First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Japanese Phonetic Typewriter Using HMM Phone Units and Syllable Trigrams

Takeshi Kawabata, Toshiyuki Hanazawa[†],
Katsunobu Itoh[††], and Kiyohiro Shikano[†††]

NTT Basic Research Laboratories
Midori-cho, Musashino-shi, Tokyo 180, Japan

ABSTRACT

This paper describes a Japanese phonetic typewriter based on HMM phone units and syllable trigrams. Even though HMM methods have considerable ability to recognize speech, it is still difficult to recognize individual phones in continuous speech without lexical information. This paper reports on a phonetic typewriter to improve HMM phone recognition performance by incorporating syllable trigrams. HMM phone units are trained using an isolated word database, and their duration parameters are modified in relation to the speaking rate. The syllable trigram tables are made from a large text database of over 35,000 syllables. Phone sequence probabilities calculated from the trigrams are combined with HMM probabilities. Limiting the number of intermediate candidates using these probabilities leads to an accurate phonetic typewriter system without excessive computation time.

1. INTRODUCTION

Speech is the most convenient communication media for human beings. Applying the speech interaction mechanism to human-machine communication requires a large vocabulary and a continuous speech recognition system.

To construct a large-vocabulary system, phone recognition approaches are promising. The system recognizes any word, phrase or sentence by recognizing the phones in the utterances. Recently, phone recognition was greatly advanced using Hidden Markov Models.[1] This has made a practical phone-based speech recognition system possible.[2]

Even though HMM methods are efficient in speech/phone recognition, it is still difficult to recognize individual phones in continuous utterances without lexical information. A kana trigram model has been shown to be effective for correcting errors in the character recognition. A kana character corresponds to a spoken syllable. Similarly, syllable trigrams limit the phone perplexity and improve phone recognition without lexical information.[3]

This paper describes an unlimited-vocabulary continuous speech recognition system based on HMM phone units and syllable trigrams. Because the system transforms any speech input into phonetic sequences, we call the system a "Phonetic Typewriter".

2. SYSTEM OVERVIEW

Figure 1 shows a schematic diagram of the phonetic typewriter. The system has two parts. The left side is for acoustic processing. The HMM phone verifier verifies each phone in continuous speech using HMM phone units trained on an isolated-word database, and HMM duration parameters are modified to match the speaking rate.

The right side is for predicting next phones. The predictive LR parser[4] generates possible phone sequences from left to right, and concatenates a corresponding HMM according to their phonetic representation. The parser calculates the phone sequence probability based on the syllable trigrams and HMM probabilities. The system recognizes speech as the phone sequence with the highest probability.

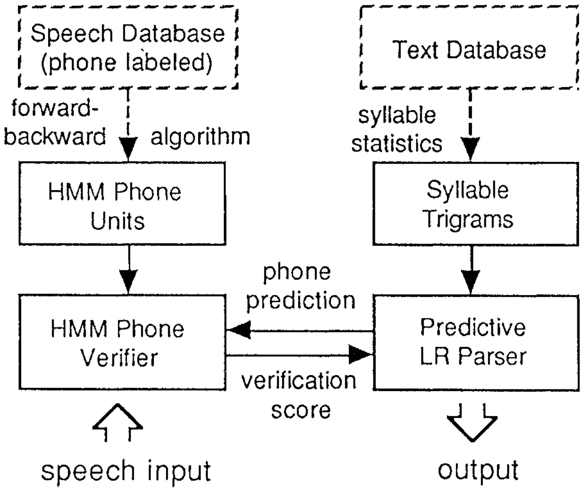

Fig.1 Schematic diagram of the phonetic typewriter

[†] Mitsubishi Electric Corporation
[††] Tokyo Institute of Technology
[†††] NTT Human Interface Research Laboratories

3. HMM PHONE UNITS

In this system, Japanese phonemes are categorized into transient or stationary phone classes. The model for a transient phone has three loops representing its time structure. The model for a stationary phone has only one loop (Fig. 2). Each loop is strongly related to an acoustical event.

These phone units are trained using the forward-backward algorithm on a large vocabulary of isolated words. A male speaker uttered 5240 important Japanese words, 216 phonetically balanced words, and 101 syllables. The speech is sampled at 12 kHz and transformed to VQ code sequences using 12th order LPC analysis and a 256-point Hamming window shifted every 9 ms.

