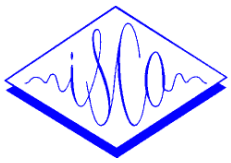

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

ISSUES CONCERNING VOICE INPUT APPLICATIONS

Tsuneo Nitta and Nobuo Sugi

Information & Communication Systems Laboratory, TOSHIBA Corporation

70,Yanagi-cho,Saiwai-ku,Kawasaki,210,JAPAN

ABSTRACT

At present, many voice input systems are unable to consistently achieve accurate recognition in practical environments. This paper describes issues and some solutions concerning voice input applications from two different standpoints of a speech recognition researcher and an application designer.

The authors point out that both the robustness in speech recognition and the well-designed user interface are particularly important for voice input applications to successfully incorporate speech recognition and to complete user's demands. Some practical applications, such as voice-activated telephones, ticket vending machines, and elevators, are also discussed, and two important items of the multimodality and well-designed prompts are emphasized.

1. INTRODUCTION

There are many as yet unsolved problems that prevent the use of the natural medium of voice for system input. At present, the recognition of unrestricted continuous speech is viewed as the most difficult of these problem to overcome. However, with the addition of simple constraints, many applications capable of accepting voice input have become available.

Voice input systems need the robustness, or stable performance in practical environments, rather than capabilities such as acceptance of a large vocabulary or continuous speech. Many voice input systems may still be unable to consistently achieve accurate recognition in noisy environments. Furthermore, many people make nonverbal noises, speak at the wrong time, and do not speak clearly. The system designer should understand these limitations of speech recognition and design the user interface for his or her applications.

In this paper, the authors first describe some issues on robust voice recognition from the standpoint of a speech researcher. Dealing with the issues of speech pattern variations and noise superposition, some solutions are discussed. Next, user interface issues and some solutions are considered from the standpoint of a voice input application designer. Through reviewing practical applications, such as voice-activated telephones, ticket vending machines, and elevators, we emphasize two important items of the multimodality and well-designed prompts.

2. ROBUST SPEECH RECOGNITION

There are many factors that give rise to speech pattern variations. we must consider these factors and some sources of superposed noise to achieve robust speech recognition.

i) Factors of speech pattern variabilities
   Environmental variations of phoneme
   Variation of a speaker's articulation with time
   Inter-speaker variation
   Speaking rate and loudness
ii) Sources of noise
   non-verbal noise made by speakers
   ambient noise

In this section, we discuss these issues and some solutions.

2.1 SPEECH PATTERN DEVIATION AND SOME SOLUTIONS

Spoken word recognition based on template matching can avoid the degradation of recognition accuracy caused by the variations in the phoneme environment. However, such tasks as discriminating between similar words, handling a large vocabulary, or rejecting unknown words with a high rejection rate are difficult for the template matching approach.

Phonetically-based word recognition has the ability to solve these problems. A phoneme is a well matched recognition unit for linguistic processing; however, because phonemes are highly affected by their environment, phonemes have wide variation. Various types of phonetic variations caused by coarticulation, devoicing, and speakers are observed in continuous speech.

In view of this, we should use multiple forms to express the variations within a phoneme rather than a single phonological unit. The authors proposed a multiple phonological unit called the phonetic segment which consist of about 600 acoustic/phonetic structures of 32~96 msec. duration (acoustic segment, phoneme, CV, VC, CC, VCV, CVC)[1],[2]. Figure 1 shows an exsample of phonetic segment lattice for a word /PIPEQTO/ (pipet). A segment of each frame is recognized using the Subspace method with multiple LPC similarity measure, in which the segment fluctuation is incorporated in an orthogonalized reference pattern set that is designed with K-L transform[1]. We can say in other words that this procedure is considered to be a statistical matrix VQ.

