2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

THE SSS-LR CONTINUOUS SPEECH RECOGNITION SYSTEM: INTEGRATING SSS-DERIVED ALLOPHONE MODELS AND A PHONEME-CONTEXT-DEPENDENT LR PARSER

Akito NAGAI *Jun-ichi TAKAMI* and *Shigeki SAGAYAMA*

ATR Interpreting Telephony Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

ABSTRACT

This paper describes a system for accurate continuous speech recognition called "ATREUS/SSS-LR". A phoneme-context-dependent LR parser drives allophonic HMMs represented by a shared-state network automatically generated by the Successive State Splitting (SSS) algorithm. In this system, the SSS principle has also been applied to duration clustering: optimal clusters of phoneme-context-dependent durations are automatically generated independently of the HMnet-based allophonic classes. ATREUS/SSS-LR achieved a phrase recognition rate of 93.2%, the best recognition result achieved in the 1000-word recognition experiments conducted at ATR. This recognition rate was obtained with a smaller beam width than used with discrete HMM(fuzzy vector quantization) and continuous mixture density HMM. This shows that the SSS-LR can realize both fast parsing and high accuracy.

1 INTRODUCTION

Recently, continuous speech recognition systems using allophonic HMMs have significantly improved performance [1][2][4]. This paper proposes a continuous speech recognition system called "SSS-LR" based on strategies of the phoneme-context-dependent modeling and parsing.

In allophonic modeling, it is very important to attain the most precise and robust model-set under the limitation of training samples [3][4]. To obtain a viable solution to this problem, we have proposed a Successive State Splitting (SSS) algorithm that automatically generates an efficient representation of allophonic continuous density HMMs, which is called a Hidden Markov Network (HMnet) [9]. SSS simultaneously finds the optimal set of phonetical context classes as well as, the optimal topology and optimal parameters for HMMs using a maximum likelihood criterion.

The SSS principle has also been applied to duration clustering: optimal clusters of phoneme-context-dependent durations are automatically generated independently of the HMnet-based allophonic classes.

In order to handle allophonic HMMs, we have also proposed phoneme-context-dependent LR parsing algorithms [7] originally based on a generalized LR parser [5]. A phoneme-context-dependent LR parser dynamically predicts the current phonetical context using a phoneme-context-independent LR table. It drives precise allophonic HMMs and duration models by exploiting phonetic context dependancy both inside words and at word junctures.

SSS-LR achieved the best phrase recognition rate of 93.2% among the 1000-word recognition results at ATR. This recognition rate was obtained with a smaller beam width than used with discrete HMM (fuzzy vector quantization) and continuous mixture density HMM. This shows that the SSS-LR can realize both fast parsing and high accuracy.