For accurate phone modeling, a multiple-codebook method is used. The following three parameters are vector-quantized separately;[2]
(i) Spectrum (WLR),
(ii) LPC cepstral difference (DCEP), and
(iii) Power,
where their codebook sizes are 256, 64, and 256, respectively.

An active state duration control mechanism is implemented in the system. After the forward-backward training, the HMM state duration distribution is determined by the Viterbi alignment on training data. Each state distribution is approximated by the Gaussian distribution. The system detects the speaking rate of input speech, and corrects the mean and variance of the duration distribution. The distribution function is used as a penalty for HMM transition and output probabilities at each loop. The modified HMM forward probability for the loop of state j at time t is as follows;[1,2,5]

$$\alpha(j,t) = \sum_{i}\sum_{\tau}\alpha(i,t-\tau-1)\cdot a_{ij}\cdot b_{ijv_{t-\tau}} \cdot\left(\prod_{k=t-\tau+1}^{t} a_{jj}\cdot b_{jjv_k}\right)\cdot d(j,\tau)^{w} \qquad (1)$$

where
$\alpha(i,t)$: forward probability of state j at time t
a_{ij} : transition probability from state i to state j
b_{ijv} : output probability of v from state i to state j
$d(i,\tau)$: corrected HMM duration distribution function of duration τ at state i
w : constant.

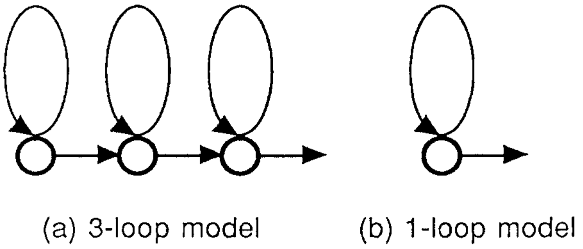
(a) 3-loop model (b) 1-loop model

Fig.2 Hidden Markov Models for phone recognition

4. SYLLABLE TRIGRAMS

The syllable trigram tables are made from a text database of over 35,000 syllables. The database consists of five blocks, as shown in Table 1. The first three blocks consist of editorial columns, which are typical of written text. The other two blocks consist of spoken language collected from keyboard or telephone dialog simulation. Two persons communicate through video display terminals or telephones. One person plays the role of an secretary at an international conference, and the other plays an applicant. Keyboard dialogs are stored directly into memory . Telephone dialogs are recorded and then transcribed. Table 1 shows the text sources and the number of syllables in each block.

The n-gram of a syllable sequence $s_0 \ldots s_{N-1}\, s_N$ is defined as follows.

$$P^{(n)} = P\left(s_i \middle| s_{i-(n-1)}, s_{i-(n-2)}, \cdots, s_{i-1}\right) \qquad (2)$$

Customarily, $P^{(1)}$, $P^{(2)}$ and $P^{(3)}$ are called unigram, bigram, and trigram. A trigram probability is estimated by

$$P\left(s_i \middle| s_{i-2}, s_{i-1}\right) = N_{s_{i-2},s_{i-1},s_i} / N_{s_{i-2},s_{i-1}} , \qquad (3)$$

where Ns_{i-2},s_{i-1},s_i is the number of occurrences of the syllable sequence s_{i-2},s_{i-1},s_i in the text database, and Ns_{i-2},s_{i-1} is that of s_{i-2},s_{i-1} .

We measure the complexity of a recognition task by "phone perplexity", which is defined as 2 to the power of the entropy per phone. A larger perplexity indicates that the task is more difficult. Table 2 shows the reduction of phone perplexity of a Japanese phrase recognition task by syllable n-grams.

Since the number of syllables in the text database is finite, the database cannot cover every actual syllable sequences. The n-gram coverage on testing data (Japanese phrases) are shown in Table 2. For example, 4% of the 3-syllable sequences in the testing data do not appear in the training database. As the n-gram probability of such a

Table 1 The number of syllables in the text database

text source	number of syllables
1. magazine	34,308
2. newspaper	38,050
3. book	42,449
4. keyboard dialogs	53,998
5. telephone dialogs	154,015
total	322,820

Table 2 Reduction of the phone perplexity by syllable n-grams

gram	no model	1 (uni-)	2 (bi-)	3 (tri-)
phone perplexity (with 1.e-5 flooring)	18.3	10.2	6.1	4.0
coverage (without flooring)	-	100%	99%	96%

sequence is zero, the system cannot recognize a speech input containing the sequence. Flooring is a convenient, if simple, technique for avoiding the problem. If the calculated value is smaller than 1.e-5, the flooring value (1.e-5) is used as the n-gram probability.

