ISSUES RELATING TO THE FUTURE OF ASR FOR TELECOMMUNICATIONS APPLICATIONS

Shigeki Sagayama, Kiyoaki Aikawa

NTT Human Interface Laboratories
1-1, Hikari-no-Oka, Yokosuka-shi, Kanagawa, 239 Japan

ABSTRACT

Issues relating to automatic speech recognition (ASR) are discussed with respect to applications in the telecommunications area in the near future. As a preliminary, we introduce an interesting discussion from a past conference in Japan about what is hindering the spread of ASR. Then, some relatively new robustness issues in telephone-based ASR applications are discussed. These include accurate voice/noise discrimination, and multiple microphones, utterance verification/rejection for flexible vocabulary systems, breath noise and hand noise, instantaneous adaptation to environmental noise, a spelling method for Japanese Kanji texts, dialog control issues, distransparency of ASR systems, children's voices, HMM training with localized data, adaptive dialog strategy.

1. INTRODUCTION

There are still a number of problems that must be solved before we can successfully apply automatic speech recognition (ASR) technology to telecommunications. Some have received a lot of attention and have been partly solved. For instance, there has been progress in robust acoustic parameters, speaker and channel adaptation, word spotting techniques, extraneous utterances, efficient search methods, dialog control, etc. Apart from these currently highly concerned problems, there are numerous other important issues. Some of them are related to already discussed issues; some others are new.

This paper focuses on topics and issues other than those often addressed in papers and at conferences. Our intent is to attract attention to topics that, while relatively rare, are still important.

Authors' E-mail addresses: {saga,aik}@nttspch.hil.ntt.co.jp, respectively.

2. GENERAL PROBLEMS OF SPEECH RECOGNITION

2.1. An E-mail-based Discussion in 1994

There has been a long-running discussion among Japanese researchers and managers working in the ASR area as to why ASR is not widely used in the real world even after the great technological progress made due to applying statistical approaches to speech recognition.

In 1994, one of the authors (Sagayama) used the opportunity provided by the formation of a new Special Interest Group of Spoken Language Processing (SIG-LSP, Prof. S. Nakagawa, chair), which is a SIG in the Information Processing Society of Japan (IPSJ), to run an internet-based discussion of what is hindering the wider use of ASR and how to cope with the problems. To start the discussion, we compiled a list of 22 likely reasons [(a)–(v) below]. In the ensuing active and exciting E-mail discussions, a number of other views were contributed and finally six more reasons [(A)–(F)] were added. We reported the entire list in the opening of a SIG-SLP conference and asked the participants to vote on these explanations. Here is the list and the votes each item received.

(a) Recognition accuracy is still too low. ASR technology is not mature and currently not useful. (36 votes)

(b) ASR will not be useful until spontaneous speech recognition is achieved. Current continuous speech recognition technology is useless. (23 votes)

(c) The recognition accuracy varies too much depending on the speaker. ASR will be useless so long as there are speakers with low recognition accuracies. (13 votes)

(d) The vocabulary is still too small. (11 votes)

(e) ASR with vocabulary constraints is not useful. Users do not feel comfortable unless arbitrary speech can be converted to text. (4 votes)

(f) A word spotting technique is essential. (18 votes)

(g) The biggest problem with ASR is poor rejection ability against erroneous and extraneous utterances. (40 votes)

(h) Robustness against noise, utterance and channel variabilities is too poor. (46 votes)

(i) The human interface is too poor for users to utilize ASR comfortably. (41 votes)

(j) Good dialog control is important: ASR is not being used because of poor dialog control. (22 votes)

(k) Humans hate to speak to machines. People intrinsically dislike ASR. (7 votes)

(l) It is not clear to users how recognition errors occur and how they can be avoided. This distransparency (opacity) greatly annoys users. (41 votes)

(m) Efficiency is not higher than other existing input methods. (17 votes)

(n) A good interface (API) has not been established between ASR and applications. (24 votes)

(o) The cost problem – ASR technology is currently too expensive. (11 votes)

(p) The speed (response time) of ASR is not sufficient yet. (6 votes)

(q) Applications are limited because the microphone has to be held close to the mouth. (10 votes)

