ATREUS: CONTINUOUS SPEECH RECOGNITION SYSTEMS AT ATR INTERPRETING TELEPHONY RESEARCH LABORATORIES

S. Sagayama, M. Sugiyama, K. Ohkura, J. Takami, A. Nagai, H. Singer, H. Hattori(†),
K. Fukuzawa, Y. Kato, K. Yamaguchi, T. Kosaka, and A. Kurematsu

ATR Interpreting Telephony Research Laboratories

ABSTRACT — This paper describes **ATREUS**, a family of a large variety of continuous speech recognition systems developed at ATR Interpreting Telephony Research Laboratories as the spoken input front-end of an interpreting telephony system. It is one of the major achievements of a seven-year automatic interpreting telephony project, which will reach its completion at the end of this fiscal year. A comparative study is given from the viewpoints of constituent technique and performance. A combination called ATREUS/SSS-LR performed best among the ATREUS systems.

INTRODUCTION

ATREUS, a family of continuous speech recognition systems, has been developed in an intensive effort to improve speech recognition performance at ATR Interpreting Telephony Research Laboratories in its seven-year interpreting telephony research project, which was initiated in 1986 and will be completed at the end of the fiscal year 1992. Speech recognition has been extensively explored as one of the three major constituent technologies for speech translation, namely, continuous speech recognition, language translation, and speech synthesis. The purpose of this paper is to summarize our research results in continuous speech recognition and to present a comparative study of structures, constituent techniques, hardware implementations, and performances.

ACOUSTIC MODELS (1) — PROBABILISTIC MODELS

ATREUS places particular emphasis on probabilistic approaches for acoustic modeling of speech. A number of acoustic models have been investigated based on HMM. The acoustic parameters are mostly LPC-based cepstral coefficients. Three of the models are as follows:

1. Multiple-codebook Fuzzy VQ-based HMM (FVQHMM) (Hanazawa et al. 1990)
   This model is based on fuzzy vector quantization and multiple codebooks. Acoustic parameters, LPC cepstrum, delta-LPC ceptrum, and delta-power, are separately vector-quantized by using three separate VQ codebooks. Coding is "fuzzy" instead of "hard" decision-type vector quantization to reduce quantization error. This helps overcome the training problem due to sufficient training data being often difficult to obtain. In the early stages of our project, this type of discrete HMM was extensively studied. Eventually, these models were gradually replaced by continuous density HMMs.
2. Continuous Mixture Output Probability Density HMM (CMHMM) (Yamaguchi et al. 1992)
   This is one of the popular models among speech researchers. This is regarded as a reference for comparison with other models. In our models, the number of mixtures are adaptive to the number of available training samples (Kosaka 1992b). The average number is 11.
3. Hidden Markov Network (HMnet) (Takami et al. 1992a)
   This is one our project's major achievements. **Hidden Markov network (HMnet)** is a highly generalized form of HMM, which incorporates context-dependent variations of phones and state sharing among different allophones. A HMnet contains a finite number of states, each containing a single Gaussian distribution, that are connected to each other to form paths representing context-dependent phones. This network is automatically derived using the **Successive State Splitting (SSS)** algorithm, which simultaneously solves three problems: network topology, allophone clusters, and the acoustic distribution of each state. The outline of the SSS algorithm is shown in Fig.1.

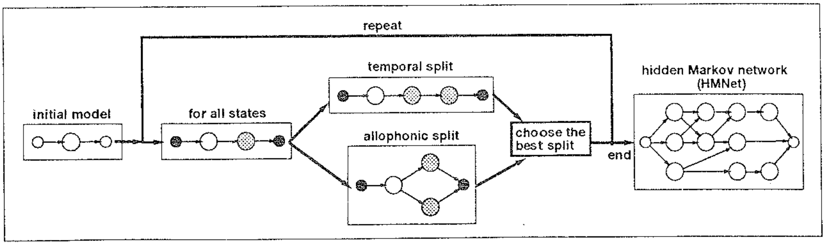

Figure 1: Outline of Successive State Splitting (SSS) algorithm for context-dependent phone modeling

ACOUSTIC MODELS (2) — NEURAL NETWORKS

Neural network approaches have been intensively investigated as alternatives to the probabilistic approaches.

