3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

ATREUS: a Speech Recognition Front-end for a Speech Translation System

Shigeki SAGAYAMA[1], *Jun-ichi TAKAMI*[2], *Akito NAGAI*[3], *Harald SINGER*[2],
Kouichi YAMAGUCHI[2], *Kazumi OHKURA*[4], *Kenji KITA*[5], *and Akira KUREMATSU*
ATR Interpreting Telephony Research Laboratories
Seika-cho, Soraku-gun, Kyoto 619-02 Japan

Abstract

This paper describes the continuous speech recognition subsystem "ATREUS" which is used as the speech input stage in the experimental speech translation system "ASURA." The speech recognition algorithm is SSS-LR/VFS which consists of context-dependent phone models (HMnet), a generalized LR parser, and vector field smoothing for speaker / environment adaptation.
Keywords: continuous speech recognition, context-dependent phone models, LR parser, speaker adaptation, speech translation

1 Introduction

Automatic interpreting telephony ("speech translation", in more general terminology), which allows speech communication between people speaking different languages, is one of the technologies most eagerly awaited by people throughout the world. ATR Interpreting Telephony Research Laboratories, a subsidiary of Advanced Telecommunications Research International, was established in 1986 to conduct basic research into speech translation (automatic interpreting telephony). Recently, the number of research organizations active in this research area has increased to the point at which mutual interconnection is feasible. On January 28, 1993, ATR Interpreting Telephony Research laboratories, Carnegie-Mellon University, and Siemens AG + Karlsruhe University successfully performed the first international joint experiment of interpreting telephony.

This paper describes the speech recognition subsystem "ATREUS," the front-end to the speech translation system "ASURA" in the international speech translation experiment.

2 ASURA: a Speech Translation System

2.1 System Outline

To permit the international interconnection of different speech translation systems, the basic components required at each site are speech recognition of own language, language translation from own language to the target, and speech synthesis of own language, as shown as the shaded components in Figure 1.

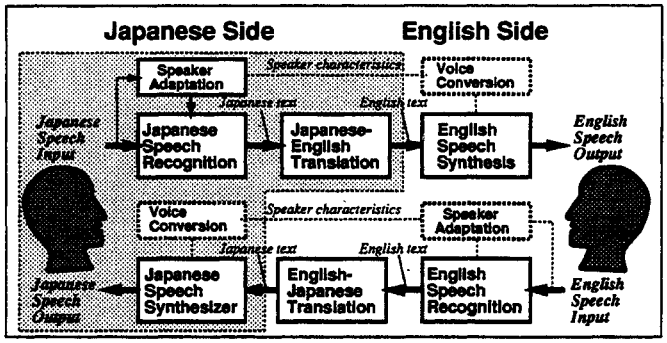

Figure 1: The General Concept of Speech Translation System (Japanese-English)

ASURA is a Japanese-to-English/German speech translation system, and the sentences uttered by the Japanese speaker are recognized by a speaker-adaptive continuous speech recognition subsystem (ATREUS) which will be described in more detail. Speech recognition is followed by language translation which is performed in three parts: analysis of the source language, transfer from source to target languages, and generation of the target language. In the analysis part, the syntactic, semantic, and pragmatic constraints are described in a unification grammar to handle the wide variety of Japanese forms such as anaphoric expressions and indirect requests that appear in Japanese conversations. The English version is transmitted to the English speaking side, and used to synthesize English speech.

Similarly, the English side recognizes the English input speech and translates it into Japanese. ASURA receives the Japanese text and converts it into speech through speech synthesis.

Although not implemented in ASURA yet, modification of speaker characteristics may be required in the future to duplicate the speaker's spectral characteristics in the synthetic speech in the target language.

2.2 Task Domain

Although the ultimate goal of interpreting telephony is universal dialogue in an unlimited domain, the immediate goal should be more feasible, such as a system which is limited to specific, task-oriented areas. We have selected the international conference registration task as a constrained task domain where the dialogue is goal-directed and the expected vocabulary limited to approximately 1,500 words. In the task, an applicant for an international conference and the conference secretariat talk with each other in different languages. In ASURA, the Japanese speech is translated into English and German. This task has a phoneme perplexity of approximately 5.9 and a vo-

[1]Currently with NTT Human Interface Research laboratories, Yokosuka, Japan; [2] Currently with ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan; [3] Currently with Mitsubishi Electric, Oofuna, Japan; [4] Currently with Sanyo Corporation, Osaka, Japan; [5] Currently with Tokushima University, Tokushima, Japan; [6] Currently with Electrical Telecommunications University, Chofu, Tokyo.

