PROCESSING UNKNOWN WORDS IN CONTINUOUS SPEECH RECOGNITION

Kenji Kita, Terumasa Ehara, Tsuyoshi Morimoto

ATR Interpreting Telephony Research Laboratories
Seika-cho, Souraku-gun, Kyoto 619-02, Japan

ABSTRACT

Current continuous speech recognition systems essentially ignore unknown words. Systems are designed to recognize words in the lexicon. However, for using speech recognition systems in real applications of spoken-language processing, it is very important to process unknown words. This paper proposes a continuous speech recognition method which accepts any utterance that might include unknown words. In this method, words not in the lexicon are transcribed as phone sequences, while words in the lexicon are recognized correctly. The HMM-LR speech recognition system, which is an integration of Hidden Markov Models and generalized LR parsing, is used as the baseline system, and enhanced with the trigram model of syllables to take into account the stochastic characteristics of a language. Preliminary results indicate that our approach is very promising.

1 INTRODUCTION

For natural language applications, processing unknown words is one of the most important problems. It is almost impossible to include all words in the system's lexicon.

In the area of written language processing, some methods for handling unknown words have been proposed. For example, Tomita (1986) shows that unknown words can be handled by the generalized LR parsing framework. In generalized LR parsing, it is easy to handle multi-part-of-speech words, and an unknown word can be handled by considering it as a special multi-part-of-speech word.

Unfortunately, in the area of continuous speech recognition, there has been little progress in unknown word processing. Unlike written language processing, in continuous speech recognition, word boundaries are not clear and the correct input is not known, so the problem is more difficult. Recently, Asadi et al. (1990) proposed a method for automatically detecting new words in a speech input. In their method, an explicit model of new words is used to recognize the existence of new words.

This paper proposes a continuous speech recognition method which accepts any utterance that might include unknown words. In our approach, the HMM-LR continuous speech recognition system for Japanese (Kita et al. 1989a; Kita et al. 1989b; Hanazawa et al. 1990) is used as the baseline system, and is an integration of *Hidden Markov Models* (HMM) (Levinson et al. 1983) and *generalized LR parsing* (Tomita 1986). The HMM-LR system is a syntax-directed continuous speech recognition system. The system outputs sentences that the grammar can accept.

The Hidden Markov Model is a stochastic approach for modeling speech, and has been used widely for speech recognition. It is suitable for handling the uncertainty that arises in speech, for example, contextual effects, speaker variabilities, etc. Moreover, if the HMM unit is a phone, then any word models can be composed of phone models. Thus, it is easy to construct a large vocabulary speech recognition system.

In our approach, two kinds of grammars are used. The first grammar is a normal grammar which describes our task. The lexicon for the task is embedded in this grammar as phone

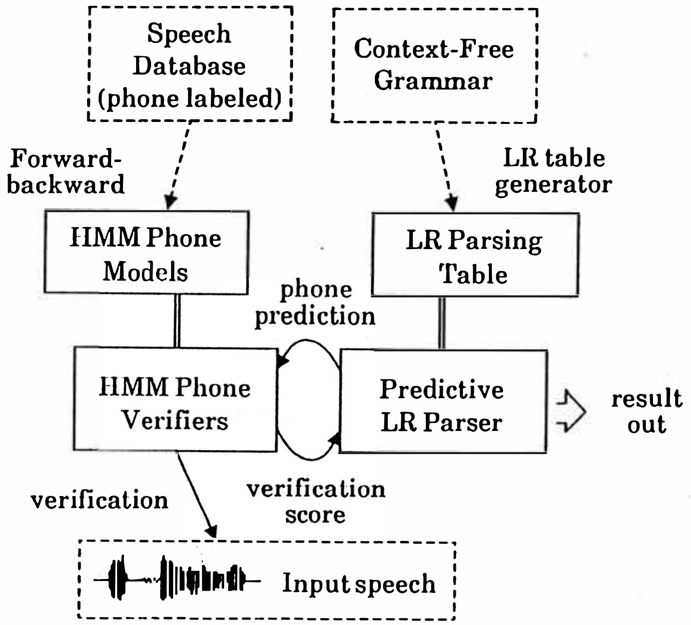

Figure 1: Schematic diagram of HMM-LR speech recognition system

sequences. The second grammar describes the Japanese phonemic structure, in which constraints between phones are written. These two grammars are merged and used in the HMM-LR system. The HMM-LR system outputs words in the lexicon if no unknown word is included in a speech input. If an unknown word is included, then the system outputs a phonemic transcription that corresponds to the unknown word. However, the second grammar by itself is too weak to get correct phonemic transcriptions. We strengthened the grammar by adding other linguistic information, the trigram model based on Japanese syllables. A trigram model is an extremely rough approximation of a language, but it is very practical and useful. By adding the trigram model of syllables, the performance of the system is improved drastically.