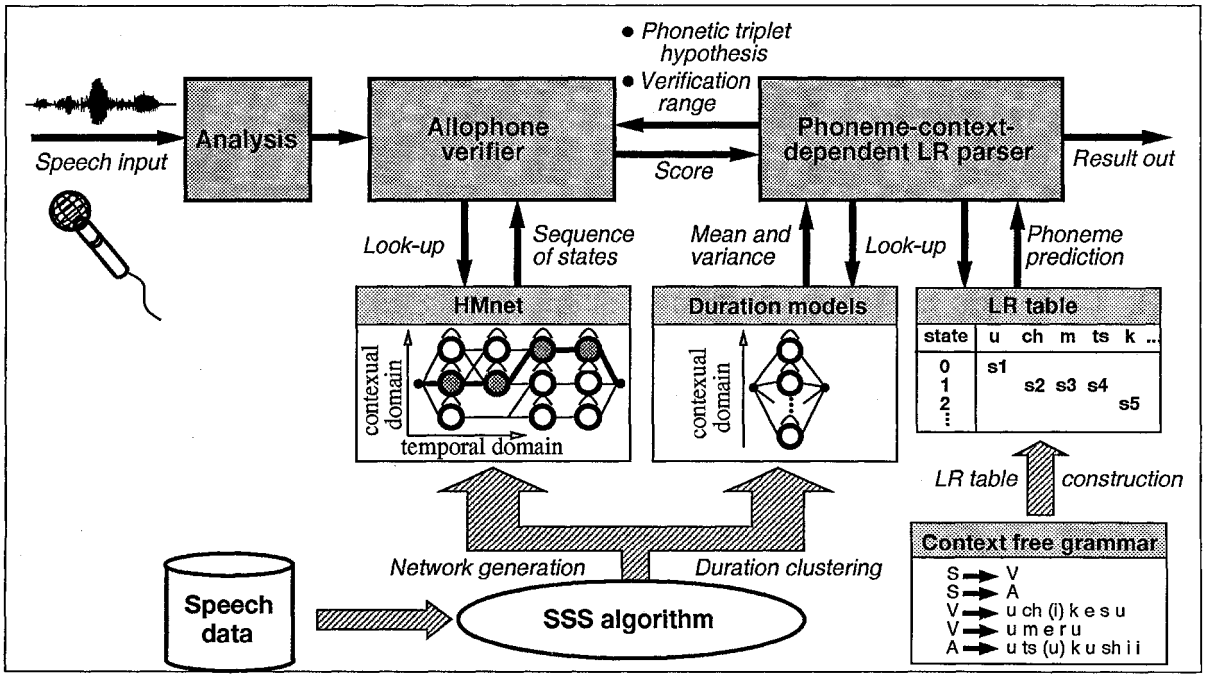

Figure 1. The SSS-LR continuous speech recognition system

This paper first outlines the SSS-LR system.It then describes SSS allphonic modeling for both HMnet and the phonetic duration model and the algorithms of phoneme-context-dependent LR parser. Finally, this paper describes experimental results of continuous speech recognition and compares them with other systems that use allophonic continuous HMMs generated by PEC, discrete HMMs (fuzzy vector quantization) and continuous mixture density HMMs.

2 SYSTEM STRUCTURE

Figure 1 illustrates the functional structure of the SSS-LR continuous speech recognition system. The structure is based on HMM-LR [6], but its recognizer and parser algorithms are entirely different from the viewpoint of phoneme-context-dependency.

In the analysis module, speech was sampled at 12 kHz, quantized to 16 bits, pre-emphasized with a transfer function of $(1 - 0.98z^{-1})$, and windowed using a 20 msec Hamming window with a 5 msec shift. 34-coefficient log-power, 16-channel cepstrum coefficients, delta log-power and 16-channel delta cepstrum coefficients were calculated every 5 ms were used as feature parameters.

In the parsing module, the phoneme-context-dependent LR parser predicts a phonetic triplet hypothesis by reference to an original LR table. It then controls phoneme durations using phoneme-context-dependent duration models. The parser drives allophonic models precisely matched to the phonetic contexts of a parse tree not only in at the intra-word level but also at the word juncture point.

In the allophone verifying module, the sequence of HMM states in the HMnet is selected corresponding to the phonetic triplet that is predicted by the parser, and likelihood is calculated.

3 PHONEME-CONTEXT DEPENDENT HMMS AND SUCCESSIVE STATE SPLITTING ALGORITHM (SSS)

It is known that speech recognition using phoneme-context-dependent HMMs is an effective approach for achieving high recognition performance despite the variations of feature parameters due to differences in phoneme contexts [1][2][3][10]. On the other hand, when the number of models becomes larger by classifying each phoneme context class into more precise divisions, the amount of free parameters increases. Therefore, it is difficult to stochastically estimate accurate phoneme-context-dependent HMMs with limited training samples. To overcome this problem, it is important to reduce the number of useless free parameters from each model as much as possible and to efficiently obtain information of training samples with a smaller number of free parameters.

To generate accurate phoneme-context-dependent HMMs, we proposed implementing a Successive State Splitting algorithm (SSS) [9]. By using the maximum likelihood criterion, SSS can simultaneously and automatically optimize the following three items that are important for constructing phoneme-context-dependent HMMs:

1. the model unit, i.e. the set of phoneme context classes;
2. the model architecture, i.e. the number of states per model and the architecture of state sharing [4];
3. the model parameters, i.e. output probability density distributions and state transition probabilities.

Using this algorithm, an efficient network of phoneme-context-dependent HMMs called a Hidden Markov Network (HMnet) is generated.

In this system, the HMnet is used as a phoneme verifier.

3.1 The Successive State Splitting Algorithm (SSS)

The concept of SSS is to successively make each model more precise by iterating the split of a probabilistic statistical signal source (i.e. a hidden Markov state) into either a phoneme contextual domain or a temporal domain based on the maximum likelihood criterion.

However, to achieve this concept directly, it is necessary to evaluate all possible combinations. Specifically, this implies determining on which state and in which domain a split can realize the maximum likelihood. after actually generating all possible networks. Such a huge computation is not practical in terms of the present computer's capabilities.

Consequently, in the actual algorithm, the following two approximations have been introduced:

1. at each iteration, a state having the largest output probability density distribution is determined as the splittee state;
2. the output probability density distribution of each state is formed as two-mixture Gaussian density distribution, and when a state is split, each of the two Gaussian density distributions of the original state is distributed to one of two new states.

