Membering^{TM}: A Conference Call Service with Speaker-Independent Name Dialing on AIN

Sung-Joon Park[1], Kyung-Ae Jang[1], Jae-In Kim[1], Myoung-Wan Koo[1], Chu-Shik Jhon[2]

[1]Service Development Laboratory, KT, Seoul, Korea
[2]School of Computer Science and Engineering, Seoul National University, Seoul, Korea
sjpak@kt.co.kr

Abstract

In this paper, we present a conference call service with speaker-independent name dialing, which allows any members of a group to communicate with all the other members by saying their names over the telephone. The service is implemented on AIN with a VoiceXML interpreter. Members' names in a group are registered over the telephone or via the Web interface. If users register their names with phones without an operator's help, automatic speech recognition is used. We use speaker-independent sub-word models for speech recognition. The name registration via the Web interface or by an operator is done with text input. In the case of name registration with speech, the average recognition rate of the names for the members who register them is 96.4%, while the average rate for the other members is 91%. When the names are registered with text input, the average recognition rate is 99.8%.

1. Introduction

Of all the applications of speech recognition, those involving recognition of speech transmitted over telephone links have received the most attention. Korea Telecom (KT) has deployed several services with speech recognition over the telephone network (Park, Koo, and Jhon 1999; Koo, Park, and Kim 2000; Kim, Koo, and Park 2001). There have been researches on feature extraction, acoustic models, pronunciation dictionaries, and other areas to improve recognition accuracy, but it is not easy to obtain high recognition accuracy with a large vocabulary. The other way, in this paper, we propose a service using a very small vocabulary, which leads to high recognition accuracy with the current technology.

Membering^{TM} is a conference call service with speech recognition developed by KT. Conference call is a service feature that allows a call to be established among three or more stations in such a manner that each of the stations is able to communicate with all the other stations. Membering^{TM} allows any member of a group to communicate with all the other members by saying their names over the telephone. For example, when a user wants to communicate with other members, he dials the phone number of his group and says his name if the number of the calling phone is not a registered one, whereupon the system is expected to recognize the name and connects the call to the other members. In case that the number of the calling phone is a registered one in the group, the system automatically connects the call to the other members without asking the caller's name. The service is implemented on advanced intelligent network (AIN) using a Voice eXtensible Markup Language (VoiceXML) interpreter.

The AIN architecture was derived from a set of functional needs associated with the provision of voice-band telecommunications services (Berman and Brewster 1992). AIN services are provided through interactions between switching systems and the systems supporting AIN service logic. The interaction with users may involve the provision of service-specific announcements to a user or the collection of digits input by a user. The participant interaction resources may reside in a switching system or may be provided by an intelligent peripheral (IP). An IP is a system which controls and manages resources such as text-to-speech synthesis, announcements, speech recognition, and digit collection. The IP in our system includes a VoiceXML interpreter that reads and processes VoiceXML pages as described by the VoiceXML language standard (Edgar 2001).

To accomplish dialing by speech, a user has to first register a set of names together with phone numbers, at which point the system creates a model for each name in the set. The registered models are used later for recognition. The models could be either speaker-dependent or speaker-independent. In general, speaker-dependent models are more accurate, but the memory required for storing the models becomes prohibitively large as the number of registered names increases. To overcome this problem, we use speaker-independent, sub-word models. This means that one need not store acoustic models for each name, resulting in significant savings in memory. Names can be registered with speech or with text. To register names with speech, a user dials the reserved phone number which is used only for registration, say all members' names including his name, entering the corresponding phone number for each name. A group number is given after registration.

In the registration with speech, the problem of automatically finding transcriptions of the unknown words as sequences of sub-word units arises because the spellings of the words are unknown and are not exploited. Some strategies have been developed in the literature for this problem (Haeb-Umbach, Beyerlein, and Thelen 1995; Jain, Cole, and Barnard 1996; Elvira and Torrecilla 1998). In our system, users' utterances are stored when a user registers names. The stored utterances are used when the system provides a feedback of the recognized name by playing them