2 HMM-LR CONTINUOUS SPEECH RECOGNITION SYSTEM

First, we will review the baseline system, the HMM-LR continuous speech recognition system (Figure 1). This system is an integration of the phone-based HMM and generalized LR parsing.

(1)	*S*	→	*NP VP*
(2)	*NP*	→	*DET N*
(3)	*VP*	→	*V*
(4)	*VP*	→	*V NP*
(5)	*DET*	→	*/z/ /a/*
(6)	*DET*	→	*/z/ /i/*
(7)	*N*	→	*/m/ /ae/ /n/*
(8)	*N*	→	*/ae/ /p/ /a/ /l/*
(9)	*V*	→	*/iy/ /ts/*
(10)	*V*	→	*/s/ /ih/ /ng/ /s/*

Figure 2: An example of a grammar with phonetic lexicon

In HMM-LR, the LR parser is used as a language source model for symbol prediction/generation. Thus, we will hereafter call the LR parser the ***predictive LR parser.*** A phone-based predictive LR parser predicts next phones at each generation step and generates many possible sentences as phone sequences. The predictive LR parser determines next phones using the LR parsing table of the specified grammar and splits the parsing stack not only for grammatical ambiguity but also for phone variation. Because the predictive LR parser uses context-free rules whose terminal symbols are phone names, the phonetic lexicon for the specified task is embedded in the grammar. An example of context-free grammar rules with a phonetic lexicon is shown in Figure 2. Rule (5) indicates the definite article "the" before consonants, while rule (6) indicates the "the" before vowels. Rules (7), (8), (9) and (10) indicate the words "man", "apple", "eats" and "sings", respectively.

The actual recognition process is as follows. First, the parser picks up all phones predicted by the initial state of the LR parsing table and invokes the HMM models to verify the existence of these predicted phones. The parser then proceeds to the next state in the LR parsing table. During this process, all possible partial parses are constructed in parallel. The HMM phone verifier receives a probability array which includes end point candidates and their probabilities, and updates it using an HMM probability calculation. This probability array is attached to each partial parse. When the highest probability in the array is under a certain threshold level, the partial parse is pruned. The recognition process proceeds in this way until the entire speech input is processed. In this case, if the best probability point reaches the end of the speech data, parsing ends successfully.

High recognition performance is attained by driving HMMs directly without any intervening structures such as a phone lattice. A more detailed algorithm is presented in (Kita et al. 1989a; Kita et al. 1989b).

3 TRIGRAM MODEL OF SYLLABLES

3.1 STOCHASTIC LANGUAGE MODELING

Language models such as context-free grammars or finite state grammars are effective in reducing the search space of a speech recogniton system. These models, however, ignore the stochastic characteristics of a language. By introducing stochastic language models, we can assign the *a priori* probabilities to word/phone sequences. These probabilities, together with acoustic probabilities, determine most likely recognition candidates.

Having observed acoustic data y, a speech recognizer must decide a word sequence $\hat{w}$ that satisfies the following condition:

$$P(\hat{w}|y) = \max_{w} P(w|y)$$

By Bayes' rule,

$$P(w|y) = \frac{P(y|w)P(w)}{P(y)}$$

Since $P(y)$ does not depend on w, maximizing $P(w|y)$ is equivalent to maximizing $P(y|w)P(w)$. $P(w)$ is the *a priori* probability that the word sequence w will be uttered, and is estimated by the *language model*. $P(y|w)$ is estimated by the *acoustic model*. Note that we are using HMM as an acoustic model.

3.2 TRIGRAM MODEL OF SYLLABLES

Word bigram/trigram models are extensively used to correct recognition errors and improve recognition accuracy (Shikano 1987; Paeseler and Ney 1989).