Considerable effort has been paid for improving the "robustness" across differences in speaking styles (e.g., word utterance vs. phrase utterance). Neural fuzzy training, PD-TDNN, neighbor integration, and KNIT (k-nearest neighbor interpolative training) are examples of such attempts.

1. Time-delay Neural Network (TDNN) (Waibel et al. 1989)
   Time-Delay Neural Netwoork (TDNN) is a four-layer perceptron-type neural network which has a special tied-link structure for time-shift capability.
2. Pairwise-Discriminant TDNN (PD-TDNN) (Takami et al. 1991)
   PD-TDNN is a TDNN-based network that consists of pairwise discriminant TDNN trained with three reference values, 0, 0.5, and 1, for all possible phoneme pairs integrated to make a majority decision.
3. Neural-Fuzzy Trained TDNN (Komori 1992)
   Instead of 0 and 1 in the standard TDNN, TDNNs are trained with fuzzy reference levels between 0 and 1 near the phoneme boundaries.
4. Fuzzy Partition Model (FPM) for Phone Modeling (Kato et al. 1992)
   FPM is another neural network architecture based on a probabilistic formulation. Each unit has multiple positive outputs whose sum is unity, unlike the conventional perceptron-type neural networks. This type proved superior to TDNN-based approaches both in training speed and in speech recognition performance.

SPEAKER-ADAPTATIVE/INDEPENDENT SPEECH RECOGNITION

ATREUS includes speaker-dependent, speaker-adaptative, and speaker-independent modes. The techniques for speaker adaptation are:

1. Codebook Mapping (CBM) (Shikano et al. 1986, Nakamura et al. 1989)
   If a pair of utterances of the same word by different speakers are provided, vector quantization codebooks for both speakers are designed independently from them. Using dynamic time warping, these utterances are aligned to each other to obtain a cross-histogram matrix that associates the pair of codebooks with each other.
2. Vector Field Smoothing (VFS) for discrete HMMs (Hattori et al. 1992)
   Vector Field Smoothing (VFS) is a principle for speaker adaptation which assumes that the difference between the reference and the new speakers can be modeled as a vector field in the feature vector space. In the discrete case, the VQ codebook of the reference speaker is transferred according to the vector field. The outline is shown in Fig. 2.
3. Vector Field Smoothing (VFS) for continuous HMMs (Ohkura et al. 1992, Takami et al. 1992b)
   VFS can be considered a principle for training phone models that uses a limited amount of training data along with its phonetic transcription where the set of phone models of a reference speaker is already trained with a large amount of data. In the continuous HMM and HMnet cases, continuous distributions are modified according to the smoothed transfer vector field between the mean vectors of HMMs before and after Baum-Welch embedded training with the given transcribed speech data.

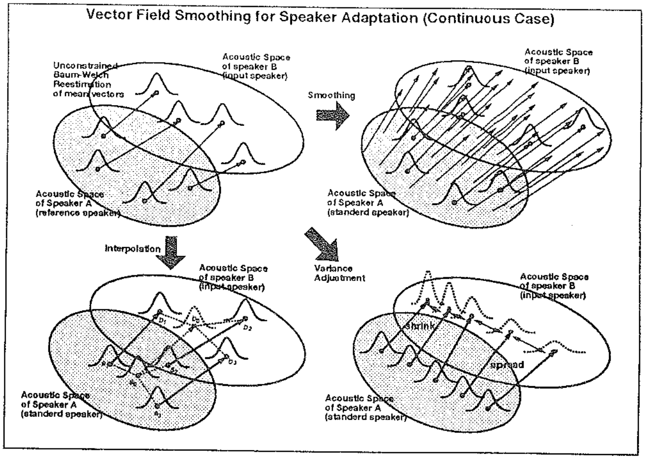
Vector Field Smoothing for Speaker Adaptation (Continuous Case)
Unconstrained Baum-Welch Reestimation of mean vectors
Acoustic Space of speaker B (input speaker)
Acoustic Space of Speaker A (reference speaker)
Smoothing
Acoustic Space of speaker B (input speaker)
Acoustic Space of Speaker A (standerd speaker)
Interpolation
Acoustic Space of speaker B (input speaker)
D1
D0
D2
D3
Acoustic Space of Speaker A (standerd speaker)
Variance Adjustment
Acoustic Space of speaker B (input speaker)
shrink
spread
Acoustic Space of Speaker A (standerd speaker)

Figure 2: Outline of Vector Field Smoothing (VFS) for speaker adaptation

4. Speaker-tied training (Kosaka et al. 1992a)
   Even one word utterance training improves the performance in **speaker-tied training** for a speaker mixture speaker-independent HMnet.
5. Segmental Speaker Mapping using Neural Network (Fukuzawa et al. 1992)
   A neural network is used to define a segment-to-segment mapping from the input speaker to the reference speaker. It is connected to a standard TDNN phone verifier.