Table 1: Some examples of Japanese to English/German translation ('/' means a short pause)

Japanese (Source Language)	English (Target Language 1)	German (Target Language 2)
Moshi moshi.	Hello.	Hallo.
Kaigi ni / mooshikomi tai no desu ga.	I'd like to apply for the conference.	Ich möchte mich zur Konferenz anmelden.
Dono yoo na / tetsuzuki wo / sure ba / yoroshii no deshoo ka.	What kind of procedure should I follow?	Wie soll ich vorgehen?
Wakaranai / ten ga / gozai mashi tara / itsu demo / okiki kuda sai.	If you have a question, please ask me at any time.	Bitte wenden Sie sich an mich jederzeit, wenn etwas unklar ist!

cabulary size of 1,500 Japanese words.

Table 1 shows some speech recognition results for this task.

3 Speech Recognition Subsystem "ATREUS" and Its Algorithm

3.1 Design Philosophy

The major part of the speech recognition algorithm is called SSS-LR/VFS. This solution was reached through a number of comparative studies including both HMM and neural network approaches [1].

In acoustic modeling, we put particular emphasis on the context dependency of phones (or allophonic variations). This context-dependent approach, however, tend to require a large amount of training data to obtain robust models in simple formulation. Here, we use a sophisticated HMM structure called "HMnet"[2] for exploiting context dependency, where allophones are allocated to paths sharing hidden states. Due to its highly sharing structure, this model performs well even with a limited amount of training data. The HMnet is automatically generated and trained using the SSS algorithm which will be described later.

Though speaker-independent speech recognition is desirable, the speaker-adaptive approach seems more practical and advantageous in attaining high recognition performance while the speaker registration load is slight. We consider the basic idea of speaker-independent phone modeling is too naive, because a sentence utterance is uttered by a single speaker and this fact ("speaker consistency") should be exploited in speaker-independent approaches. Moreover, the combination of speaker differences, different ambient noise, microphone characteristics, and environmental acoustic conditions creates very large variations that cannot be covered by a reasonable amount of training data. In ATREUS, we use an algorithm called VFS[3][4][5] for speaker adaptation.

Speech translation systems demand high speech recognition performance, since the recognition results are directly translated into another language without manual correction. On the other hand, we consider that the grammar should cover most of the common expressions in spoken Japanese even within a specific task. We have chosen a general Japanese syntax with a task-specific vocabulary.

To increase performance, we have chosen phrase-by-phrase (*bunsetsu*) utterances as the speech input instead of sentence speech utterances. Intra-phrase and inter-phrase syntaxes carefully written by linguists are used in the generalized LR parser and N-best hypotheses are obtained.

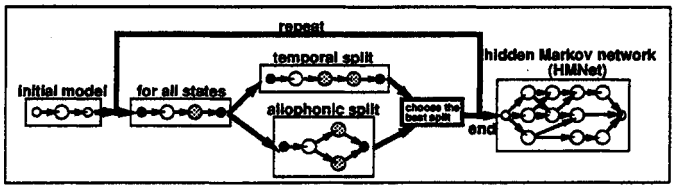

Figure 2: SSS algorithm to automatically obtain an Hidden Markov Network

3.2 Acoustic Analysis

Acoustic analysis is the normal LPC-based cepstrum analysis. The speaker utters input sentences with pauses between phrases. The sentence speech is sampled at 12 kHz and divided into phrase utterances. Each phrase speech is analyzed with a 20 ms frame length and 10 ms shift. The resulting feature vector sequence consists of 16 cepstral and delta-cepstral coefficients, and power and delta-power components.

3.3 Context-dependent Acoustic Model

The acoustic model used in this system is a Hidden Markov Network (HMnet). This includes three major ideas: allophone clusters each represented by a path connecting hidden states, shared states among different allophone paths, and the Successive State Splitting (SSS) algorithm for automatically determining the network topology.