(r) We lack good ideas for ASR applications. Good ideas would make it possible to fully utilize ASR at the current technological level. (37 votes)

(s) Linguistic processing is the problem. Even though acoustic processing has reached a considerably high level, language models, or linguistic processing, (e.g., syntactic analysis, semantic processing, pragmatics, knowledge, common sense) are far behind. (18 votes)

(t) The requirements ASR must meet at are often too demanding. ASR technology went from simple to more complex; so should the applications. (0 votes)

(u) Humans should be trained in how to use ASR. Keyboards required more than a day. (6 votes)

(v) There is no social consensus how ASR should be used. (11 votes)

(A) ASR researchers are not serious about the ASR into practical use. Belief and devotion are essential in technology development. (22 votes)

(B) Intelligent dialog is not the problem; instead, intelligent service provided through ASR *is* important. (21 votes)

(C) The following 3 items are not going well yet:
1. The achievment of a certain level of ASR performance and functionality (depending on the application).
2. Availability of support tools that enables the user (system integrator) to easily integrate ASR into a system.
3. The development of specific applications that utilize ASR. (e.g., Computer-aided Instruction) (34 votes)

(D) Progress in ASR research is not sufficiently reported to the public. In addition to fundamental research, we need good ideas, technologies, devotion, and capital to use ASR in products and services to build up the market. (7 votes)

(E) ASR is a great technological human dream. The ideal picture is much too far ahead of the reality. This is the reason why ASR is not well reputed. The world needs to change its thinking. (4 votes)

(F) No manner has been established to fill the gap between the mental model (how the system is perceived by the user) and the real system. (36 votes)

From the above votes, one can see that the robustness issue was of most concern to the attendants. These explanations for low public acceptance were compiled almost three years ago, so some of them now seem trivial (e.g., the cost of ASR is not a big issue thanks to rapid progress in personal computers). Most, however, are still interesting and stimulating.

2.2. Advantages of ASR

The benefits ASR promises are the points we should always come back to and have in mind every time we discuss ASR problems and issues. In attempt to clarify some of these benefits show how to get the most use out of current ASR technology, the author wrote in [2]:

Telephone channels: A very important aspect ASR allows information exchange through telephone lines without any special equipment at the users' side. The ANSER system was the first service in this area.

Comfortableness: It has long been assumed that users will no longer need keyboards (which are cumbersome for Japanese people) when they have ASR capability. In addition to this advantage, sophisticated dialog control and system implementation of ASR applications provide more comfortable communications between humans and machines than a keyboard and mouse.

Case of no alternatives: ASR is an effective way to enter information particularly when the hands are occupied with another task. One potentially large market is in car navigation systems where eyes and hands are occupied while driving. The physically challenged and senior citizens represent another important market.

Efficiency: It has been said that ASR provides faster information entry than typing. Multimodal input seems more effective for higher efficiency, i.e., spoken commands with drawing.

Synergy: Synergy with other input modes is very important. Not ASR alone, but ASR with a keyboard is the right picture. Both hands and voice will be used for work efficiency rather than usability.

Downsizing: The merging computers and mobile telephones will mean removing keyboard and display to minimize the size. Speech input/output functionality is the only possible way to do this.

Low cost: In the future, there will be a very small inexpensive computers consisting of a CPU, a microphone, an earphone, and a battery with ASR functionality. Computers equipped with a keyboard and a display will be considered high-class but expensive. Computers with ASR will be popular only because they are small, light, and cheap.

Fun: ASR is fun. In game machines, the game will be made more exciting by talking with the machines.

Conversation with a robot: Not classic robots, but virtual persons in cyberspace. These virtual persons must be able to talk and listen to the human user. ASR will be vital for clerks to talk with users in "cybershops".

Low cost and downsizing were new views several years ago, but they sound quite reasonable now due to boosts in personal computers speed.

3. ACOUSTIC AND TRANSMISSION ISSUES

Acoustic issues in ASR include robustness against variability of transmission characteristics and of the characteristics of telephone microphones including those used in mobile telephones.

3.1. Multi-microphone, Room Acoustics

For hands-free and position-free ASR, ASR should be combined with acoustic signal processing. It may be possible to obtain even better speech recognition ability than humans have in a reverberant room using an array microphone. There are big possibilities of gaining significant improvements by combining techniques other than speech recognition techniques.

