A LARGE-VOCABULARY CONTINUOUS SPEECH RECOGNITION ALGORITHM AND ITS APPLICATION TO A MULTI-MODAL TELEPHONE DIRECTORY ASSISTANCE SYSTEM

Yasuhiro Minami, Kiyohiro Shikano, Osamu Yoshioka, Satoshi Takahashi, Tomokazu Yamada and Sadaoki Furui

NTT Human Interface Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo, 180 Japan

ABSTRACT

This paper describes an accurate and efficient algorithm for very-large-vocabulary continuous speech recognition based on an HMM-LR algorithm. The HMM-LR algorithm uses a generalized LR parser as a language model and hidden Markov models (HMMs) as phoneme models. To reduce the search space without pruning the correct candidate, we use forward and backward trellis likelihoods, an adjusting window for choosing only the probable part of the trellis for each predicted phoneme, and an algorithm for merging candidates that have the same allophonic phoneme sequences and the same context-free grammar states. Candidates are also merged at the meaning level. This algorithm is applied to a telephone directory assistance system that recognizes spontaneous speech containing the names and addresses of more than 70,000 subscribers (vocabulary size is about 80,000). The experimental results show that the system performs well in spite of the large perplexity. This algorithm was also applied to a multi-modal telephone directory assistance system, and the system was evaluated from the human-interface point of view. To cope with the problem of background noise, an HMM composition technique which combines a noise-source HMM and a clean phoneme HMM into a noise-added phoneme HMM was investigated and incorporated into the system.

1. INTRODUCTION

One of the main problems with very-large-vocabulary continuous speech recognition is how to accurately and efficiently reduce the search space without pruning the correct candidate. Our speech recognition system is based on the HMM-LR algorithm [1] which utilizes a generalized LR parser [2] as a language model and hidden Markov models (HMMs) as phoneme models. Applying this algorithm to large-vocabulary continuous speech requires: (1) accurate scoring for phoneme sequences, (2) reduction of trellis calculation, and (3) efficient pruning of phoneme sequence candidates.

For the first requirement, several speech recognition algorithms that calculate the backward trellis likelihood from the end of the utterance, as well as the forward trellis likelihood, have been proposed [3][4]. We also use forward and backward trellis likelihoods for accurate scoring. For the second requirement, we use an adjusting window, which chooses only the probable part of the trellis according to the predicted phoneme. For the third requirement, we use an algorithm for merging candidates which have the same allophonic phoneme sequences and the same context-free grammar states [5]. In addition, candidates are also merged at the meaning level [6].

Speech HMMs are sensitive to incoming noise and this often results in a large decrease in the recognition. One solution is to train HMMs on noisy speech to obtain the corresponding optimum HMMs. For large-vocabulary continuous speech recognition, however, the computation load of this solution becomes too high, because all the HMMs need to be retrained each time the characteristics of the background noise (such as its level) change. Taking inspiration from HMM decomposition [7], we proposed an HMM-composition technique to easily adapt the speech recognition system based on clean-speech HMMs to background noise [8]. This technique is similar to the technique of Nolazco Flores et al. [9] which was investigated independently.

Providing access to directory information via spoken names and addresses is an interesting and useful application of large-vocabulary continuous speech recognition technology in telecommunication networks. Although many systems based on recognizing spoken spelled names are being investigated, it is unreasonable to expect users to correctly spell the names of the persons whose telephone number they want. In addition, there are several sets of letters having similar pronunciations, such as the English E-rhyming letters, and pronunciation of the spelling of another person's names is often unstable, since this is not a familiar action for people. Therefore, it is not easy to correctly recognize alphabetically spelled names, and a more successful approach might be to recognize naturally spoken names, even if the machine has to recognize hundreds of thousand names. We applied our speech recognition technology to a directory assistance system recognizing names and addresses continuously spoken in Japanese. This system was evaluated from the human-machine-interface point of view.

2. SPEECH RECOGNITION ALGORITHM

2.1. Two-Stage LR Parser

Figure 1 shows the structure of our continuous speech recognition system for telephone directory assistance. We have developed a two-stage LR parser that uses two classes of LR tables: a main grammar table and several sub-grammar tables. These grammar tables are separately compiled from a context-free grammar. The sub-grammar tables deal with semantically classified items, such as city names, town names, block numbers, and subscriber names. The main grammar table controls the relationships between these semantic items.