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review

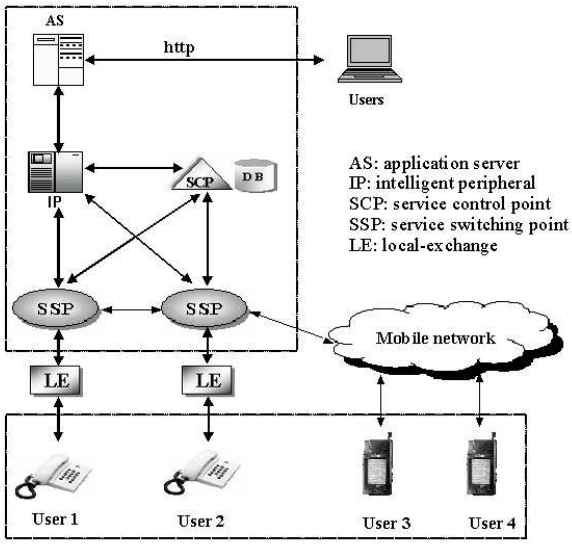

Figure 1: *AIN system architecture of Membering^{TM}.*

and prompting for a confirmation during recognition.

Another problem is that a lowering of recognition rates can be caused by the wrong transcriptions. Temporal variations of speech also may decrease the rates because the utterances used for creating the models may not represent all the other utterances at the stage of recognition. So, the system also provides registration with text. The system translates text names into sub-word sequences and stores them in the dictionary.

This paper is organized as follows. Section II describes system architecture, a VoiceXML interpreter and speech recognizer. In Section III, two main phases of the service, registration and recognition, are presented. Database and experimental results are presented in Section IV and V. Our conclusions are given in Section VI.

2. System overview

2.1. System architecture

AIN is a telecommunication network services control and management architecture. Fig. 1 shows the architecture of the Membering^{TM} service on AIN.

The service switching point (SSP) is an enhanced switching system. This node has extra call-processing software to enable the switching system to recognize when call-processing at the external service control point (SCP) is required (Sharp and Clegg 1994). When the conditions for external processing are detected, a query/begin transaction message is launched to the service control point, where the application service logic will decide how to process the call. The SCP is a network database which is one of the intelligent network (IN) physical elements that also contain service control logic.

Intelligent peripherals (IPs) handle specialized interactions between the user and the intelligent network. The IP has resources such as tones and announcements, speech recognition and voice synthesis. As a result, it has both signaling and voice circuits to the service switching point.

If the SCP instructs the SSP to route a call to an IP, the IP performs a particular function, for example 'play announcements and collect digits'. To support this function, the IP in our system includes a VoiceXML interpreter that has the role of interpreting the VoiceXML documents submitted from the Web, and then generates the interactive voice dialing services. VoiceXML documents are stored in the application server (AS). The AS also includes the VoiceXML documents for other services as well as Membering^{TM}.

2.2. VoiceXML Interpreter

The VoiceXML is an interactive markup language, which supports an environment to develop applications that search Web information via voice and phone. It is designed for creating audio dialogs that feature synthesized speech, recognition of spoken and dual-tone multi-frequency (DTMF) key input, telephony, and mixed initiative conversations. Its major goal is to bring the full power of Web development and content delivery to voice response applications, and to free the authors of such applications from low-level programming and resource management. The VoiceXML architecture has following key components: Implementation platform, VoiceXML interpreter context, VoiceXML interpreter, and document servers. The relationship between components is shown in Fig. 2.

The implementation platform checks the inbound and outbound calls, and generates events in response to user actions and system status. These events are acted upon by the VoiceXML interpreter or the VoiceXML context. The VoiceXML interpreter context may monitor user inputs and manage the environment variables related to the sessions. VoiceXML interpreter parses the VoiceXML document responded from a document server, and controls the entire logic according to the dialog. The document server processes the requests from the VoiceXML interpreter. The server produces the VoiceXML documents in reply, which are processed by the interpreter. The VoiceXML interpreter can request any file format, such as documents, CGI scripts, and audio files.