3.2. Accurate Voice/Noise Discrimination

It is often pointed out that speech detection errors cause serious degradation of overall ASR performance. In some applications, such as the electronic secretary, voice/noise discrimination for accurately capturing speech input is particularly important. Speech detectors should be accurate and computationally inexpensive so that they can be run continuously until the computationally intensive ASR is activated.

3.3. Telecommunication Issues

In a discussion with R. Rose of AT&T, he pointed out the following telecommunications issues:

Customer Premises Equipment: How will the evolution of Voice/Data telephone terminal equipment along with computer/telephony integration affect ASR?

Globalization: Telecommunications services and technology suppliers are expanding internationally. What are the issues associated with providing service offerings to a multi-lingual, multi-cultural customer base? What are the issues associated with supporting a variety of networks all with different transmission characteristics?

Network Architecture: Depending on the network architecture, a speech recognizer might be placed in different parts of the network or its functionality may be distributed. We may perform feature analysis in a handset and transmit features. What are the possibilities, and what new problems arise as a result?

Transducers and Transmission Channels: Will the variability of channels and transducers decrease as a result of an increase in digital transmission lines, or will it increase as a result of a larger number of service providers?

As for the third problem, mobile telephones actually contain speech analysis as a part of digital coding of speech. We can think of a conversion algorithm from coded speech into speech recognition feature parameters.

4. SPEAKER AND ENVIRONMENTAL ISSUES

Robustness against speaker and environmental differences between the model and the target speech has been one of most concerned problems in ASR research. Speaker, environmental noise, and channel adaptation is recently investigated actively. In speaker-independent ASR, the out-lier problem is one of most important issues and it is still true that "there is no data like more data" for training acoustic models.

4.1. Children's and Seniors' Voice

The performance of ASR with young children's speech is often discouraging. One reason is that it is generally difficult to collect a sufficient amount of training data from young children because they can not read utterance texts nor maintain a stable speaking style throughout the data collection session. This contradicts the expectation that ASR is supposed to be particularly useful to people who can not read and write texts. This performance degradation is also seen with senior citizens, another group to whom ASR is supposed to be useful.

4.2. Instantaneous Adaptation to Environmental Noise

In telecommunications applications, every time ASR is used, it could be by a different speaker in a different environmental noise. Even good acoustic models for speaker-independent ASR may not provide acceptable performance with different background noise. Existing techniques for adaptation of acoustic models to environmental noise (e.g., NOVO or PMC) are often computationally too expensive to execute the algorithm every time an utterance is entered. We need a novel algorithm that lets the acoustic models adapt instantaneously as to make them ready to recognize the input speech well before the recognition phase.

One possible algorithm for this purpose is based on the Jacobian adaptation (JA) of acoustic models [3] which requires small amount noise observation and computational cost for good performance. In noisy speech recognition at 10 dB SNR, JA experimentally gave even better performance with a 0.2-second noise observation than NOVO (or, equivalently, PMC) with a 60-second noise observation, while having only 1/34 computational complexity.

In contrast to the existing idea of composition of clean speech HMMs and noise HMMs (e.g., PMC or NOVO), the JA approach is based on an idea of "adaptation of HMMs with noise A to HMMs with noise B". The composite cepstrum of speech and noise is a non-linear function of the speech cepstrum, $\boldsymbol{C}_S$, and the noise cepstrum, $\boldsymbol{C}_N$, such that

$$\boldsymbol{C}_{S+N} = \boldsymbol{F}^{-1}\log(\exp(\boldsymbol{F}\boldsymbol{C}_S) + \exp(\boldsymbol{F}\boldsymbol{C}_N))) \qquad (1)$$

where $\boldsymbol{F}$ is the Fourier transform operator. When $\boldsymbol{C}_N$ slightly changes to $\tilde{\boldsymbol{C}}_N$, the composite cepstrum also slightly changes from $\boldsymbol{C}_{S+N}$ to $\tilde{\boldsymbol{C}}_{S+N}$ and these small changes are related by a linear approximation using a Jacobian matrix:

$$\tilde{\boldsymbol{C}}_{S+N} = \boldsymbol{C}_{S+N} + \frac{\partial \boldsymbol{C}_{S+N}}{\partial \boldsymbol{C}_N}(\tilde{\boldsymbol{C}}_N - \boldsymbol{C}_N) \qquad (2)$$

where the Jacobian is obtained by

$$\frac{\partial \boldsymbol{C}_{S+N}}{\partial \boldsymbol{C}_N} = \boldsymbol{F}^* \frac{\boldsymbol{N}}{\boldsymbol{S}+\boldsymbol{N}} \boldsymbol{F} \qquad (3)$$

from Eq. (1) where $\boldsymbol{S}$ and $\boldsymbol{N}$ represents the speech and noise spectra. The Jacobian matrices are calculated beforehand in the training phase, a likely noise cepstrum being assumed. In recognition, the small difference between the assumed and actual noises is detected and used to compensate the model parameters using Eq. (2).

If the environmental noise is observed for a short time (e.g., 0.2 sec at the pause in the guidance) after the telephone is connected to the user, the acoustic models are promptly adapted to the environmental noise and become ready to recognize the input speech. Adaptation of cepstral mean vectors is most effective while delta-cepstral means and variances can also be adapted [3]. In the case of barge-in, however, the noise observation timing is not a simple problem.

4.3. HMM Training with Localized Data

When a dialog is spoken between the user and the system, the system will obtain a certain amount of user's speech. It is desirable to utilize this to improve the recognition accuracy using the speaker adaptation technique. However, the training data are often too localized, i.e., they cover only a small variety of the vocabulary and do not balance well for training the acoustic model. One issue here is HMM training with very localized training data. This issue will become more important when the ASR function is equipped in more personal computers as a common input modality.

4.4. Speaker Consistency

There is no speaker-independent speech signal; all speech signals are speaker-dependent. In most ASR systems, this fact is entirely neglected and "speaker-independent" acoustic models are used.

The speaker is consistent throughout an utterance. A speaker change half way through the utterance of a sentence is quite unlikely. This fact should receive more attention. One simple application of this principle is the gender-dependent model where the speaker's identity is only male or female. Only few studies have been done on the speaker consistency principle. One has investigated "speaker-consistent parsing", where the phoneme sequence and the speaker class are simultaneously parsed to find the best hypothesis [4].

5. UTTERANCE ISSUES

Utterances are often made in different ways from those expected or assumed in the design stage, e.g., sometimes too casually and sometimes too nervously. Robustness against different ways of utterances is important. Out-of-vocabulary words, extraneous words, barge-in, the Lombard effect and other phenomena related to spontaneous speech have been actively explored recently.

5.1. Breath Noise, Handling Noise, etc.

ASR errors are often found in degraded utterances with breath noises. The errors are caused by turbulent air flow onto microphones and telephone set. The speech signal is partly masked by a loud noise, other parts being normal. This situation looks like it requires a new search technique that skips the degraded parts and traces the reliable matching path.

Another problem is noise made by friction between the user's hand and the telephone handset. This is a practical problem like coded speech in mobile telephones. It is not clear at present whether we can derive a specific technique for the recognition of digitally coded speech.

5.2. Utterance Verification or Accurate Rejection of Erroneous Utterances in the Flexible Vocabulary Framework

Rejection of extraneous utterances (e.g., eh, ah, well, hmm, laughing, coughs, sighs, ...) and out-of-vocabulary words is sometimes even more important than the high ASR accuracy within the vocabulary for a reliable and smart dialog between the user and the system. This problem involves verification that the recognition result is reliable and verification that the utterance is really within the vocabulary, or in other words, the detection of recognition errors (confidence measure), and the detection of out-of-vocabulary utterances. If these detections can be done with a very high accuracy, even isolated-word ASR will gain a wide range of applications through a reliable dialog.

This issue has often been discussed in past research papers on utterance verification and rejection. However, the task is often fixed and acoustic model training is done to enhance the utterance verification power within a closed domain.