The general idea of a trigram model can be easily applied to Japanese syllables. A typical syllable in Japanese is in the form of a CV, namely one consonant followed by one vowel, and the number of syllables is very small (about one hundred). Moreover, Japanese syllables seem to have a special stochastic structure. Araki et al. (1989) suggest that a statistical method based on Japanese syllable sequences is effective for ambiguity resolution in speech recognition systems. Thus, a syllable trigram model is effective for recognizing Japanese syllable sequences.

In our syllable trigram model, the *a priori* probability $P(S)$ that the syllable sequence $S = s_1, s_2, \ldots, s_n$ will be uttered is calculated as follows (Kita et al. 1990).

$$P(s_1, \ldots, s_n) =$$
$$P(s_1 \mid \#)P(s_2 \mid \#, s_1) \prod_{k=3}^{n} P(s_k \mid s_{k-2}, s_{k-1})$$
$$P(\# \mid s_{n-1}, s_n)$$

$$P(s_k \mid s_{k-2}, s_{k-1}) = q_1 f(s_k \mid s_{k-2}, s_{k-1}) + q_2 f(s_k \mid s_{k-1}) + q_3 f(s_k) + q_4 C$$

$$q_1 + q_2 + q_3 + q_4 = 1$$

$$f(s_k \mid s_{k-2}, s_{k-1}) = \frac{N(s_{k-2}, s_{k-1}, s_k)}{N(s_{k-2}, s_{k-1})}$$

In the above expressions, "#" indicates the phrase boundary marker, and C is a uniform probability that each syllable will occur. The function N counts the number of occurrences of its arguments in the training data. The optimal interpolation weights q_i are determined using *deleted interpolation* (Jelinek and Mercer 1980). Given a collection of training data, the interpolation weights are estimated as follows (Kawabata et al. 1990).

1. Make an initial guess of q_i that $\sum_i q_i = 1$ holds.

2. Calculate i-gram probabilities f_i^j when the j-th data is removed from the training data.

3. Re-estimate q_i by the following formula.

   $$\hat{q}_i = \frac{1}{N} \sum_{j=1}^{N} C_i^j$$

   where N is the number of syllables in training data, and

   $$C_i^j = \frac{q_i f_i^j}{\sum_k q_k f_k^i}$$

4. Replace q_i with $\hat{q}_i$ and repeat from step 2.

4 PROCESSING UNKNOWN WORDS IN AN HMM-LR SPEECH RECOGNITION SYSTEM

4.1 GRAMMAR FOR JAPANESE PHONEMIC STRUCTURE

The HMM-LR system is a syntax-directed continuous speech recognition system. If we use a grammar which describes the Japanese phonemic structure, we can then construct a *phonetic typewriter* for Japanese. This grammar includes rules like "a sequence of consonants doesn't appear" or "the syllabic nasal /N/ doesn't appear at the head of a word". This grammar does not include phonemic spellings for each word, so this grammar is suitable for transcribing an unknown word as a phone sequence.

However, because the *perplexity* [1] of this grammar is quite large, the trigram model of Japanese syllables is used at the same time. By adding the trigram model of syllables, the perplexity of the grammar is reduced from 18.3 to 4.3 (Kawabata et al. 1990).

4.2 UNKNOWN WORD PROCESSING

For processing unknown words, two kinds of grammars are used. The first grammar is a normal grammar which describes our task. The phonemic spellings for each word are also included in this grammar. The second grammar is a grammar for Japanese phonemic structure, mentioned in the previous subsection. Hereafter, these two grammars are referred to as the *task grammar* and the *phonemic grammar*, respectively.

These two grammars are merged and used in the HMM-LR system. When merging two grammars, the start symbol of the phonemic grammar is replaced with pre-terminal names that might include unknown words (in our experiments, *proper-noun* is allowed to include unknown words).

If a speech input includes an unknown word, then a segment of speech input does not match well with any word in the system's lexicon. In this case, the grammar for phonemic structure produces the phone sequence that matches well with the unknown word. If the speech input includes no unknown word, then the HMM-LR system outputs words in the lexicon.

[1] Perplexity is a measurement of language model quality. It represents the average branching of the language model. In general, as perplexity increases, speech recognition accuracy decreases. For more details, see (Jelinek 1990).

