSTIC MODELS WITH SEGMENTAL STATISTICS[1]

The standard HMM cannot adequately represent time variant features during staying at a state. In order to improve the weakness, various methods have been studied so far. Here, segmental unit input HMMs were compared with conditional density HMMs or HMMs with regression coefficients. Experiments showed segmental unit input HMMs even in the basic structure generally showed better recognition rate than those by traditional methods. And the combination of segmental unit by mel-cepstrum frames and their regression coefficients showed the best.

2.1 Segmental Unit Input HMM

For an input symbol sequence $y = y_1 y_2, ..., y_T$ (T is the length of the sequence) and a state sequence $x = x_1 x_2, ..., x_T$, equations for output probability computation of HMM is given by the followings:

$$
\begin{aligned}
&P(y_1 \cdots y_T) \\
&= \sum_x \prod_i P(y_i|y_1 y_2 \cdots y_{i-2} y_{i-1}, x_1 x_2 \cdots x_{i-1} x_i) \times P(x_i|x_1 x_2 \cdots x_{i-1}) \\
&\simeq \sum_x \prod_i P(y_i|y_{i-3} y_{i-2} y_{i-1}, x_{i-1} x_i) P(x_i|x_{i-1}) && (1) \\
&= \sum_x \prod_i \frac{P(y_{i-3} y_{i-2} y_{i-1} y_i|x_{i-1} x_i)}{P(y_{i-3} y_{i-2} y_{i-1}|x_{i-1} x_i)} P(x_i|x_{i-1}) && (2) \\
&\simeq \sum_x \prod_i \frac{P(y_{i-1} y_i|x_{i-1} x_i)}{P(y_{i-1}|x_{i-1} x_i)} P(x_i|x_{i-1}) && (3) \\
&= \sum_x \prod_i P(y_i|y_{i-1}, x_{i-1} x_i) P(x_i|x_{i-1}) && (4) \\
&\simeq \sum_x \prod_i P(y_{i-1} y_i|x_{i-1} x_i) P(x_i|x_{i-1}) && (5) \\
&\simeq \sum_x \prod_i P(y_i|x_{i-1} x_i) P(x_i|x_{i-1}) && (6)
\end{aligned}
$$

Eq.(1) or Eq.(2) is a conditional density HMM of 4-frame segments; Eq.(3) or Eq.(4) is a conditional density HMM of 2-frame segments; and Eq.(5) is a segmental unit input HMM of 2-frame segments. The usual predictional HMM is realized in Eq.(1). Eq.(6) is a standard HMM. The segmental unit input HMM proposed here is given as approximation of Eq.(2), that is, we use only the numerator of Eq.(2) as follows:

$$
\begin{aligned}
&P(y_1 \cdots y_T) \\
&\simeq \sum_x \prod_i P(y_{i-3} y_{i-2} y_{i-1} y_i|x_{i-1} x_i) P(x_i|x_{i-1}) && (7)
\end{aligned}
$$

2.2 Continuous Syllable Recognition Experiments

Continuous syllable recognition experiments were designed on a speaker-independent mode. Syllable-based HMMs were trained using the ATR database uttered by 6 male speakers and retrained using additional data from the ASJ database uttered by 30 male speakers. The test data were from "Mt. Fuji Sightseeing Guide task", where 104 sentence were uttered by 6 new male speakers.

Results showed that the segmental unit input HMMs provided better results than the traditional HMMs as well as the the combination of the segmental unit HMMs with delta cepstrum coefficients indicated best performance.

2.3 Sentence Recognition Experiments

Sentence recognition experiments were also performed on a speaker-independent mode. The models used in these experiments were the same models as used in 2.2. The test data were also the same as used in the previous experiments. Word vocabulary size was 500 and the test set perplexity was calculated to be about 29.

The experimental results were generally similar to those in 2.2. And the best performance here was comparable to that previously obtained in speaker adaptation mode by cepstrum and delta cepstrum coefficients[2].