Speaker-independent speech recognition approaches are also investigated. They are:

1. Fuzzy Vector Quantization HMM as the baseline system
   A speaker-independent speech recognition system based on fuzzy vector quantization HMM has been developed as the baseline system for comparison.
2. Continuous mixture HMM (Kosaka et al. 1992a)
   In continuous output mixture density HMMs, the number of mixture components are adjusted by the "variance uniformty" principle (Kosaka et al. 1992a) for representation of speaker-independent phone models.
3. Speaker-mixture HMnet (Kosaka et al. 1992b)
   Speaker mixture is a context-dependent, speaker-independent hidden Markov network that is composed of speaker-dependent HMnets each derived from the SSS algorithm. This has attained the highest performance among our speaker-independent models.
4. Speaker-independent trained FPM (Kato et al. 1992b)
   FPM also performs well in the speaker-independent mode if trained with multiple speaker data.

LANGUAGE MODELS

Language models are also investigated from three different approaches, namely, syntactic, stochastic, and their combination. Some of them are:

1. Generalized LR Parser
   A **generalized LR parser** (Tomita 1987) is extensively used in combination with various acoustic models. The grammar is written in a context-free grammar style and used in the LR parser which is combined with both HMM- and neural network-based phone models (Kita et al. 1990, Nagai et al. 1992, Sawai 1990). The parsing process consists of LR table look-up, phoneme verification, and hypotheses pruning. Two of its major advantages are that it can obtain multiple recognition hypotheses as the result and that explicit phoneme duration control is possible. An outline of LR

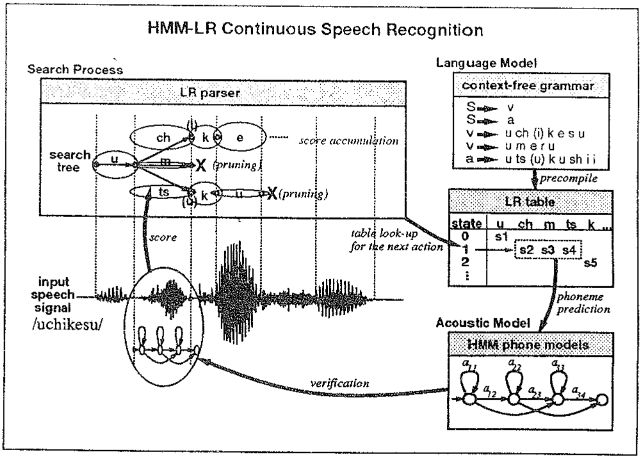

Figure 3: HMM-LR continuous speech recognition

parsing for speech recognition is illustrated in Fig. 3.

2. Stochastic grammar approaches
   Bigram and other statistical approaches to language modeling are investigated and combined with phone models, although they are not treated here.

The combination of fuzzy VQ HMM and the LR parser has already been successfully implemented on a devoted hardware consisting of 33 DSP chips. This system is capable of 1,000-word vocabulary real-time continuous speech recognition.

SYSTEM PERFORMANCES

ATREUS, a family of various combinations of the above mentioned components, has been evaluated for Japanese Bunsetsu (phrase) speech recognition in a task domain of "international conference registration." This domain contains an approximately 1,500 word vocabulary and phoneme perplexity of 5.9 (equivalent word perplexity of over 100). The recognition rates are shown in Fig. 4.

CONCLUSION

Among the ATREUS family systems, **SSS-LR continuous speech recognition** provided the best performance in both speaker-dependent and speaker-adaptive modes. In the speaker-independent mode, **speaker mixture HMnet** performed better than the other methods. These were all based on phoneme-context dependent phone models combined with a phoneme-context-dependent LR parser.
Even though the performances did not exceed the best performance of HMM, neural network approaches have shown high potential. Future works will include the combination of neural networks and HMM approaches.