4.3 RECOGNITION LIKELIHOOD

The HMM-LR continuous speech recognition system uses the *beam-search* technique to reduce the search space. A group of likely recognition candidates are selected using the likelihood of each candidate. The likelihood S is calculated as follows.

$$S = (1 - \lambda)S^{(HMM)} + \lambda S^{(SYLLABLE)}$$

$S^{(HMM)}$ and $S^{(SYLLABLE)}$ are the log likelihoods based on the HMM and the trigram model of syllables, respectively. The scaling parameter λ is introdued to adjust the scaling of the two kinds of likelihoods, as determined by preliminary experiments.

At the end of recognition, the likelihood of recognition candidates that include unknown words are penalized a small value to reduce the false alarms.

5 EXPERIMENTS

5.1 HMM PHONE MODELS

HMMs used in the experiments are basically the same as reported in (Hanazawa et al. 1990). HMM phone models based on the discrete HMM are used as phone verifiers. A three-loop model for consonants and a one-loop model for vowels are trained using each phone data extracted from the *ATR isolated word database* (Kuwabara et al. 1989).

Duration control techniques and separate vector quantization are used to achieve accurate phone recognition.

5.2 SPEECH DATA

The experiments were carried out using 25 sentences including 279 phrases uttered by one male speaker.

The speech is sampled at 12kHz, pre-emphasized with a filter whose transform function is $(1-0.97z^{-1})$, and windowed using a 256-point Hamming window every 9 msec. Then, 12-order LPC analysis is carried out. Spectrum, difference cepstrum coefficients, and power are computed. Multiple VQ codebooks for each feature were generated using 216 phonetically balanced words. Hard vector quantization without the fuzzy VQ was performed for HMM training. Fuzzy vector quantization (fuzziness = 1.6) was used for test data.

5.3 LINGUISTIC DATA

Syllable trigrams were estimated using a large number of training texts extracted from the *ATR dialogue database* (Ehara et al. 1990). This database contains not only raw texts but also various kinds of syntactic/semantic information, such as *parts of speech*, *pronounciation* and *conjugational patterns*, etc. The training texts includes approximately 73,000 phrases and 300,000 syllables.

5.4 GRAMMARS

As stated earlier, the task grammar and the phonemic grammar are merged into one grammar and used in the HMM-LR system. The task grammar describes the domain of an *International Conference Secretarial Service* and has 1,461 rules including 1,035 words. Of course, all the words which appear in the test data are included in this grammar.

To evaluate the unknown word processing method, all proper nouns (8 words), such as a person's name and a place name, were removed from the task grammar.

5.5 RESULTS

Table 1 shows the transcription rates for phrases that include unknown words. Here the transcription rate is equal to *phone accuracy* (Lee 1989), which can be calculated as follows.

$$phone\,accuracy = \frac{total - sub - ins - del}{total} \times 100$$

where *total* indicates the total number of phones in test data, and *sub*, *ins* and *del* are the number

Table 1: Transcription rates for phrases that include unknown words

Without syllable trigrams	*With syllable trigrams*
66.1%	95.3%

Table 2: Examples of recognition results that include unknown words

Input		*Results*	
Correct	*Meaning*	*Without syllable trigrams*	*With syllable trigrams*
higashiku	higashiku (*place name*)	shigashiku	higashiku
ichitaroudesu	(I am) Ichitarou	ishitaoouutsusu	ishitaroudesu
takarasamadesune	(You are) Mr. Takara (aren't you)	takaasabautsunu	takarasamadesune
kyoutoekikara	from Kyoto station	hyotorekitaafu	kyoutoekikara
kitaoojiekimade	to Kitaooji station	shitaouziekimare	kitaoojiekimade

Table 3: Phrase recognition rates (with syllable trigrams)

rank	*Task grammar*	*Task grammar + Phonemic grammar*
1	87.5%	81.7%
2	93.5%	86.4%
3	94.6%	87.5%

of phones recognized as incorrect, deleted and inserted, respectively.

Table 2 shows examples of recognition results that include unknown words. By using the trigram model of Japanese syllables, the system can output very close phonemic transcriptions for unknown words.

Table 3 shows the phrase recognition rates for two kinds of grammars, the task grammar and a merged grammar consisting of the task grammar and the phonemic grammar. These grammars are both enhanced with the trigram model of syllables. By adding the phonemic grammar, the phrase recognition rate dropped from 87.5% to 81.7%. This is because the phonemic grammar sometimes causes a word to be recognized as a phone sequence despite the word being in the lexicon.