When phonemes expressed with multiple forms are adopted as a recognition unit, there are some strategies for converting a segment of continuous speech into discrete linguistic unit. One is an approach using phoneme spotting, or time-shift-invariant phoneme recognition such as a neural network approach[3]. Spotting phonemes with high accuracy is one of the most difficult tasks, however the resultant discrete phoneme sequences make it easy to apply linguistic processing. In another approach, stochastic phoneme models such as HMM phoneme models are connected to construct word or sentence models, and phonemes are not recognized explicitly[4]. This approach has been used widely for speech

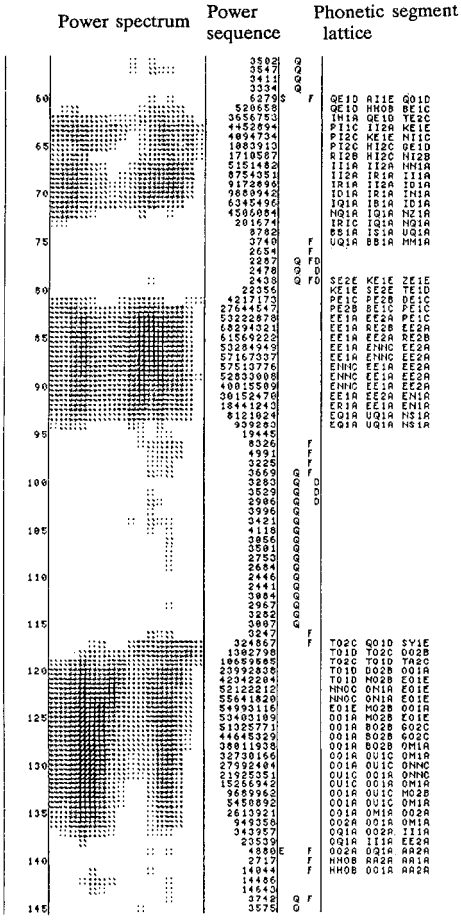

Figure 1 Phonetic segment lattice
(/PIPEQTO/ (pipet))

recognition, however it needs a large amount of training data.

In order to efficiently incorporate speech pattern variations into a recognition mechanism, we select the other approach shown in the following two steps. First, the input speech is converted into the phonetic segment lattice as mentioned above. Next, the phonetic segment lattice is entered into word models of HMM or STN(Structured Transition Network). This approach has been successfully examined in the speaker independent (connected) word recognition [1],[5].

2.2 NOISE SUPERPOSITION AND SOME SOLUTIONS

When a voice input system is trained and tested in different noise environments, system performance often degrades to an unacceptable degree. Also, many speakers make various types of unvocalized sounds when talking such as tongue clicking, sighs, breathing, lip smacking, etc. There are two fundamentally different approaches to cope with the recognition of speech contaminated by noise.

The first approach directly handles contaminated speech signals. Spectral subtraction is one of the most popular preprocessing techniques[6]. Word boundary detection is another important technique to achieve noise-robust recognition. Considering the difficulty of separating speech segments from background noise, we need to introduce the word boundary hypothesis method in which multiple word boundaries are hypothesized and tested through word verification[7],[8]. If condition can be determined under which voice input systems will be used, training with contaminated speech in the target environment is also an effective approach for improving performance[9]. Noise reduction using neural network is in the experimental stage, and practical applications lie in the near future[10].

The second approach applies some noise reducing acoustic techniques. The signal-to-noise ratio of an incoming speech signal depends on the distance between the speaker and the microphone. We can effectively obtain a high quality speech signal by using a close-talking microphone. However, since this type of microphone must be worn on a headband or used like a telephone handset, we cannot always use this type of microphone in situations where the speaker is freely moving around or using the system quickly. In such situations, adaptive noise cancellation techniques using microphone array are effective[11],[12]. However, these techniques do not perform yet reliably in the presence of acoustic reverberation and speaker movement.

3. USER INTERFACE DESIGN

Speech recognition technology needs to be more robust, however the evaluation of a voice input system not only depends upon recognition capability but also on overall system performance. The following are important items to design into the user interface for speech recognition applications.

(1) Multimodality : A developer should provide alternative paths to complete user's demands.
(2) Well-designed prompts : A developer must design, test, and modify the prompts for completing a task.
Strategies for handling failures are particularly important.