Dividing the grammar into two classes has two advantages: since each grammar can be compiled separately, the time needed for compiling the LR table is reduced, and the system can easily be adapted to many types of utterances by changing the main grammar rules.

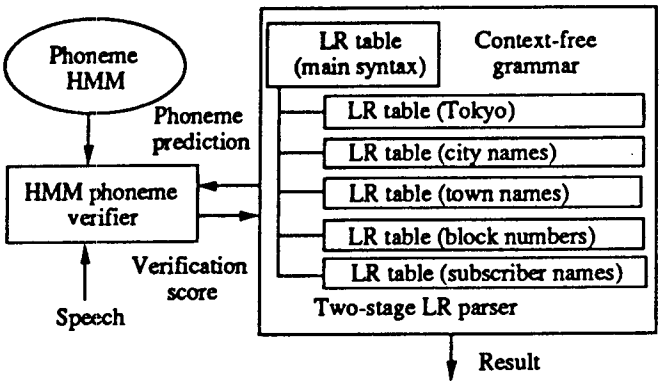

Figure 1: Structure of the continuous speech recognition system.

2.2. Accurate Scoring

Figure 2 shows the search algorithm. Our algorithm uses a backward trellis as well as a forward trellis so as to accurately calculate the score of a phoneme sequence candidate. The backward trellis likelihood is calculated without any grammatical constraints on the phoneme sequences; it is used as a likelihood estimate of potential succeeding phoneme sequences.

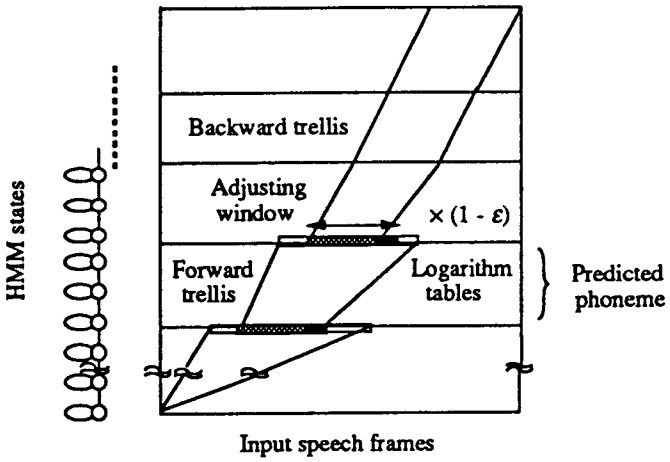

Figure 2: Search algorithm for the continuous speech recognition system.

2.3. Adjusting Window

We proposed an algorithm for determining an adjusting window that restricts calculation to a probable part of the trellis for each predicted phoneme. The adjusting window (shaded rectangle in Fig. 2) has a length of 50 frames (400 ms). The score within the adjusting window is calculated by taking the convolution of the forward and backward trellises. In this procedure, the likelihood in the backward trellis is multiplied by $(1-\epsilon)$, where ϵ is a small positive value.

2.4. Merging Candidates

The LR tables need multiple pronunciation rules to cover allophonic phonemes, such as devoicing and long vowels in Japanese pronunciation. These multiple rules cause an explosion of the search space. To make the search space smaller, we merge phoneme sequence candidates as well as grammatical states when they are phonetically and semantically the same. We further merge the candidate word sequences having the same meaning, ignoring the differences in non-keywords.

3. RECOGNITION EXPERIMENTS

3.1. Experimental System

We developed a telephone directory assistance system that covers two cities and contains more than 70,000 subscriber names. The vocabulary size is roughly 80,000. The grammar used in this system has various rules for interjections, verb phrases, post-positional particles, etc. It was made by analyzing 300 sentences in simulated telephone directory assistance dialogs. Figure 3 gives an example of an inquiry that can be accepted by the system. The word perplexity was about 70,000. In this task, no constraints were placed on the combination of addresses and subscriber names by the directory database, since users may sometimes input wrong addresses.

"Sumimasen etto, Tokyo no Mitaka-shi, etto Amari-san no denwabangou wo oshietekudasai"

("Excuse me, uh could you give me the phone number of Mr. Amari in Mitaka Tokyo?"

(in English))

Figure 3: Example of inquiry that can be accepted by the system.