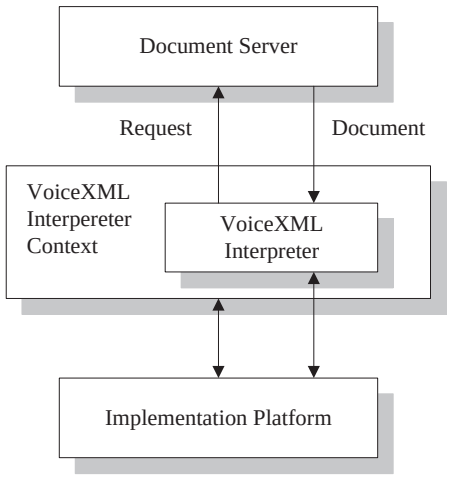

Figure 2: *Relations between components in VoiceXML.*

Accepted after full paper review

2.3. Speech recognizer

The system uses the familiar phonetically-based, hidden Markov model (HMM) approach. Logically it is divided into two modules: the front-end and the decoder. The front-end computes 12 dimensional linear predictive coding (LPC) cepstral coefficients and log energy from 16-bit PCM speech data sampled at 8kHz. Speech samples are partitioned into overlapping frames of 20 ms duration with a frame-shift of 10 ms. Each frame of speech is windowed with a Hamming window and represented by a 12 dimensional cepstral vector. With the calculated cepstral coefficients and log energy, the front-end computes first and second differences of the 13-dimensional vectors, and concatenates these with the original elements excluding log energy to yield a 38-dimensional feature vector. To reduce the effects of channel, the front-end performs cepstral mean removal. The cepstral mean is calculated using all the frames of the utterance and is subtracted from each frame.

The decoder computes the log likelihood of each feature vector according to observation densities associated with the states of the HMMs. The results are used in a synchronous Viterbi search over the active vocabulary. Words are represented as sequences of context-dependent phonemes, with each phoneme modeled as a three-state HMM.

The HMMs of each state are created using tree-based state tying (Young, Kershaw, Odell, Ollason, Valtchev, and Woodland 2000). A phonetic decision tree is a binary tree in which a yes/no phonetic question is attached to each node. The question at each node is chosen to maximize the likelihood of the training data given the final set of state tying. The question set of our system has 158 questions in total. Initially all states in a given item list are placed at the root of a tree. Depending on each answer, the pool of states is successively split and this continues until the states have trickled down to leaf-nodes. The final system consists of 14 mixture component context-dependent HMMs.

3. Registration and recognition

Any user can get a group number consisting of four digits by registering members' names and phone numbers, and can connect to other members by dialing service identification number 1646 followed by the group number. The number of members is currently limited up to four. If the phone number of the calling phone is registered, the call is automatically connected to the other numbers. Otherwise the system requests user's name and determines with speech recognition whether the name is a registered one in the group. If the recognized name is a registered one, the system connects the call to other members. The group number expires if it is not used for a week.

There are two methods for registration. One is the registration via the Web interface. A user just enters members' names and phone numbers, and gets a group number. Input names are transcribed to phoneme sequences and later used in the grammar for speech recognition.

In the other method, telephone is used. The phone number 1646-0000 is reserved for the registration. The scenario is shown in Fig. 3.