For wider application capabilities, it is absolutely necessary that the system has both flexible vocabulary capability and accurate utterance verification ability. However, this is not a simple problem. The phone-based ASR approach is essential for flexible vocabulary, but it is not easy to enhance the discrimination power between the vocabulary words and the out-of-vocabulary words in terms of phone model training. Once the active vocabulary is changed, the vocabulary-dependent discrimination power is not necessarily applicable to the new vocabulary. This will be one of most important problems in the area of acoustic model training.

With accurate utterance verification, reliable services are possible by combining ASR and human backup. If the utterance verification function detects an utterance not recognized by ASR, it is supported by the human operator.

6. DIALOG ISSUES

ASR robustness for information entry is another important aspect. Dialog control includes a number of interesting unsolved problems. We discuss a few of them here.

6.1. Adaptive Dialog Strategy

In dialog between the user and system, the user becomes skillful in handling the ASR function. When we speak to another person, we estimate the listener's background knowledge and experiances and adjust our speaking style, speed, and prosody (e.g., in dialog between a customer and the customer service center). The dialog system also should have the ability to adaptively change its dialog strategy depending on the user's skill and experiance. When the system can estimate the user's skill and experiances, it can skip lengthy guidance procedures and can encourage the user to utilize the barge-in function. How to estimate the user's skill and experiances through dialog is a new challenge.

Barge-in timing is another related problem. If the system allows barge-in, the timing of switching from training the echo cancellor to accepting the barge-in utterances is an important factor in dialog design.

6.2. Distransparency

One problem is the big gap between ASR errors and human-felt distances. The ASR system sometimes gives a recognition result that looks very strange and very unlike what the user expected as the output. This is often due to breath noise and other causes which the user is unconscious of. This sometimes confuses the user. Since ASR aims at a good human interface, ASR errors should also be human-like – at least, they should be reasonable to the user.

Among the possible hindrances discussed in 1994, one convincing explanation is "distransparency". Even if users work very co-operatively with the ASR system and are very motivated to avoid the recognition errors and ready to change their utterance styles, they do not know how they can help the system.

All successful systems are transparent at certain level, namely, users can form mental models (what they should do to get what they want). Users can think of ways to improve system utility (throughput), controlling themselves according to their "mental model". Intelligent systems that are not really intelligent, such as ASR, present a serious problem. It is generally difficult to model from the users' point of view what the relationship is between the input and the output.

When an ASR error occurs, users often try to improve their utterances by speaking more clearly and loudly. When the same error, or a difficult error occurs, they speak even more clearly and more loudly; too clear and too loud for correct recognition, which results in a hopeless vicious circle. An important question is how we will be able to make speech recognition systems transparent to users.

6.3. Speech-to-Text Conversion

Potential ASR users often have the illusion that ASR be able to convert whatever the user speaks into text because this is fairly true for human listeners. The common way of ASR usage, where the vocabulary and the grammar are given beforehand, narrows the imagination of users' application possibilities.

Just like text-to-speech technology, speech-to-text technology without defining the vocabulary and grammar is one of most awaited technologies, whatever its implementation may be. We do not know how such a system can be realized. A simple phonetic decoder can produce a transcription of input speech, but with a number of errors.

6.4. Spell-Out Method for Japanese Kanji Texts

One problem with the Japanese language is that Kanji (Chinese-origin) characters are essential for describing and exchanging ideas. The rich phonetic variety of Chinese has degenerated into the numerous homonyms of Japanese Kanji characters over time. The same thing is happening in this century for English words. English pronunciations are incorporated into the Japanese phonetic system and become numerous homonyms (e.g., *vote*, *boat*, *bort* and *bought* are equally pronounced). But, the problem with Kanji characters is much more serious. There is no alphabetical labels for Kanji characters, nor are their common phonetic codes as in the English system: *alpha, bravo, Charlie, delta, echo, foxtrot, golf, hotel.* Therefore it is not trivial for Japanese to transfer Kanji spellings to others in the spoken language. Describing Kanji spelling in spoken language is very awkward[1].

As Japanese language has no way of spelling out Kanji words, people try to describe the Kanji spelling using the common knowledge between the two parties. Therefore, a very intelligent automatic dialogue system for accepting Kanji spelling from the user is awaited. If one could be created, even though it might run very slowly, it would be a very robust communications channel that would allow us to tell the machine the Kanji text.