We prepared two speaker-independent HMM types to evaluate our algorithm: 56 context-independent phoneme HMMs, and 358 context-dependent phoneme HMMs. Each HMM has 3 states, each with 4 Gaussian distributions. We evaluated our proposed algorithm by using 51 sentences that included

184 keywords. These utterances were prepared as text with various interjections and verb phrases. They were "spontaneously" uttered by eight different speakers.

3.2. Experimental Results

The average sentence understanding and key-word recognition rates are shown in Fig. 4. These results confirm the effectiveness of merging at the meaning level and of context-dependent HMMs. These techniques achieved an average sentence understanding rate of 65% and an average keyword recognition rate of 89%.

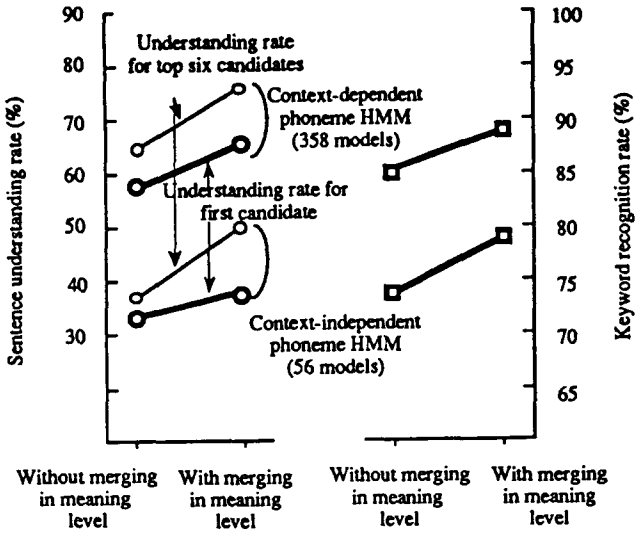

Figure 4: Sentence and keyword recognition rates.

4. HMM COMPOSITION

4.1. Principle

The HMM composition assumes that the NOVO HMM (NOVO means voice mixed with noise) obtained by combining two or more "source HMMs" will adequately model a complex signal (i.e. noisy speech) resulting from the interaction of these sources. The source HMMs may model clean speech recorded in noise-free conditions or various noise sources, such as stationary or non-stationary noises, background voices, etc. In HMM decomposition [7], recognition is carried out by decomposing a noisy observation in a multi-dimensional state-space (at least 3 dimensions), whereas in HMM composition the noisy observation is modeled before the recognition so the computation load is much smaller than for HMM decomposition.

Let R,S, and N represent the noisy-speech, clean-speech, and noise signals. X_{cp}, X_{lg}, and X_{in} are the variables corresponding to signal X in the cepstrum, the logarithm and the linear spectrum; μ and Σ are the mean vector and the covariance matrix of the Gaussian variable, respectively; Γ is the cosine transform matrix; and c is the vector of LPC cepstrum coefficients. An example of the HMM composition process to create a NOVO HMM as the product of two source HMMs is shown in Fig. 5. Initial probabilities and transitional probabilities of the NOVO HMM can be deduced directly from the source HMMs as the product of the corresponding parameters of the source HMMs.

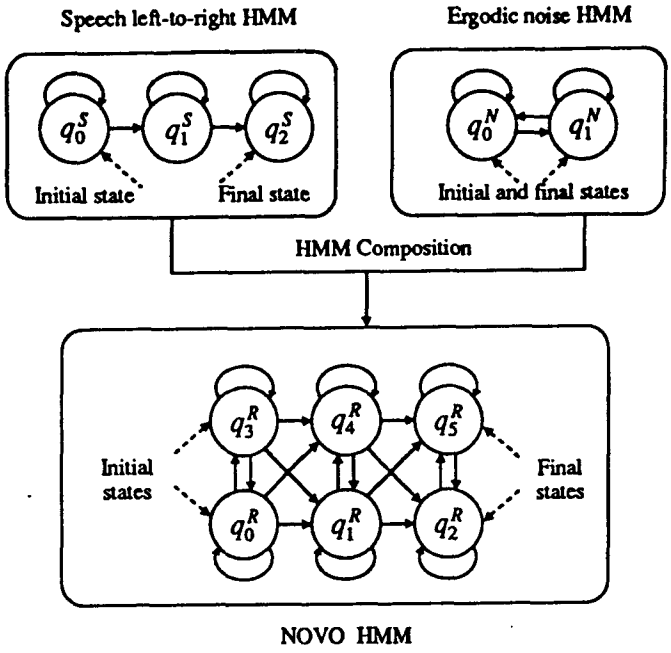

Figure 5: Example of the HMM composition process to create a NOVO HMM as the product of two source HMMs.