By adding a speech recognition unit combined with other input/output modalities, a system designer must improve efficiency or user-friendliness in his or her applications. In this section, we discuss some problems and associated solutions through a review of practical applications.

3.1 VOICE-ACTIVATED TELEPHONES[13]

Voice dialing requires a person to only pronounce the name of the party to be called. This function frees the user from having to memorize telephone numbers and abbreviated dialing numbers. Figure 2 illustrates operation of voice dialing. Registration and dialing are both cued by an LED display.

To register a name, the abbreviated number and telephone number of the name are entered. A voice lamp or ready indicator cues the speaker three times and insures that the telephone records the speaker's utterances correctly for confirming pronunciation. If any of the utterances are differ greatly, the NG lamp signals the need for a new set of utterances. Finally, the entry is confirmed through reproduction of the third utterance in the handset. This confirmation ends number registration.

To dial a number, the user lifts the handset and says the desired name after being cued by the voice lamp. The telephone checks the utterance and indicates rejection with

the NG lamp if the utterance is garbled by noise. Once the OK lamp is lit, the name scoring the best match is reproduced through the handset for confirmation. If the name is correct, the telephone automatically dials the registered number corresponding to the name.

If the name is misrecognized, the next best matching name is reproduced through the handset when the next button is pressed. If reproduced name is correct, the telephone dials the corresponding number . This operation can be repeated up to three times. If the correct name is not among the three best matched names, the user can make another attempt to pronounce the name after the handset has been replaced.

It is important for this application to provide simple instructions to inform the user how to use it. As the system also has a separate training mode, a beginner can practice voice dialing beforehand. In the case of dialing applications, quick system response and the next button are effective in correcting errors since dialing numbers takes time.

3.2 VOICE–ACTIVATED TICKET VENDING MACHINE[14]

If a voice–activated ticket vending machine is used at stations, ticket purchases can be done quickly because there would be no need to refer to any map for the location of destination.

Figure 3 shows the configuration for a system equiped with an infrared sensor for adapting to a station's noisy environment. When a user approaches the vending machine, the sensor is activated, and the system is ready to accept voice commands. To purchase a ticket, a user utters the destination as guided by instructions shown on a CRT. The fare is then automatically deducted from a prepaid card, and the ticket is issued. If the station name is misrecognized, the user has to restate the destination or choose the station name on a touch panel.

A prototype voice–activated ticket vending machine successfully completed a field trial at a station on the Kinki Nippon Railway in Osaka in 1988. In the three month trial, a number of reject cases were observed which were caused by the following.

- Users speaking as they approached the system
- Users saying "For Nara, please" instead of "Nara"
- Users using synonyms like "Nihonbashi" instead of "Nipponbashi"

A phonetically–based speech recognizer should be developed to cope with these problems and extend vocabulary.

3.3 ELEVATOR APPLICATION[15]

Voice input elevator systems permit people to tell the elevator where to stop. By adding speech recognition capabilities, people with physical disabilities or people