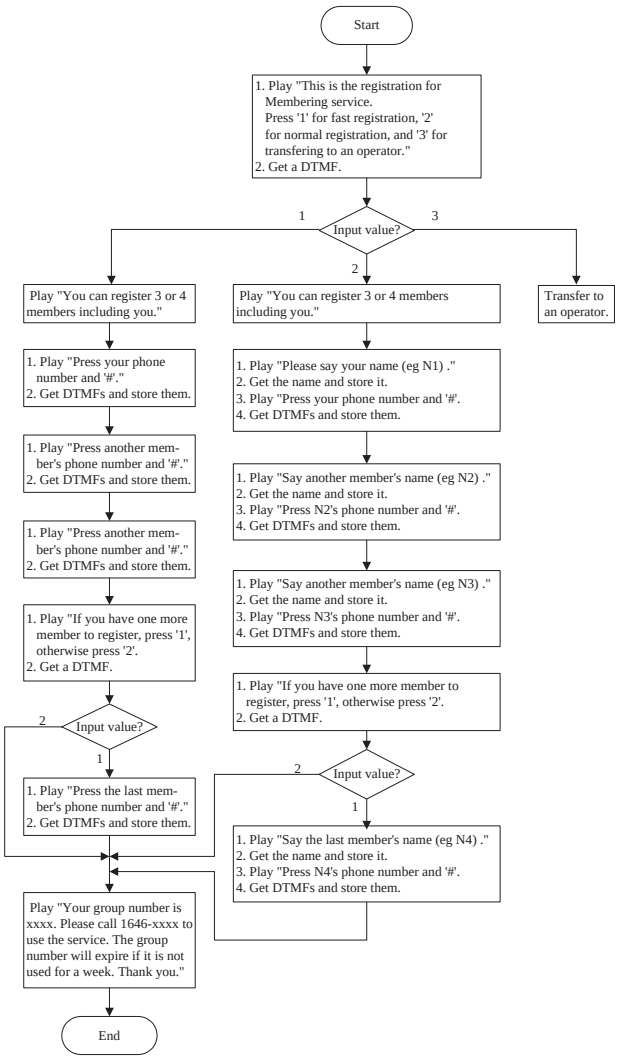

Figure 3: *Scenario of registration.*

When connected to an operator, users can register their names and phone numbers by providing necessary information to the operator. In the fast registration, members' names are not registered. Accordingly, the phones whose numbers are not registered cannot be used for the service. In the normal registration, the system uses speech recognition to register members' names. The registration with speech recognition is done by the recognition of the names spoken by the user. The recognizer produces a sequence of syllables for each name. This sequence is used later as a grammar for recognition.

After registration, any member can use the service by calling the service number 1646 followed by the group number. The scenario is shown in Fig. 4.

If the calling phone is a registered one, the system automatically connects the call to the other members. Otherwise, the caller is requested to say his name for confirmation. The speak-independent phone models along with the grammars obtained during registration are used for recognition. The complete registration-recognition scheme is shown in Fig. 5.

Accepted after full paper review

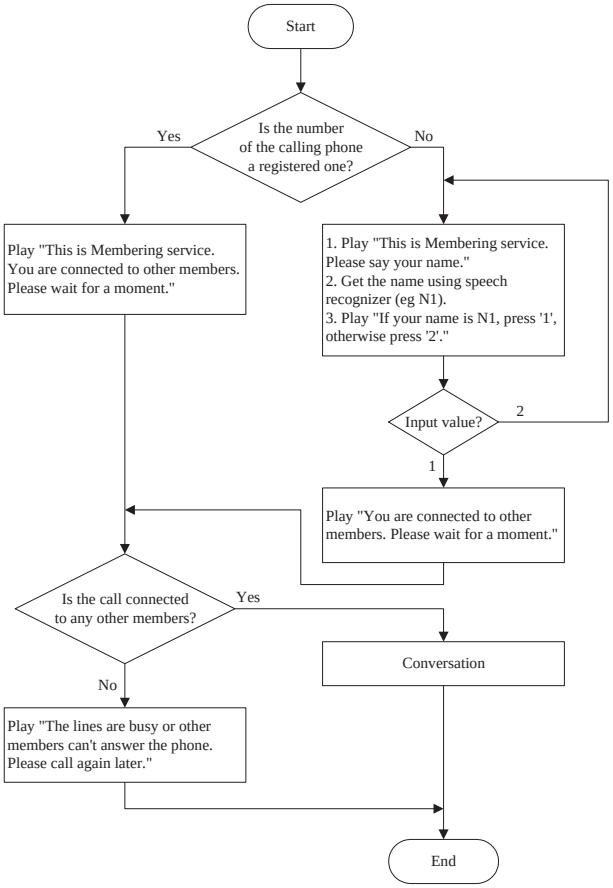

Figure 4: *Scenario of the service after registration.*

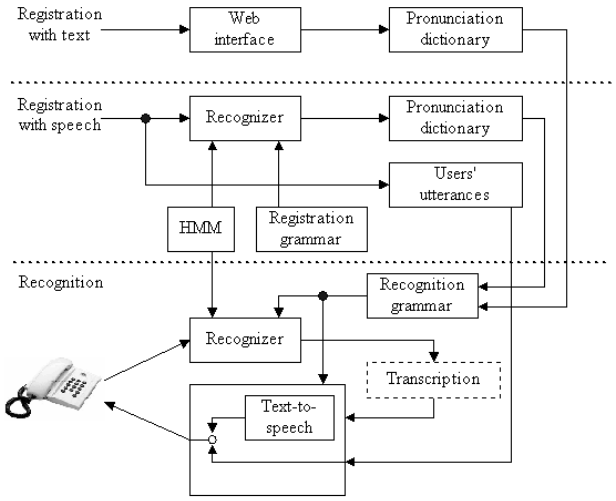

Figure 5: *Scheme of registration and recognition with speech recognition.*

4. Database

For the experiments, a set of 150 Korean names was selected among directory assistance database that includes 470,000 names. The selected 150 names includes PLUs (phonelike units) which cover over 90% of the occurrences.