In acoustic modeling, context-dependent models (or allophonic models) have already been intensively studied and proved quite effective in speech recognition. However, since the variety of context-dependent phones become very large, they tend to require a large amount of training data. Even in Japanese, possible phoneme triplets have over 3,000 varieties, and the training data is often quite insufficient to cover the whole variety of all possible triplets.

As some triplets may be acoustically similar, a smaller number of allophone classes can effectively cover the whole variety. This is the need for allophone clustering. Furthermore, HMM states of the allophone classes may be shared, if the output probability densities are similar[6]. The number of free parameters can be reduced by allophone clustering and state sharing. The resulting HMM topology is a network which contains hidden states associated with output probability densities and paths representing allophonic classes. The authors call this a hidden Markov network (HMnet). Both allophone clustering and state sharing are advantageous in enhancing acoustic modeling robustness, since the required degree of freedom of the model is rather small.

The acoustic model used in ATREUS is a single HMnet which can be considered a kind of highly sophisti-

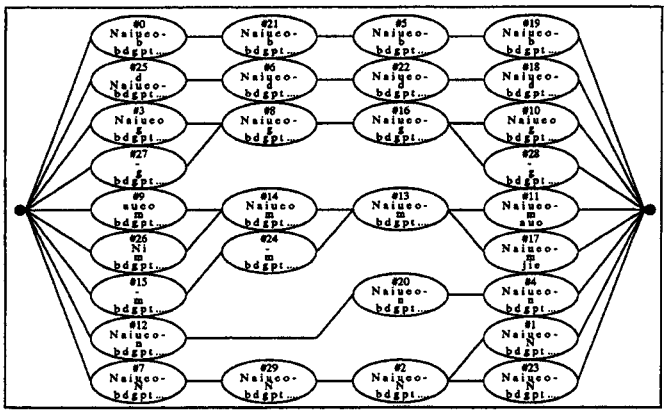

Figure 3: An example of hidden Markov network for 6 consonants /b, d, g, m, n, N/; each state is assocoated with the state number, preceding, center, and following phonemes

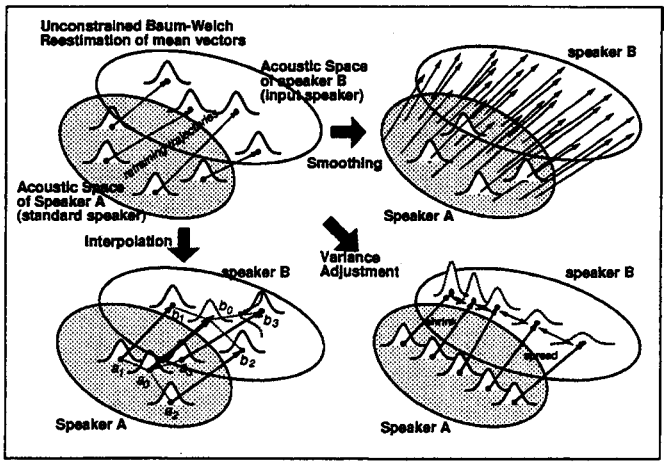

Figure 4: Vector Field Smoothing for adaptation to speaker, environment, microphone, and noise

cated HMM. It contains a number of hidden states, each of which is associated with an output probability density, and are linked by transition probabilities. An allophone or an allophone group is represented by a path across the hidden states. For example, as many as 1,700 different allophone classes can be represented by only 600 states, each of which is represented by a single Gaussian distribution of the 34 dimensional input vector.

The HMnet topology is automatically generated by the Successive State Splitting (SSS) algorithm and the model parameters are estimated at the same time. This algorithm is applied to phonetically labeled single-speaker speech data because the difference among speakers might be larger than allophonic variability so that allophonic variation may be masked as signal masked by noise.

Figure 2 shows the outline of the SSS algorithm. Initially, all training tokens are modeled by a single state. It is then split into two states by either temporal or allophonic splitting to better represent the whole set of tokens in terms of maximum likelihood. All states are tested in respect to two types of split, allophonic and temporal, and the best state and split type is chosen and applied to the network. Incrementing the number of states one by one yields a network that represents each allophone cluster by a path across multiple states where one state may be shared by different paths. This algorithm is a stepwise approximation of over-all optimization of allophone clustering, state sharing, and parameter estimation. The total number of states is determined experimentally since it is a trade-off between robust modeling with a limited amount of training data and precise classification of allophonic variations.