ACKNOWLEDGEMENT

The ATREUS family is a compilation of long continuous research effort involving the authors and a number of other researchers: Kiyohiro Shikano, Hisao Kuwabara, Tetsuo Umeda, Takeshi Kawabata, Shin-ichi Tamura, Hidefumi Sawai, Kazuya Takeda, Satoshi Nakamura, Masanori Miyatake, Kaichiro Hatazaki, Toshiyuki Hanazawa, Katsuteru Maruyama, Masami Nakamura, Yasuhiro Komori, Kazuki Katagishi, Shingo Fujiwara, Jin-ichi Murakami, Yasunaga Miyazawa, Ryosuke Isotani, Alex Waibel, Hiroaki Saito, David Rainton, Helmut Lucke, Patrick G. Haffner, and Yasuhiro Minami.

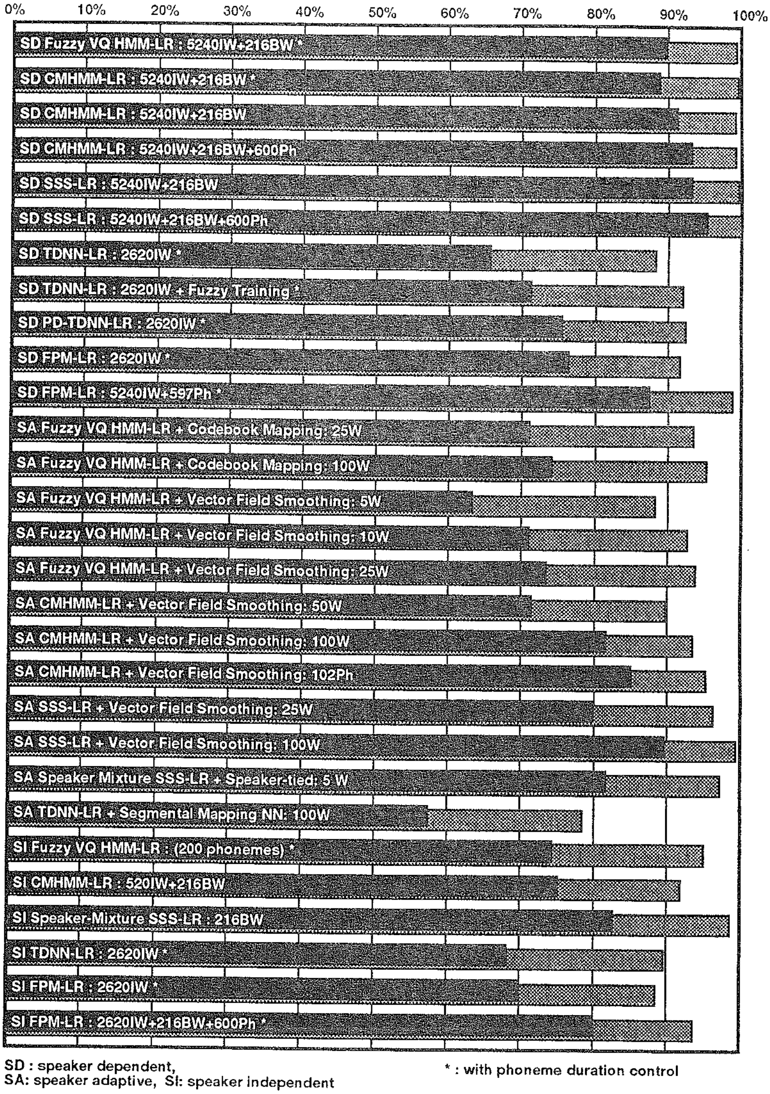

Figure 4: Summary of Speaker-dependent, adaptive, and independent phrase speech recognition performances. (Black and gray bars represent recognition rates for the first candidate and for five candidates. The task domain is international conference registration with phoneme perplexity 5.9. W, IW, and Ph stands for words, isolated words, and phrase utterances, respectively, used for training. '*' means "with phoneme duration control".)

NOTE

(†) H. Hattori is currently with C & C Information Technology Research Laboratories, Nippon Electric Corporation.

REFERENCES

The following references are mostly English papers, and it should be noted that most of them are preceded by earlier reports in Japanese by the same authors.