6 CONCLUSION

In this paper, we described a continuous recognition method that can process unknown words. The key idea is merging a task grammar and a phonemic grammar. If no unknown word is included in the speech, then the system uses the task grammar and outputs a correct result. If an unknown word is included, then the system uses the phonemic grammar and outputs a phonemic transcription for the unknown word. We also showed that the trigram model of Japanese syllables is very effective in getting phonemic transcriptions for unknown words.

This is our first approach. There are many problems that must be resolved. Further development to improve the system is currently in progress.

ACKNOWLEDGMENTS

The authors are deeply grateful to Dr. Kurematsu, the president of ATR Interpreting Telephony Research Laboratories, all the members of the *Speech Processing Department* and the *Knowledge and Data Base Department* for their constant help and encouragement.

REFERENCES

[1] Araki, T.; Murakami, J.; and Ikehara, S. 1989 Effect of Reducing Ambiguity of Recognition Candidates in Japanese Bunsetsu Units by 2nd-Order Markov Model of Syllables. *Transactions of Information Processing Society of Japan.* Vol. 30, No. 4 (in Japanese).

[2] Asadi, A.; Schwartz, R. S.; and Makhoul, J. 1990 Automatic Detection of New Words in a Large Vocabulary Continuous Speech Recognition System. *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing.*

[3] Ehara, T.; Ogura, K.; and Morimoto, T. 1990 ATR Dialogue Database. *Proceedings of the International Conference on Spoken Language Processing.*

[4] Hanazawa, T.; Kita, K.; Nakamura, S.; Kawabata, T.; and Shikano, K. 1990 ATR HMM-LR Continuous Speech Recognition System. *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing.* Also In: Waibel, A. and Lee, K. F. (eds.) *Readings in Speech Recognition.* Morgan Kaufmann Publishers.

[5] Jelinek, F. and Mercer, R. L. 1980 Interpolated Estimation of Markov Source Parameters from Sparse Data. In: Gelsema, E. S. and Kanal, L. N. (eds.) *Pattern Recognition in Practice.* North Holland.

[6] Jelinek, F. 1990 Self-Organized Language Modeling for Speech Recognition, In: Waibel, A. and Lee, K. F. (eds.) *Readings in Speech Recognition.* Morgan Kaufmann Publishers.

[7] Kawabata, T.; Hanazawa, T.; Itoh, K.; and Shikano, K. 1990 HMM Phone Recognition Using Syllable Trigrams. *IEICE Technical Report.* SP89-110 (in Japanese).

[8] Kita, K.; Kawabata, T.; and Saito, H. 1989a HMM Continuous Speech Recognition Using Predictive LR Parsing. *Proceedings of the 1989 International Conference on Acoustics, Speech, and Signal Processing.*

[9] Kita, K.; Kawabata, T.; and Saito, H. 1989b Parsing Continuous Speech by HMM-LR Method. *First International Workshop on Parsing Technologies.*

[10] Kita, K.; Kawabata, T.; and Hanazawa, T. 1990 HMM Continuous Speech Recognition Using Stochastic Language Models. *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing.*

[11] Kuwabara, H.; Takeda, K.; Sagisaka, Y.; Katagiri, S.; Morikawa, S.; and Watanabe, T. 1989 Construction of a Large-Scale Japanese Speech Database and its Management System. *Proceedings of the 1989 International Conference on Acoustics, Speech, and Signal Processing.*

[12] Lee, K. F. 1989 *Automatic Speech Recognition: The Development of the SPHINX System.* Kluwer Academic Publishers.

[13] Levinson, S. E.; Rabiner, L. R.; and Sondhi, M. M. 1983 An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition. *Bell System Technical Journal.* Vol. 62, No. 4.

[14] Paeseler, A. and Ney, H. 1989 Continuous-Speech Recognition Using a Stochastic Language Model. *Proceedings of the 1989 International Conference on Acoustics, Speech, and Signal Processing.*

[15] Shikano, K. 1987 Improvement of Word Recognition Results by Trigram Model. *Proceedings of the 1987 International Conference on Acoustics, Speech, and Signal Processing.*

[16] Tomita, M. 1986 *Efficient Parsing for Natural Language: A Fast Algorithm for Practical Systems.* Kluwer Academic Publishers.