Figure 3 shows an example of HMnet topology generated by SSS for six consonants.

The basic idea of the SSS algorithm is applicable to other similar problems. Actually, in ATREUS, context-dependent phoneme duration was clustered and estimated with the SSS algorithm.

3.4 Vector Field Smoothing for Speaker/ Environment Adaptation

Fast speaker/environment adaptation is one of the features of this system. Vector Field Smoothing (VFS), a speaker adaptation algorithm, is well suited for adapting the HMnet and thus as few as 10 words is effective to adapt the system to a new speaker, the microphone, and the room noise at the same time.

Assume that only a limited amount of training data is given but a set of good initial acoustic models is available. Speaker adaptation is one such case where the initial model is derived from a large amount of training data from another speaker. Vector Field Smoothing (VFS) assumes that the difference between speakers can be considered as a smooth vector field in the acoustic feature space.

VFS speaker adaptation is performed as follows. Given the new speaker's arbitrary utterance with its phonetical transcription, mean vectors of constituent distributions in the HMnet are retrained through embedded Baum-Welch training using the initial models. The difference between mean vectors before and after retraining is derived and the difference vector field is spatially smoothed to interpolate untrained mean vectors, correct the retrained mean vectors, and adjust covariances according to local space shrinkage and spread. This is conceptually illustrated in Figure ??.

This algorithm is performed in the cepstral multidimensional space. The difference between linear microphone characteristics can be normalized by parallel shifting the acoustic models in the cepstrum multidimensional space since a linear system represented by convolution in the time domain is transformed into multiplication in the spectral domain, and addition in the logarithmic spectral domain and the cepstral domains. The spectral change from noiseless to noisy environments is also supposed to be continuous and smooth in the cepstral space. Thus, VFS adapts the system to the new speaker, new microphone, new noise environment, and some other factors affecting the speech spectra.

3.5 Generalized LR Parser for Context-free Grammars

Linguistic constraints are described by a two-level context-free sentence grammar consisting of intra-phrase and inter-phrase grammars. The vocabulary size is 1,000 words (ultimately 3,400words). A generalized LR parser searches the acoustically most likely word sequence for the input

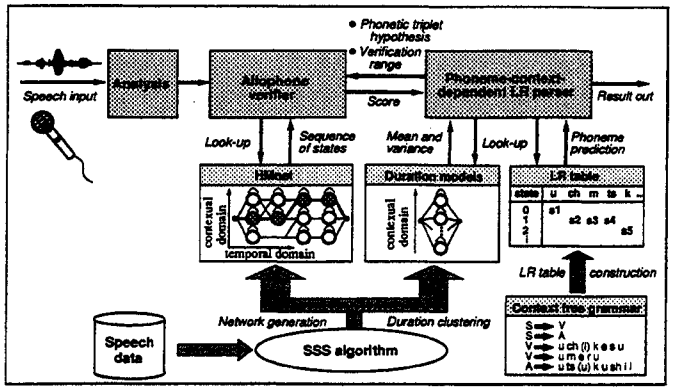

Figure 5: SSS-LR continuous speech recognition system

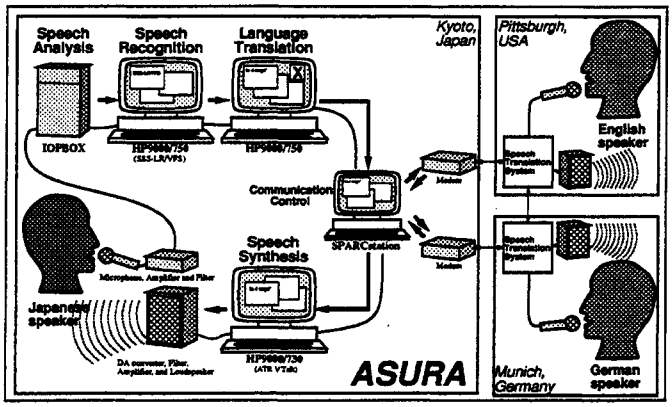

Figure 6: ASURA – ATR's experimental interpreting telephone prototype system connected with other systems in USA and Germany

phrase speech within the phrase grammar based on the beam search technique which produces multiple candidates [7]. To combine the HMnet derived from the SSS algorithm and the LR parsing procedure, the parser has been modified to handle phoneme-context-dependent acoustic models and is called the SSS-LR continuous speech recognition algorithm. Figure 5 shows the outline of SSS-LR algorithm [8].