Fukuzawa, K., Komori, Y., Sawai, H. and Sugiyama, M. (1992) "A Segment-based Speaker Adaptation Neural Network Applied to Continuous Speech Recognition," Proc. ICASSP92 (San Francisco), Vol.1, pp.433–436 (1992.3).

Hanazawa, T., Kita, K., Nakamura, S., Kawabata, T. and Shikano, K. (1990) "HMM-LR Speech Recognition System," Proc. ICASSP90 (Albuquerque), pp. 53 – 56.

Hattori, H. and Sagayama, S. (1992) "Vector Field Smoothing Principle for Speaker Adaptation," Proc. ICSLP92 (Banff), We.fPM.1-4, to appear.

Kato, Y. and Sugiyama, M. (1992a) "Fuzzy Partition Models and their effects in continuous speech recognition," IEEE Workshop Neural Networks for Signal Processing (Elsinore), pp. 111–120.

Fukuzawa, K., Kato, Y., and Sugiyama, M. (1992b) "A Fuzzy Partition Model Neural Network Architecture for Speaker-Independent Continuous Speech Recognition," Proc. ICSLP92 (Banff), Th.PM.P-14, to appear.

Kita, K., Kawabata, T., and Hanazawa, T. (1990) "HMM Continuous Speech Recognition Using Stochastic Language Models," Proc. ICASSP90 (Albuquerque), pp.581–584.

Komori, Y. (1992) "Neural Fuzzy Training Approach for Continuous Speech Recognition Improvement," Proc. ICASSP92 (San Francisco).

.Kosaka, T., Takami, J. and Sagayama, S. (1992a) "Speaker-independent and Speaker-adaptive Speech Recognition Using Speaker-mixture SSS," IEICEJ Technical Report, Sept. 1992 (in Japanese).

Kosaka, T. and Sagayama, S. (1992b) "An Algorithm for Automatic HMM Structure Generation in Speech Recognition," Proc. SST92 (Brisbane), to appear.

Nagai, A., et al. (1992) "Hardware Implementation of Realtime 1000-word HMM-LR Continuous Speech Recognition," Proc. ICSLP92 (Banff), We.fAM.1-4, to appear.

Nagai, A., Takami, J. and Sagayama, S. (1992) "The SSS-LR Continuous Speech Recognition System: Integrating SSS-derived Allophone Models and a Phoneme-Context-Dependent LR Parser," Proc. ICSLP92 (Banff), We.fAM.1-4, to appear.

Nakamura, S. and Shikano, K. (1989) "Spectrogram normalization using fuzzy vector quantization," Journal of Acoust. Soc. of Japan, Vol.45, No.2, pp.107–114.

Ohkura, K., Sugiyama, M. and Sagayama, S. (1992) "Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs," Proc. ICSLP92 (Banff), We.fPM.1-1, to appear.

Sawai, H. (1990) "The TDNN-LR Large-Vocabulary and Continuous Speech Recognition System," Proc. ICSLP90, S31.4, pp.1349–1352.

Shikano, K., Lee, K-F., and Reddy, R. (1986) "Speaker Adaptation through Vector Quantization," Proc. ICASSP86, 49.5, pp.2643–2646.

Takami, J., Sagayama, S. and Kai, A. (1991) "Speech Recognition by Combining Pairwise Discriminant Time-Delay Neural Networks and Predictive LR-Parser," Proc. NNSP91 (Princeton), pp.327-336.

Takami, J. and Sagayama, S. (1992a) "A Successive State Splitting Algorithm for Efficient Allophone Modeling," Proc. ICASSP92 (San Francisco), 66.6.

Takami, J., Nagai, A. and Sagayama, S. (1992b) "Speaker Adaptation of the SSS (Successive State Splitting)-Based Hidden Markov Network for Continuous Speech Recognition" Proc. SST92 (Brisbane), to appear.

Tomita, M. (1987) "An Efficient Augmented-Context-Free Parsing Algorithm," Computational Linguistics, Vol.13, No.1-2, pp.31–46.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. (1989) "Phoneme Recognition Using Time-Delay Neural Networks," IEEE Trans. on ASSP, Vol.37, No.3, pp.328–329 (1989.3).

Yamaguchi, K. and Sagayama, S. (1992) "Continuous Mixture HMM-LR Using the A* Algorithm for Continuous Speech Recognition," Proc. ICSLP92 (Banff), We.sAM.1-2, to appear.