In applications of speech synthesis, "describing Kanji characters in voice" is being actively explored for remote access to Kanji-based data and the vision-impaired.

[1]For example, a relatively rare Japanese personal name *Sagayama* is spelled as "嵯峨山" in Kanji. It is often described as follows: the first character "嵯" /sa/ consists of "山" /yama/ for "mountain" at the lefthand side and "差" /sa/ for "difference" at the righthand side; the second character "峨" /ga/ consists of "山" /yama/ for "mountain" and "我" /ware/ for "myself"; the last character is "山" /yama/ for "mountain" alone. Alternative possible descriptions are: the first 2-character part "嵯峨" /saga/ is spelled same as the name of a popular touristic place in Kyoto; or "嵯峨" /saga/ is identical to the name of a past Japanese emperor almost 1200 years ago.

7. PC-BASED SYSTEM ISSUES

7.1. Electronic Secretary Metaphor

In future telecommunications, personal computers will play the crucial role as the agent mediating between the user and the telecommunications network. Concerning PC-based ASR applications, one of largest market exists in the computer-telephony integration (CTI). "CTI" has multiple meanings. One of the multiple aspects of CTI is that computers and telecommunications will be combined more tightly where ASR will be found useful as a user-friendly human interface.

As for the question: how will the the evolution of voice/data telephone terminal equipment along with the computer/telephony integration affect ASR, one answer is an electronic secretary metaphor. As a result of the integration of high-speed PCs and telephony, ASR will be one of the basic functions of personal systems. When numerous application softwares are combined with the ASR function, we need a simple metaphor that provides a unified idea of a user-friendly interface between the user and PC. The electronic secretary as a speech I/O agent for the computer-telephony system human-machine interface is based on such an idea.

With an animated image of a secretary, it accepts the user's requests such as "Call Aikawa-san", "Do I have E-mail messages?", "What is today's schedule?", etc, through a smart dialog. At the first stage, ASR will be used as an alternative to mouse and clicked menus. For example, the complicated mouse clicks in a the hierarchical menu and scrolling numerous menu items will be replaced by a simple "click by voice". This will be one of the secretary functions. The electronic secretary is a convenient platform on which future functions will be implemented just as human secretaries keep acquiring new skills of office works.

7.2. Cyber Society

In the near future, the so-called "cyber society" will become popular. In a cyber-shopping mall, the user can go around the mall and enjoy shopping while talking with shop clerks. That means a cyber-clerk should have an animated human image as well as the ability to talk with customers using ASR. This issue involves not only continuous speech recognition but also spontaneous speech phenomena, conversational dialog control, speech synthesis with a wide range of speaking styles and prosody control, and intelligent natural language processing. Intelligent robots in cyber space may be the largest market for ASR in the future.

8. CONCLUSION

We discussed some relatively new issues related to ASR for telecommunications applications. Some of them will be important in the future of ASR.

We would like to express special thanks to Rick Rose, AT&T Labs, for discussing telecommunication issues, and Mikio Kitai, NTT Human Interface Labs, for discussing dialog issues in ASR.

9. REFERENCES

[1] Shigeki Sagayama: "Why Is Speech Recognition Not Widely Used ? How Can It be Used More Often ?," IPS Technical Report, Vol. 94, No. 40, pp. 23-30, 1994. (in Japanese)

[2] Shigeki Sagayama: "Discussion toward the Practical Use of Speech Recognition Technology," Vol. 36, No. 11, pp. 1047–1053, Nov 1995. (in Japanese)

[3] Shigeki Sagayama, Yoshikazu Yamaguchi, Satoshi Takahashi, and Jun-ichi Takahashi: "Jacobian Approach to Fast Acoustic Model Adaptation," to appear in Proc. IEEE International Conference of Acoust., Speech, and Signal Processing (ICASSP97) (Munich), Apr 1997.

[4] Kouichi Yamaguchi, Harald Singer, Shoichi Matsunaga, Shigeki Sagayama: "Speaker-consistent Parsing for Speaker-independent Continuous Speech Recognition," Trans. IEICE Inf. & Syst., Vol. E78-D, No. 6, June 1995.