By these approximations, the splittee state and the output probability density distributions of the new states can be determined without any training process. As a result, a great amount of computation has been reduced. Figure 2 shows the principle of SSS.

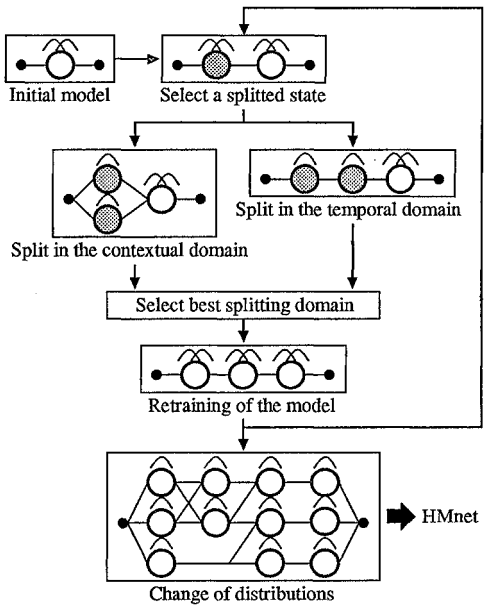

Figure 2. The Successive State Splitting algorithm

3.2 The Hidden Markov Network (HMnet)

The HMnet is a network of multiple hidden Markov states, and each state has the following information:

- state index;
- acceptable contextual class;
- lists of preceding states and succeeding states;
- parameters of the output probability density distribution;
- state transition probabilities.

In the HMnet, if a phoneme context of a sample is given, the model corresponding to the context can be determined by concatenating several states, each of which can accept the context, by applying the restrictions of the preceding state list and the succeeding state list. Since this model is equivalent to a common HMM, we can use the forward-pass algorithm to calculate the likelihoods for input samples as well as for common HMMs.

4 PHONEME DURATION MODELING

The phoneme-context-dependent LR-parser used in this system determines the phoneme verification scope with phoneme duration models. Therefore, in order to achieve efficient parsing, it is important to estimate accurate duration models. Accordingly, accurate duration models are generated with SSS independent of HMnet generation.

For this purpose, SSS can be used without any modification. A set of scalar values, each of which shows the duration of the phoneme, is given to SSS as the training samples. These values are in contract with the set of vectors consisting of *K-dimensions* × *N-frames*, which is used for HMnet generation.

In this case, state splitting by SSS is done only into the phoneme contextual domain because the training samples do not have any temporal structures. Each phoneme duration model consists of one state and a single Gaussian density distribution. An individual phoneme context class is constructed in each state. The mean value and the standard deviation of the single Gaussian density distribution of each state indicate the mean of phoneme duration and its standard deviation, respectively.

5 PHONEME-CONTEXT-DEPENDENT PARSING ALGORITHMS

To realize the phoneme-context-dependent LR parser, we proposed three algorithms as follows (Figure 3):

1. *Parser level realization* [7]

   A modified LR parser predicts the phoneme context dynamically, using a phoneme-context-independent LR table. After the phoneme-context-dependent LR parser hypothesizes a phoneme triplet, it evaluates the likelihood for the current phoneme with the allophonic HMM that corresponds to the phonemic context.

2. *Table level realization* [7]

   An original LR parsing table is converted into a phoneme-context-dependent LR parsing table. This table has sufficient information for the LR parser to handle context-dependent allophonic HMMs without changing the LR parser algorithm.

3. *Grammar level realization* [8]

   Original context independent CFG rules are converted into phoneme-context-dependent rules with nonterminal symbols prepared for allophonic models. In this case, the algorithm can be applied not only to the LR parser but to any parser based on a CFG by replacing this CFG with a phoneme-context-dependent CFG.

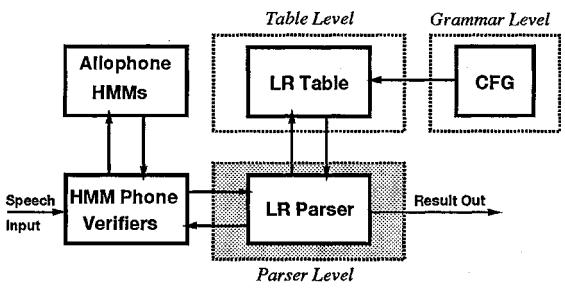

Figure 3. Three Different Phoneme-Context-Dependent LR Parsers

The algorithm at the parser level was implemented in SSS-LR. Figure 4 shows a simple example of dynamically predicting the phonemic context with a phoneme-context-independent LR table. This is done by searching succeeding phonemes at the next state and repeating this process until parsing is complete. After the LR parser hypothesizes a phoneme triplet, it evaluates the likelihood for the current phoneme with the allophonic model which corresponds to the phonemic context.