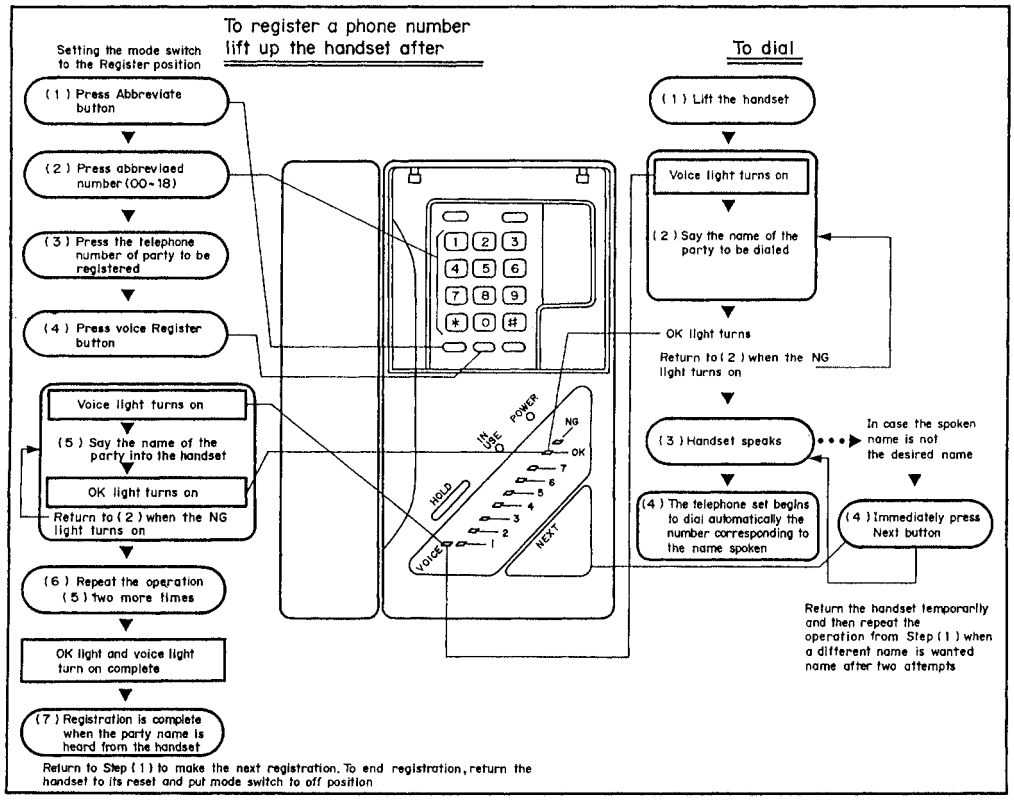

Figure 2 Voice dialing operation

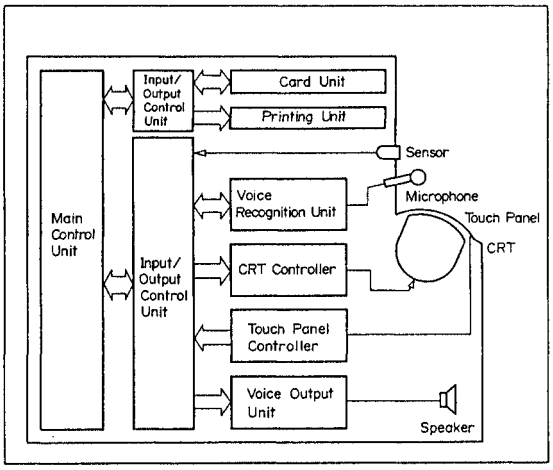

Figure 3 Blockdiagram of voice-activated vending machine

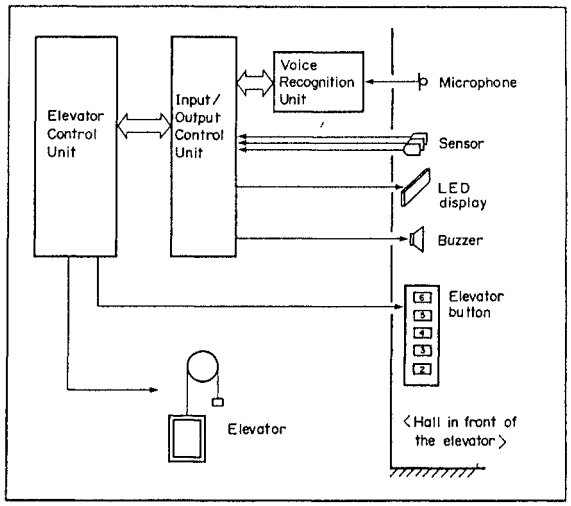

Figure 4 Blockdiagram of elevator system with speech recognition unit

whose hands are busy with other tasks can use elevators easily.

Figure 4 shows the configuration of a system equiped with three infrared sensors. In this case, the sensors perform the important role of detecting people approaching a microphone from various directions. To assign a floor, a user utters the floor number when cued from an LED display. If the utterance is misrecognized, the user has to restate the destination within a few seconds or push the desired button on the elevator button panel. The user signals correct recognition by doing nothing for a few seconds or stepping back from the sensor detection area one step.