This paper also tests another approach based on "deleted interpolation" (D.I.).[6] An n-gram probability is interpolated from k(=0..3)-gram probabilities. Each k-gram probability is weighted by the reliability of the k-gram estimation. The reliability factor is determined by the deleted interpolation technique.

[Deleted Interpolation]

The n-gram probability is approximated as

$$P^{(n)} = \sum_{k=0}^{3} \lambda_k P^{(k)} \qquad (4)$$

where λ_k is a weighting coefficient for the k-gram probability. The problem to be solved is an optimization of λ_k with the condition $\Sigma \lambda_k = 1$.

Now, suppose that the training text is a phrase sequence b_1 ... b_N . The generation probability of the phrase b_j , which is a syllable sequence s_0 ... s_K , is

$$P(b_j) = \prod_{i=1}^{K} P_i^{(n)} \qquad (5)$$

The mean phrase-generation probability is the geometrical mean of Equation 5. The probability is maximized through the following procedure.

(I) Repeat steps (II) and (III) for each phrase ($i=1..N$).

(II) Delete phrase b_j from the training data, and calculate $P^{(0)}$, $P^{(1)}$, $P^{(2)}$, and $P^{(3)}$ from the remaining data.

(III) Calculate the contribution factor of the i-th k-gram on the phrase b_j as

$$c_i^{(k)}(b_j) = \frac{\lambda_k P_i^{(k)}}{\lambda_0 P_i^{(0)} + \lambda_1 P_i^{(1)} + \lambda_2 P_i^{(2)} + \lambda_3 P_i^{(3)}} , \qquad (6)$$

where i is the position of the syllable in the phrase b_j .

(IV) Update λ_k by the formula:

$$new\ \lambda_k = \sum_{j=1}^{N} \sum_{i=1}^{K_j} c_i^{(k)}(b_j) \Big/ \sum_{j=1}^{N} K_j \qquad (7)$$

where K_j is the number of syllables in phrase b_j .

The procedure is iterated until the mean phrase-generation probability stops increasing.

5. PHONETIC TYPEWRITER

The phonetic typewriter system is based on the HMM-LR continuous speech recognizer.[2,4] The predictive LR parser generates legal Japanese phone sequences from left-to-right and forms an HMM corresponding to their phonetic representation.

The phone sequence probability is calculated from the HMM probability and the sequence generation probability.

$$\ln P^{(total)} = 0.75 \times \ln P^{(hmm)} + 0.25 \times \ln P^{(gram)} \qquad (8)$$

$P^{(hmm)}$ is the HMM probability of the generated phone sequence.

$P^{(gram)}$ is the sequence generation probability calculated by the recurrence formulas.

Let the syllable sequence be s_0 ... s_{i-1} s_i . If the recognized phone is a vowel, calculate the formula:

$$\ln P_i^{(gram)} = \frac{(i-1)\ln P_{i-1}^{(gram)} + \ln P(s_i | \ldots, s_{i-2}, s_{i-1})}{i} \qquad (9\text{-}1)$$

If the recognized phone is a consonant, calculate

$$\ln P_i^{(gram)} = \frac{(i-1)\ln P_{i-1}^{(gram)} + \sum_{s} \ln P(s_i | \ldots, s_{i-1})}{i} \qquad (9\text{-}2)$$

where Σ is the sum over all syllables which contain the recognized consonant.

The system uses a beam search. At each recognition state, the LR parser predicts the next phones and concatenates them onto phone sequence candidates. Consequently, the number of possible sequences explodes quickly. The system limits the number of candidates (beam width) using the phone sequence probabilities calculated by Equation 8. Only 250 phone sequence candidates are maintained in the system. This permits an acceptable computation time.

6. RECOGNITION EXPERIMENTS

The phonetic typewriter system is tested by phone recognition and phrase recognition experiments using 25 sentences containing 279 phrases uttered by a male speaker. The sentences are collected through simulation of a secretarial service of an international conference. Therefore, the contents are conversational. The speech is analyzed under the same conditions as in section 2.