3. PROCESSING OF ILLEGAL UTTERANCES OCCURRED IN SPONTANEOUS SPEECH[3]

A new method for processing illegal utterances is proposed here to realize robust spontaneous speech recognition. Unknown words, interjections(filled pauses), and repairs are detected as arbitrary sequences of subword units by a subword-unit based decoder and the method is incorporated into the One-Pass search algorithm. Experiments showed that the method is sufficiently effective for dealing with the illegal linguistic expressions.

3.1 Incorporating Unknown Word Processing into One-Pass Search

In the dynamic time warping or the Viterbi scoring method, e.g. the One-Pass algorithm, the accumulated likelihood score of the hypothesis which arrives at a grammatical state q is represented as

$$Lq(i) = max(p, m, n)Lp(m) + L^n(m + 1 : i) \qquad (8)$$

where Lq(i) is the maximum verification score among the Viterbi paths which arrive at state q in i-th frame, $L^n(m+1 : i)$ is the accumulated verification score of word n summed up from m+1-th up to i-th frame and delta(p,n)=q (word n is outputted at state transition from p to q) should be satisfied if a syntactic constraint is imposed. The accumulated likelihood score of the optimal syllable sequence without lexical constraints can be represented as

$$L(i) = max(m, s)L(m) + L^s(m + 1 : i) \qquad (9)$$

where s is a syllable. We approximate the second term in equation (8) by equation (9) when n is an unknown word. That is, only the optimal boundaries obtained by equation (9) are considered as candidates of m in equation (8) and $L^n(m + 1 : i)$ is calculated as L(i)-L(m). In practice, we should impose a penalty on the likelihood score of unknown words because the score is always equal to or higher than the likelihood of any of the vocabulary words.

The above method can be extended by imposing general constraints for unknown words. The following two constraints are considered.

1) length of unknown word (number of syllables),

2) assumption that an unknown word can occur once per an utterance at most.

3.2 Performance of Unknown Word Processing

Experiments were carried out to examine the validity of the above method to verify unknown words. In these experiments a speech understanding system SPOJUS-SYNO-X[4] was used, where HMMs(5 states, 4 Gaussian densities and 4 discrete duration distributions) were adopted as 113 syllable acoustic models. The CFG grammar was used here and and the task domain was fixed to "Sightseeing guidance around Mt.Fuji" with a vocabulary size of 500 words.

Several proper nouns were removed from the lexicon to produce unknown words intentionally. And a special terminal symbol for unknown words was added to the original rewriting rules at the same time. This operation yielded 10 kinds of unknown words and they were scattered in 18 sentences while the total number of sentences was 104. These were uttered by 6 male speakers and, consequently, the total test set consisted of 624 sentence utterances(104 X 6). The test set perplexity was about 29 when all the unknown words were registered in the lexicon.

Recognition experiments were performed for the sentences including unknown words(UWs) and those not including ones under several conditions; 1)with no constraints("none"), 2)with length constraint("L"), 3)with occurrence constraint("N") and 4)their combination. The recognition experiment when UWs were registered and UW processing was not used (the 5-th condition) were also performed. The results showed that more than 70 % of unknown words were correctly detected if the successfully recognized utterances in the 5-th condition are only taken into consideration. And the added constraints improved both the detection accuracy of unknown words and the false alarm rate.

For utterances including interjections, the above unknown word processing(UWP) was evaluated in comparison with the conventional method, where frequent interjections were pre-registered in the lexicon. Although the results can't be described here in detail for space, the combination of the UWP and the conventional method showed the best performance on interjections.

4. INTERPRETATION OF SENTENCES WITH MISRECOGNIZED WORDS AND ILLEGAL EXPRESSIONS[5]

By incorporating the proposed technique in section 3, the system came to correctly recognize sentences with illegal expressions. If the interpretation module can't handle these sentences adequately, however, the improvement of the recognizer does not lead directly to the increase of overall performance as a spoken dialogue system. Based upon these considerations, a robust interpreter was developed which can accept misrecognized words as well as illegal linguistic expressions.