Because the first prototype system had only one sensor, the system could not detect people standing near the sensor but misaligned with the sensor's main axis. At present, elevator systems with speech recognition capability are in practical use.

4. CONCLUSION

Issues and some solutions concerning voice input applications have been discussed. Robustness in speech recognition and a well-designed user interface are important for the practical applications.

We believe that in the future, we can expand the communication channel between people and computers by using multiple modalities. Much work is needed to realize such a future however, we will be able to choose the best combination from the multimodal interface in relative to each task as people do.

REFERENCES

[1] T.Nitta, K.Uehara and S.Watanabe, "Connected Word Recognition Based on Word Transition Network and Selective Scoring of Phonetic Segments", IECE Japan Trans. Vol.J71-D, No.9 pp.1640-1649, 1988 (Japanese)

[2] N.Sugi, J.Iwasaki, H.Matsu'ura, T.Nitta, A.Fukumine and A.Nakayama, "Speaker Independent Word Recognition System Based on the Structured Transition Network of Phonetic Segments", Proc. ICSLP-90 in Kobe, 1990

[3] M.Miyatake, H.Sawai, Y.Minami, and K. Shikano, "Integrated Training for Spotting Japanese Phonemes Using Large Phonemic Time-Delay Neural Networks" , Proc. 1990 Int. Conf. Acoust., Signal, Speech Processing, pp.449-452, 1990

[4] K-F.Lee, "Automatic Speech Recognition - The Development of the SPHINX System", Kluwer Academic Publishers, 1989

[5] H.Matsu'ura, J.Iwasaki and T.Nitta, "Speaker Independent Word Recognition Based on SM-HMM Using Matrix Quantization for Phonetic Segments" , Proc. 1990 Autumn Meeting of the Acoustic Society of Japan, pp.69-70, 1990 (Japanese)

[6] S.F.Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction",IEEE Trans. ASSP, Vol.27 No.2, pp.113-117, 1979.

[7] L.F.Lamel, L.R.Rabiner, A.E.Rosenberg and J.G.Wilpon ,"An Improved Endpoint Detector for Isolated Word Recognition", IEEE Trans. Acoust., Speech & Signal Processing, Vol.29, No.4, pp.777-785, 1981

[8] T.Nitta, K.Uehara and S.Watanabe, "Telephone Speech Recognition Based on Word Boundary Hypothesis and Multiple LPC Similarity Measures", IECE Japan Trans. Vol.J71-D, No.1 pp.59-66, 1988 (Japanese)

[9] A.Nakayama, N.Sugi and T.Nitta, Microphone Characteristics and Noise Adaptation Techniques for Speaker-independent Word Recognition", Trans. Speech Research, IEICE japan, SP89-104, 1990 (Japanese)

[10] S.Tamura and M.Nakamura "Improvements to the Noise Reduction Neural Network", Proc. 1990 Int. Conf. Acoust., Signal, Speech Processing, pp.825-828, 1990

[11] B.Widrow and S.D.Stearns. "Adaptive Signal Processing" , Prentice-Hall,1985

[12] Y.Kaneda and J.Ohga "Adaptive Microphone Array System for Noise Reducation", IEEE Trans. Acoust. , Speech & Signal Processing, Vol.34, No.6, pp.1391 -1400, 1986

[13] N.Ohguchi, "Toshiba's Voice Activated Phone Incorporates Artificial Intelligence", Telecommunications, pp.94-96, June 1989

[14] N.Sugi, T.Nitta, Y.Mimata, T.Shimada and T.Nishimura ,"A Voice Activated Ticket Vending Machine", Trans. Speech Research, IEICE, SP89-24, 1989 (Japanese)

[15] N.Sugi, T.Nitta, Y.Nakajima and Y.Harada,"Development of a Speech recognition Unit for Elevator", Proc. 1990 Convension of IEICE, A-242, 1990 (Japanese)