We define the phone recognition rate taking phone substitutions, insertions, and deletions into account. First, a recognized phone sequence is compared with the reference phone sequence using DP matching.

```
     ----S-----------------------D---------------I-----
Ref: y   o   k   o   u   sy  u   u   d   a   i       o
Out: y   u   k   o   u   sy  u       d   a   i   y   o
     -----------------------------------------------------
        S:Substitution    D:Deletion    I:Insertion
```

After counting the number of substitutions, insertions, and deletions, calculate the phone recognition rate;

$$Rate = \frac{N_{phones} - N_{subst} - N_{ins} - N_{del}}{N_{phones}} \qquad (10)$$

where

N_{phones} : the total number of phones

N_{subst} : the number of substitutions

N_{ins} : the number of insertions

N_{del} : the number of deletions.

The recognition experiments are carried out under two different trigram training conditions.
(A) Training syllable trigrams with the written and spoken language text databases (blocks 1,2,3,4,5).
(B) Training syllable trigrams with only the written language text databases (blocks 1,2,3).

Two sequence models are used in each experiment.
(1) Syllable trigram model with 1.e-5 flooring.
(2) 0,1,2,3-gram models weighted by D.I.

Table 3 shows the phone and phrase recognition scores under the condition (A). The table contains the mean phrase-generation probability, phone perplexity, phone recognition rate, and five best cumulative phrase recognition rates. Generally, D.I. gets better results than the simple flooring method. Consequently, 94.9% phone recognition and 78.5% phrase recognition rates were achieved without lexical information.

Table 4 shows the recognition scores under the condition (B). The recognition rates were uniformly lower, proving that the training with spoken language is necessary for recognizing conversational speech. Even with this constraint, phone recognition rates of over 80% were achieved. Again, D.I. improves the performance.

7. CONCLUSIONS

A Japanese phonetic typewriter was constructed. The key technologies are accurate phone recognition using HMMs, and phone sequence modeling based on syllable trigrams. Using a spoken language text database and deleted interpolation to train the syllable trigrams, enabled a high recognition rate without lexical information.

ACKNOWLEDGEMENT

This research was performed while the authors were at the ATR Interpreting Telephony Research Laboratories. The authors wish to thank Dr. Akira Kurematsu, President of the ATR Interpreting Telephony Research Laboratories, for his guidance.

REFERENCES

[1] Levinson,S.E., Rabiner,L.R., and Sondhi,M.M.: "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech recognition", BSTJ, Vol.62, No.4, pp.1035-1074 (1983)
[2] Hanazawa,T., Kita,K., Nakamura,S., Kawabata,T., and Shikano,K.: "ATR HMM-LR Continuous Speech Recognition System", ICASSP90, S2.4, pp.53-56 (1990)
[3] Araki,T., Murakami,J., and Ikehara,S.: "Effect of Reducing Ambiguity of Recognition Candidates in Japanese Bunsetsu units by 2nd-order Markov Model of Syllables", Trans. of Inf. Processing Soc. Japan, Vol.30, No.4, pp.467-477 (1989)
[4] Kita,K., Kawabata,T., and Saito,H.: "HMM Continuous Speech Recognition Using Predictive LR Parsing", ICASSP89, S13.3, pp.703-706 (1989)
[5] Kawabata,T., and Shikano,K.: "Island-driven Continuous Speech Recognizer Using Phone-Based HMM Word Spotting", ICASSP89, S9.7, pp.461-464 (1989)
[6] Jelinek,F., and Mercer,R.: "Interpolated Estimation of Markov Source Parameters from Sparse data", Pattern Recognition in Practice, pp.381-397, E.S.Gelsema and L.N.Kanal, ed., North-Holland Publishing Company (1980)

Table 3 Phone and phrase recognition scores (Condition A: Syllable trigram models trained using the written and spoken language text databases)

syllable sequence model	phrase generation probability (log scale)	phone perplexity	phone recognition rate (%)	phrase recognition rate (%)				
				1	~2	~3	~4	~5
trigram (with 1.e-5 flooring)	-2.08	4.2	94.0	76.7	81.4	86.3	87.5	88.5
0,1,2,3-gram & Deleted Interpolation	-1.98	3.9	94.9	78.5	86.7	90.7	91.8	92.1

Table 4 Phone and phrase recognition scores (Condition B: Syllable trigram models trained using the written language text databases)

syllable sequence model	phrase generation probability (log scale)	phone perplexity	phone recognition rate (%)	phrase recognition rate (%)				
				1	~2	~3	~4	~5
trigram (with 1.e-5 flooring)	-3.30	9.8	82.8	46.2	52.3	55.2	57.3	58.8
0,1,2,3-gram & Deleted Interpolation	-2.68	6.4	84.6	51.3	62.4	65.6	68.5	70.6