4.1 Development of a Robust Interpreter

In making our previously developed interpreter more robust, we consider it necessary to refer to HUMAN strategies to handle sentences including illegal expressions or misheard words. These consideration brought us to conduct some experiments to examine whether humans can adequately correct the output sentences from the recognizer, which can have misrecognized words as well as illegal linguistic expressions. The detail results are discussed in [5]. Here, we focus especially on how to apply the findings obtained in the experiments.

In the developed interpreter, the following procedures are carried out.

1. Syntactic and semantic analyses, which are conducted with different capabilities to accept illegal expressions. The input sentences from the recognizer go though the following analyses in the order of (a) to (d). If one of the analyses succeeds in producing a semantic representation for the input, go to process 2 omitting the rest of the analyses. Without any semantic representation produced after the analysis (d), go to process 3.

    (a) analysis for legal sentences.

    (b) analysis for sentences
    allowing omission of post-positions.

    (c) analysis for sentences
    allowing omission of post-positions and word order inversion.

    (d) analysis for sentences
    allowing invalid (misrecognized) post-positions and word order inversion.

2. Semantic filtering, where the validity of the semantic representation obtained in 1. is checked by domain knowledge.

   (a) If it is judged as correct, the semantic representation is produced as the final output from the interpreter.

   (b-1) Otherwise, if applicable heuristic rules are prepared, the representation is corrected and produced as the final output from the interpreter.

   (b-2) With no rules for correcting, go to process 3.

3. Keyword analysis (mentioned below) is performed by using partial results of the above analysis.

4.2 Evaluation of the Developed Interpreter

Evaluation experiments were carried out with the same domain and task as in 3.2, where seven novice subjects were asked to use this system with spontaneous speech. Before the experiments, the subjects were explained about this system and shown some examples of scenario actually uttered in a previous experiment.

Through the experiments, 232 utterances were recorded in total at a computer room. About 70 % of the utterances were correctly understood and about 13 % were rejected. Simulating a perfect recognizer by inputs of transcription of the above utterances, shows about 78 % as understanding rate. Further, the achievement rate of the task was 98.2 %. These results indicates the validity of the developed interpreter.

5. CONCLUSIONS

In order to realize a robust recognition/understanding system for spontaneous speech, several studies have been made mainly from the three viewpoints. Firstly, segmental statistics was introduced into the acoustic feature modeling in HMM paradigm. Through the experiments, segmental unit HMMs constantly provided better recognition rates than those by conventional HMMs. Secondly, to directly improve the robustness of our recognizer, we proposed a new method, where illegal linguistic expressions were detected as arbitrary sequences of subword units by a subword-unit based decoder. In implementing this method, it was incorporated into the One-Pass search algorithm. Evaluation experiments showed that more than 70 % of unknown words were correctly detected and that this method was also valid in detecting interjections. Lastly, a robust interpreter was developed so that it could accept and correctly understand the outputs from the improved recognizer. In the development, human strategies for correcting illegal sentences were considered. Consequently, the syntactic/semantic analysis modules with different capabilities to accept illegal phenomena and the semantic filtering module were implemented. By utilizing these modules adequately, it became possible to understand spontaneous speech with higher rate.

REFERENCES (PUBLISHED PAPERS)

[1] K.Yamamoto and S.Nakagawa, "Comparative evaluation of segmental unit input HMM and conditional density HMM," Proc. Eurospeech'95, pp.1615-1618 (1995).

[2] Y.Tsurumi and S.Nakagawa, "An unsupervised speaker adaptation method for continuous parameter HMM by maximum a posteriori probability estimation," Proc. ICSLP'94, pp.431-434 (1994).

[3] A.Kai and S.Nakagawa, "Investigation on unknown word processing and strategies for spontaneous speech understanding," Proc. Eurospeech'95, pp.2095-2098 (1995).

[4] S.Nakagawa and A.Kai, "A context-free grammar-driven, One-Pass HMM-based continuous speech recognition method," Systems and Computers in Japan, Vol.25, No.4, pp.92-102 (1994).

[5] M.Hidano, T.Itoh, M.Yamamoto and S.Nakagawa, "Spontaneous speech understanding for a dialogue system," Proc. ESCA Workshop on Spoken Dialogue Systems, pp.25-28 (1995).