The output probabilities of the NOVO HMM are inferred as shown in Fig. 6. Since source HMMs are defined in the cepstrum domain, and clean speech and noise are additive in the linear spectrum domain, the normal distributions defined in the cepstrum domain are transformed into lognormal distributions in the linear spectrum domain and summed. In the figure, $k(SNR)$ is a weighting factor that depends on the estimated SNR of the noisy speech. The distributions obtained in the linear spectrum domain are finally converted back into the cepstrum domain. The process shown in the figure has to be repeated for all states and for all mixture components of the noise and clean-speech HMMs.

4.2. Experimental Results

The effectiveness of the HMM composition technique was evaluated by the telephone directory assistance system, using the 56 context-independent phoneme HMMs. The clean-speech and the noisy-speech HMMs had 3 states, each with 4 Gaussian distributions. The noise model had one state and one Gaussian distribution. Thus the NOVO HMMs had 3 states, each with 4 Gaussian distributions; there was no increase in the decoding time. Two kinds of noise were used for this experiment: computer-room noise (stationary) and

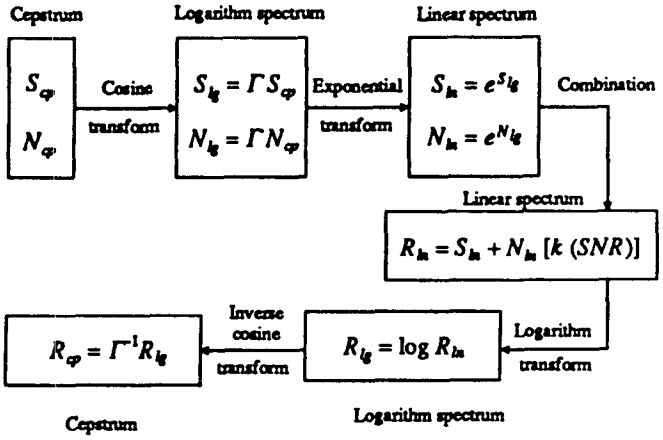

Figure 6: NOVO transform to infer the output probabilities of the NOVO HMM.

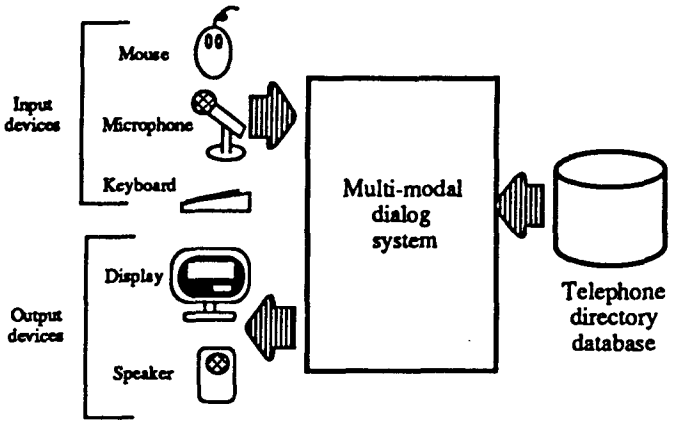

Figure 7: Basic structure of the multi-modal dialog system for telephone directory assistance.

crowd noise (nonstationary). The same sentences as in Chapter 3 were used for testing.

The experimental results showed that the NOVO HMMs could be obtained very rapidly and gave similar recognition rates to those of HMMs trained by using a large noise-added speech database. The efficiency and flexibility of the algorithm and its adaptability to new noises and various SNRs make it suitable as a basis for a real-time speech recognizer resistant to noise.

5. MULTI-MODAL TELEPHONE DIRECTORY ASSISTANCE SYSTEM

5.1. System Structure

We designed a multi-modal speech dialog system for telephone directory assistance with three input devices (microphone, keyboard and mouse) and two output devices (speaker and display), based on the above-mentioned continuous speech recognition and NOVO HMM techniques. Figure 7 shows the basic structure of the dialog system [10]. Since the interaction time, that is, the recognition speed is crucially important in testing dialog systems, we reduced the number of subscribers to 2,300 in this experimental system. The vocabulary size was roughly 4,000 as shown in Table 1. The corresponding beam width was also reduced to 200. This recognition system uses context-independent HMMs and does not use merging at the meaning level. Implemented on an HP-9000-735, the recognition currently takes about 20 seconds per sentence.