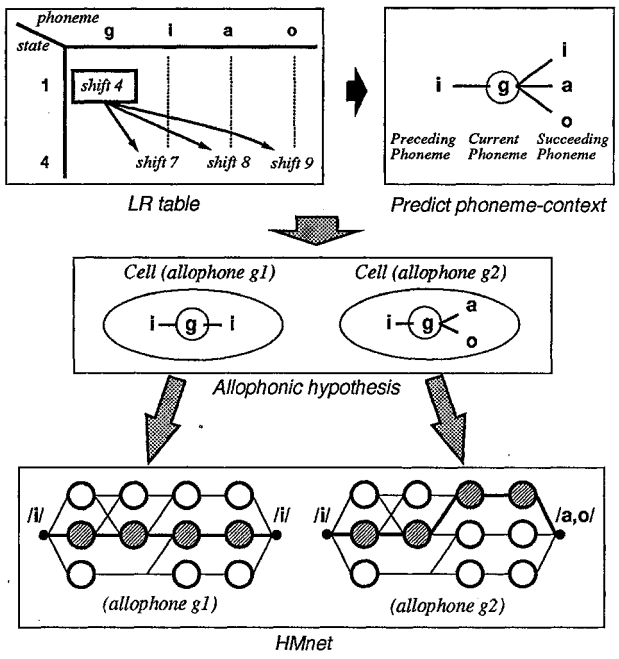

Figure 4. Generating Allophonic Hypotheses

6 EXPERIMENTS

Continuous speech recognition experiments with a 1000 word-vocabulary were carried out.

In these experiments, a diagonal-covariance single Gaussian distribution was used as the output probability density distribution of each state. Isolated words of the 5240 common Japanese word set (5.68 mora/s) were used for training data, and 279 phrase utterances (7.14 mora/s) for testing data (SB3 in the ATR database). Allophones number 300, clustered by Phoneme Environment Clustering (PEC) [10] with a phonemic triplet context. The grammar included 1407 rules,task entropy was 17.0, and phonetic perplexity was 5.9 [11]. The beam-search technique was used.

The results of speaker-dependent speech recognition are shown in Tables 1, 2 and 5. The speaker was a professional announcer (MAU in the ATR database).

Table 1 shows the performances of four speech recogniton systems developed at ATR. PEC-LR is for single gaussian continuous density HMMs based-on PEC, CDHMM-LR for continuous mixture density HMMs and DDHMM-LR for discrete density HMMs. In this experiment, SSS-LR successfully reduced the error rates of first rank 6.8% from the 11.1 % of the CDHMM-LR.

The error rates of Table 2 were achieved by adding phrase utterance data (SB1,2,4) to the word training data. This addition, however, did not include testing data (SB3). This condition provided the adaptation of a speaking rate that compensates for the different rates of words and phrases. The first rank error rate of SSS-LR was reduced to 4.7%.

Figure 5 shows the results of the beam width evaluation. The abscissa is the beam width and the ordinate is the recognition rate. In the figure, the solid lines show the first rank rates, and the dotted lines show the accumulative rates to the fifth rank.

SSS-LR's best recognition rate of 89.3% was obtained with a smaller beam width (32) than used for discrete HMM (fuzzy vector quantization) and continuous mixture density HMM. This shows that fast parsing with high accuracy can be realized because of the high accuracy of allophonic models.

Table 1. Error rates (%)

system	SSS-LR	PEC-LR	CDHMM-LR	DDHMM-LR
HMM	HMnet (single)	CDHMM (single)	CDHMM (mixture)	DDHMM
training	5240 words	5240 words	5240+216* words	5240+216* words
# phone models	1688	300	71	71
rank 1	6.8	13.3	11.1	11.8
~2	1.8	3.2	2.9	1.1
~3	1.1	2.2	1.4	0.7
~4	0.7	1.1	1.1	0.7
~5	0.4	0.7	0.7	0.4

* : 216 phonetically balanced words, CDHMM : Continuous Density HMM, DDHMM : Discrete Density HMM

Table 2. Error rates (%)

system	SSS-LR	CDHMM-LR
HMM	HMnet (single)	CDHMM (mixture)
rank 1	4.7	6.8
~2	0.7	1.4
~3	0.4	1.1
~4	0.0	0.7
~5	0.0	0.7

training: 5240 words+216 phonetically balanced words+phrase(SB1,2,4)
testing: phrase(SB3)