With the VFS algorithm, the total SSS-LR/VFS algorithm attains both fast speaker/environment adaptability and high continuous speech recognition performance. Inter-phrase grammar is used to construct sentence hypotheses from multiple phrase recognition candidates. The multiple sentences are further examined by semantic analysis before language translation.

3.6 System implementation and performance

Figure 6 shows the ASURA system as interconnected to the equivalent English and German systems. A DSP-based front-end processor calculates the output probabilities of hidden states in the HMnet for each frame in real time and transfers them to a workstation (HP9000/750) which performs LR parsing and inter-phrase post processing to achieve a processing time near real-time.

System performance including language translation is shown in Table 2. 25 words were used for speaker adaptation and 259 sentences were recognized and translated. The linguistic processing part chooses only semantically meaningful sentences among the recognition hypotheses,

Table 2: Speech recognition and translation performances to English

speaker	phrase recognition (%)			sentence recognition (%)			translation accuracy (%)
	1	~2	~3	1	~2	~3	
MIK	91.3	96.2	97.2	84.9	90.3	91.1	90.7
MST	87.2	93.2	95.6	76.4	83.4	86.9	86.1
FAK	95.3	97.4	98.6	90.7	94.2	95.0	91.1
FNY	87.9	95.5	96.4	77.6	89.2	91.0	86.1

and this results in higher translation accuracy than is possible with just the first candidate. This suggests the viability semantic postprocessing multiple candidates.

4 Conclusion

This paper described "ATREUS" the speech recognition subsystem used as the front-end in the ASURA speech translation system. The major algorithm is SSS-LR/VFS which is a speaker-adaptive continuous speech recognition based on context-dependent acoustic models and a generalized LR parser. It was implemented on a DSP-based hardware and a workstation and attained high accuracy and speed.

The authors thank all members of the ATR Interpreting Telephony Research Laboratories for their comments and suggestions in enhancing the ASURA system.

References

[1] A. Nagai, K. Yamaguchi, S. Sagayama, and A. Kurematsu, "ATREUS : A Comparative Study of Continuous Speech Recognition Systems at ATR," Proc. ICASSP93 (Minneapolis, USA), .

[2] J. Takami, S. Sagayama, "A Successive State Splitting Algorithm for Efficient Allophone Modeling," ICASSP92(San Francisco), 66.6 (1992.3).

[3] H. Hattori, S. Sagayama: "Vector Field Smoothing Principle for Speaker Adaptation," Proc. of 1992 International Conference on Spoken Language Processing, pp. 381–384 (1992.10).

[4] K. Ohkura, M. Sugiyama and S. Sagayama: "Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs," Proc. of 1992 International Conference on Spoken Language Processing, pp. 369–372 (Oct.1992).

[5] J. Takami, A. Nagai, S. Sagayama: "Speaker Adaptation of the SSS (Successive State Splitting)-Based Hidden Markov Network for Continuous Speech Recognition," Proc. of SST92 (Fourth Australian International Conference on Speech Science and Technology) (Brisbane), pp. 437-442, (1992.12).

[6] X. D. Huang, K. F. Lee, H. W. Hon and M. Y. Hwang: "Improved Acoustic Modeling with the SPHINX Speech Recognition System," Proc. ICASSP'91, pp.345-348 (1991).

[7] T. Hanazawa, K. Kita, S. Nakamura, T. Kawabata, and K. Shikano: "ATR HMM-LR Continuous Speech Recognition System," Proc. of ICASSP90, Albuquerque 1990.

[8] A. Nagai, J. Takami, S. Sagayama: "The SSS-LR Continuous Speech Recognition System: Integrating SSS-derived Allophone Models and a Phoneme-Context-Dependent LR Parser," Proc. of 1992 International Conference on Spoken Language Processing, Vol. 2, pp. 1511–1514, Canada (1992.10).