Figure 8 shows an output window example in our system. The numbers on the left side of the window show the order of candidates. This window displays five potential subscriber candidates. For each candidate, the system displays five slots: city, town, block number, subscriber name, and telephone number. A simple example of how this dialog system is used is as follows:

1. After clicking the speech button, a user states an address and subscriber name.
2. The system recognizes the input speech and displays five candidates.
3. If one of the candidates is correct, the user obtains the telephone number by clicking the telephone number slot of the correct candidate.

5.2. Dialog Controller

The main functions of the dialog controller are as follows.

Display Candidates After speech recognition, four potential candidates are displayed in order of their likelihood scores. The telephone directory assistance database constraint is not usually used in selecting these candidates. However, the fifth candidate is the candidate that satisfies the constraint in the telephone directory assistance database, because there is a high possibility that the candidate that satisfies the constraint is correct, even if it has a low likelihood score.

Table 1: Vocabulary size of the multi-modal telephone directory assistance system.

Semantic item	City names	Town names	Block numbers	Subscriber names
Vocabulary size	2	27	620	3504 (Full name: 2287) (Last name: 1217)

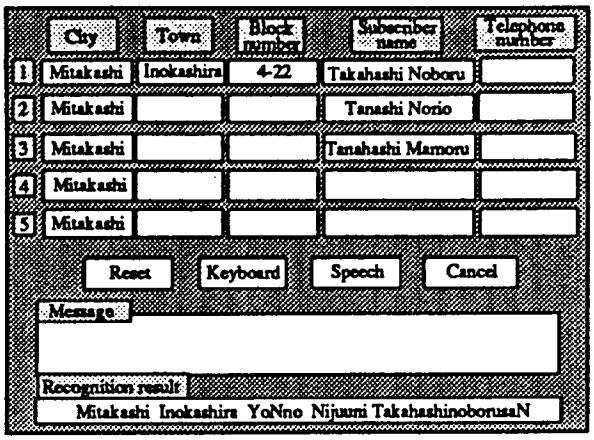

Figure 8: Example of a window in the multi-modal dialog system for telephone directory assistance.

Error Correction If there is no correct candidate among the five candidates, the user corrects the input error by choosing the candidate closest to the correct subscriber address and name, clicking the wrong keyword slot, and uttering a sentence with the specified semantic item. In the error correction mode, the system switches the main grammar to the grammar in which the clicked item must be uttered. For example, if a user clicks the subscriber name slot, the system switches the main grammar to the grammar for utterances that need to include a subscriber name. The user can include some new information in the sentence, in addition to the specified item. The beam width is also increased to raise the recognition accuracy.

5.3. Evaluation

This system was evaluated from the human-machine-interface point of view. We asked 20 researchers in our laboratory to try to use this system. Dialog experiments were performed to evaluate the following issues:

1. System performance (task completion rate, sentence understanding rate, task completion time, etc),
2. User evaluation of the system,
3. Content and manner of user utterances, and
4. Problems encountered with the system.

Training The users were first requested to practice operating this system by themselves using a tutorial system, which was an interactive system implemented on a workstation. The tutorial system was designed to control and unify the guidance as well as knowledge given to each user. One sequence of the practice, including examples of correct recognition and incorrect recognition, takes roughly 10 minutes, in which users operate the system following instructions displayed on the screen. A typical way of speaking is also displayed and practiced in this stage. Pauses and speaking rates are not controlled.

Testing 20 sheets of paper indicating the tasks using sketch maps were given to each user. Each task was indicated by the name and location of the person whose telephone number had to be requested on the map. Figure 9 shows an example of a sheet. The amount of information indicated on the sheet varied; for example, the first name or the town name of the person was sometimes not given. The users were requested to make inquiries based on the information given in each sheet. We used maps for indicating the tasks to avoid controlling the structure of the spoken sentences. When the user could obtain the desired telephone number, he/she wrote down the number on the answer sheet, and proceeded to the next task. Even if the user could not get the telephone number after all efforts, he/she was requested to proceed to the next task.