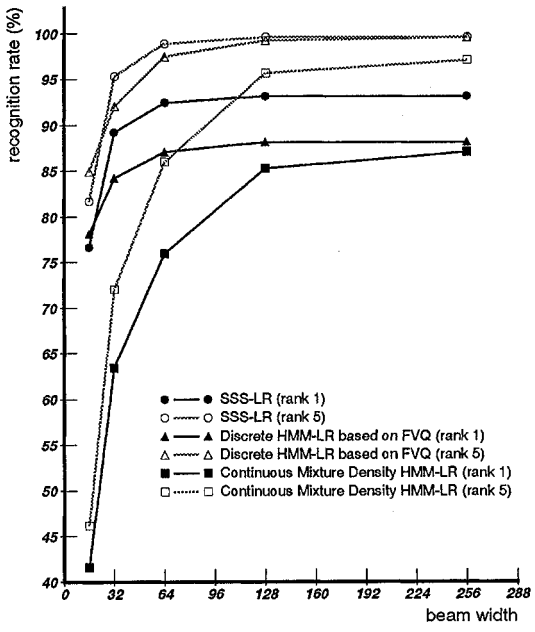

Figure 5. Evaluation of beam width

7 CONCLUDING REMARKS

SSS-LR achieved the best performance with the smallest beam width in the 1000-word recognition experiments conducted at ATR. This was probably due to the high accuracy of the HMnet-based representation.

Future work includes the following enhancements: (1) recognition of sentence utterances by SSS-LR; (2) the use of minimum error classification retraining [12] for more accurate HMMs.

SSS-LR will play an important role in ATREUS, the final speech recognition system of ATR's seven year interpreting telephony project.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Kurematsu, president of ATR Interpreting Telephony Research Laboratories, for his encouragement and support of this research, and all of the members of the *Speech Processing Department* and the *Knowledge and Data Base Department* for their their valuable technical suggestions.

REFERENCES

[1] Y.L. Chow, M.O. Dunham, O.A. Kimball, M.A. Krasner, G.F. Kubala, J. Makhoul, P.J. Price, S. Roucos and R.M. Schwartz, "BYBLOS: The BBN Continuous Speech Recognition System", ICASSP'87, pp.89-92 (1987).

[2] K.F. Lee, H.W. Hon, M.Y. Hwang, S. Mahajan and R. Reddy, "The SPHINX Speech Recognition System", ICASSP'89, pp.445-448 (1989).

[3] K.F. Lee, S. Hayamizu, H.W. Hon, C. Huang, J. Swartz and R. Weide, "Allophone Clustering for Continuous Speech Recognition", ICASSP'90, pp.749-752 (1990.4).

[4] X.D. Huang, K.F. Lee, H.W. Hon and M.Y. Hwang, "Improved Acoustic Modeling with the SPHINX Speech Recognition System", ICASSP'91, pp.345-348 (1991.5).

[5] M. Tomita, "Efficient Parsing for Natural Language: A Fast Algorithm for Practical Systems", Kluwer Academic Publishers (1986).

[6] K. Kita, T. Kawabata and H. Saito, "HMM Continuous Speech Recognition Using Predictive LR parsing", ICASSP'89, pp.703-706 (1989).

[7] A. Nagai, S. Sagayama and K. Kita, "Phoneme-context-dependent LR parsing algorithms for HMM-based continuous speech recognition", Eurospeech'91 (Genova), 48.3, pp.1397-1400 (1991.9).

[8] A. Nagai, H. Kikuchi, S. Sagayama and K. Kita, "An Algorithm of Converting CFG into Phoneme-Context-Dependent Grammar", The Acoustic Society of Japan Fall Meeting Proc., 3-1-6, pp.81-82 (1992.3) (In Japanese).

[9] J. Takami and S. Sagayama, "A Successive State Splitting Algorithm for Efficient Allophone Modeling", ICCASP'92, pp.573-576 (1992.3).

[10] S. Sagayama, "Phoneme Environment Clustering for Speech Recognition", ICASSP'89, pp.397-400 (1989).

[11] T. Kawabata, K. Shikano and K. Kita, "Task entropy and Phone Perplexity", The Acoustic Society of Japan Spring Meeting Proc., (1989.3) (In Japanese).

[12] D. Rainton and S. Sagayama, "Optimal Error Criterion Selection For HMM Minimum Error Missclassification Training ", ICSLP'92 (1992.10).