The 150 names were divided into five groups that contain thirty words respectively. 100 speakers recorded them. Ten male and ten female speakers were assigned to each group and each speaker recorded 30 words six times over the telephone. To reflect temporal variations of pronunciation, the recordings were done at two sessions with an interval of over one week. In each session, three utterances for each word were recorded. From the first session recordings, one utterance for each word was used for grammar. One set or two sets of the first session recordings was used for registration with speech and 100 pronunciation dictionaries were created.

5. Experimental results

One group who recorded a set of 30 names consists of 20 people. All the subsets consisting of four names were created for each speaker in the group, registered, and tested by the member who registered the names and the other members of the group. The other four groups followed the same steps for the experiments.

The first recognition tests were performed using the data of the first session recordings. The data used for the registration were excluded in the test. In case of registration with speech, the average recognition rates for the members who registered names were 96.8% while the rates for the other members were 91%. The data of the second session recordings showed similar results. The members who registered names showed the rates of 95.9% while the other members showed 91%.

In case of registration with text, the recognition rates are higher than those of registration with speech. Because the grammars from the registered names are independent of who registered the names, the recognition rates were averaged for all the members. The rates are 99.7% and 99.8% for the first session recordings and the second session recordings. All the results are summarized in Table 1.

In case of registration with speech, the system is made to ask the user to speak each word just once for users' convenience because the registration is an additional burden to users. For the purpose of comparison, however, more experiments were made for the case of using two utterances for the registration with speech. If two utterances for each name are used for the registration, the grammars have two transcriptions for each name and the recognition rates may be improved when the transcriptions are different and reflects the speaker's variations of the utterances. Table 2 shows the results of the experiments in which two utterances were used for registration of a name.

The recognition rates of registration with text are the

Table 1: *Recognition rates of the registration with speech and with text.*

		Registrant	The other members
Registration with speech	1st	96.8%	91%
	2nd	95.9%	91%
	Average	96.3%	91%
Registration with text	1st	99.7%	
	2nd	99.8%	
	Average	99.8%	

Table 2: *Recognition rates of the registration with two utterances for each name.*

		Registrant	The other members
Registration with speech	1st	97.3%	94.7%
	2nd	97.9%	94.6%
	Average	97.8%	94.7%

highest. Accordingly, to improve the rates, the system allows users to modify the dictionary via the Web interface even after the names have been registered with speech.

6. Conclusions

In this paper, we proposed the Membering^{TM} service which is a conference call service with speech recognition. The architecture of AIN and the components of the system were described. We also presented the experimental results of registrations with speech and with text. While the average recognition rates according to registration with speech are 96.3% and 91%, respectively for the registrant and the other members, the average rate of registration with text is 99.8%. High recognition accuracy could be obtained with a very small vocabulary.

References

Berman, R. K. and J. H. Brewster (1992). Perspectives on the AIN Architecture. *IEEE Communications Magazine*, 27–32.

Edgar, B. (2001). *The VoiceXML Handbook : Understanding and Building the Phone-Enabled Web*. New York: CMP Books.

Elvira, J. and J. Torrecilla (1998). Name dialing using final user defined vocabularies in mobile (GSM & TACS) and fixed telephone networks. *Proceedings of IEEE ICASSP*, 849–852.

Haeb-Umbach, R., P. Beyerlein, and E. Thelen (1995). Automatic transcription of unknown words in a speech recognition system. *Proceedings of IEEE ICASSP*, 840–843.

Jain, N., R. Cole, and E. Barnard (1996). Creating speaker-specific phonetic templates with a speaker-independent phonetic recognizer: implications for voice dialing. *Proceedings of IEEE ICASSP*, 881–884.

Kim, H.-G., M.-W. Koo, and Y.-K. Park (2001). TeleGateway: A Spoken Dialogue System Based on VoiceXML. *KT Journal 6*, 56–63.

Koo, M.-W., S.-J. Park, and H.-K. Kim (2000). An Experimental Study on a Speech Recognition System for Barge-In and Non-Barge-In Utterances on the Telephone Network. *Proceedings of the 7th Western Pacific Regional Acoustics Conference 1*, 73–78.

Park, S.-J., M.-W. Koo, and C.-S. Jhon (1999). An Implementation of Continuous Speech Recognition for a Stock Information Retrieval System. *Proceedings of International Conference on Speech Processing 2*, 461–464.

Sharp, C. and K. Clegg (1994). Advanced intelligent networks : now a reality. *Electronics & Communication Engineering Journal*, 153–162.

Young, S., D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland (2000). *The HTK Book, Version 3.0*. Microsoft Corporation.