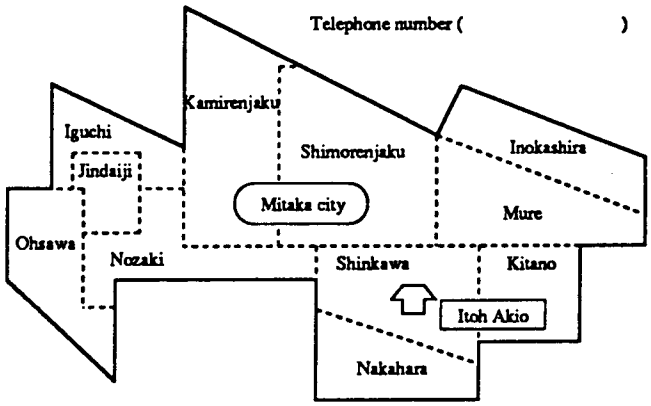

Figure 9: Example of the sheet indicating a directory inquiry task.

Questionnaires After testing, each user was requested to answer several questions, and the information obtained was compared with various logs recorded during the test.

Results The results of the experiments gave the task completion rate as 99%, which means that, in most of the trials, the users could get the correct telephone numbers. The average number of utterances for each task was 1.4, and the average sentence understanding rate was 57.8%. The average rate for the correct recognition result being indicated in the top five candidates was 77.5%. We found that the higher the top five recognition rate was, the lower the average number of utterances became.

The average time needed to complete each task was 57.2 seconds, and it decreased as the users became more experienced. About 75% of the users said that they prefered using the computer-based dialog system to a telephone directory. About 55% of the users said that the system was easy to use. The main reason for negative answers to this question was highly related to the feeling that the response time of the system was too slow.

We have collected a speech database through these experiments for future analysis and experiments.

6. CONCLUSIONS

We proposed a very-large-vocabulary speaker-independent continuous speech recognition algorithm and applied it to a telephone directory assistance system including 70,000 subscriber names. The algorithm is accurate and efficient, using a two-stage LR parser with phoneme HMMs. The sentence understanding and keyword recognition rates with context-dependent phoneme HMMs and merging at the meaning level are 65% and 89%, respectively, demonstrating that our algorithm works well for large-vocabulary continuous speech recognition. A multi-modal dialog system that uses this recognition algorithm was implemented, and evaluated from the human-machine-interface point of view. Although experimental results show that the smaller-scale system containing 2,300 subscribers works very well, we still need to improve the performance of the system; in particular, to speed up the processing time.

References

1. K. Kita, K. Kawabata, and H. Saito, "HMM Continuous Speech Recognition Using Predictive LR Parsing," Proc. ICASSP 89, pp.703-706 (May 1989).
2. M. Tomita, "Efficient Parsing for Natural Language: A Fast Algorithm for Practical Systems," Kluwer Academic Publishers (1986).
3. S. Austin, R. Schwartz, and P. Placeway, "The Forward-Backward Search Algorithm," Proc. ICASSP 91, pp.697-700 (May 1991).
4. R. Schwartz and S. Austin, "A Tree-Trellis Based Fast Search for Finding N Best Sentence Hypotheses in Continuous Speech Recognition," Proc. ICASSP 91, pp.705-708 (May 1991).
5. Y. Minami, T. Matsuoka, and K. Shikano, "Very Large Vocabulary Speech Recognition Algorithm for Telephone Directory Assistance", Proc. 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (Octorber 1992).
6. Y. Minami, K. Shikano, S. Takahashi, T. Yamada and O. Yoshioka, "Large-Vocabulary Continuous Speech Recognition System for Telephone Directory Assistance", Proc. ICASSP 94, 75.5 (April 1994) (to be published)
7. A. P. Varga and R. K. Moore, "Hidden Markov Model Decomposition of Speech and Noise", Proc. ICASSP 90, pp. 845-848 (April 1990)
8. F. Martin, K. Shikano and Y. Minami, "Recognition of Noisy Speech by Composition of Hidden Markov Models", Proc. Eurospeech '93, pp. 1031-1034 (September 1993)
9. J. A. Nolazco Flores and S. J. Young, "Adapting a HMM-Based Recogniser for Noisy Speech Enhanced by Spectral Subtraction", Proc. Eurospeech '93, pp. 829-832 (September 1993)
10. O. Yoshioka, Y. Minami and K. Shikano, "Development and Evaluation of a Multi-Modal Dialogue System for Telephone Directory Assistance", Technical Report of IEICE, SP93-128 (January 1994) (in Japanese)