R&D Spirits

Laying the Foundation for New Communication Technologies

Dr. Shigeru Katagiri
Director
NTT Communication Science Laboratories
NTT R&D Fellow

NTT Communication Science Laboratories is a place of diverse research activities toward information-communication that reflects the very essence of human communication. One area of particular interest is speech recognition. NTT's speech recognition technology boasts advanced realtime capabilities and the world's largest vocabulary. The director of these laboratories, Dr. Shigeru Katagiri, is a recognized authority in speech recognition and an NTT R&D Fellow. We asked him about the current state of progress in speech recognition and what benefits it might provide to society.

Pattern recognition research: endowing computers with intelligence

—Dr. Katagiri, could you first tell us about the mission of NTT Communication Science Laboratories?

As the name implies, our mission is to perform research in the field of communication science. However, as these are research laboratories of a telecommunication carrier, the approach that we take does not fit the typical "academic" image centered on linguistics, psychology, and the like. Rather, we take a "scientific" approach that aims to clarify human communication functions and develop technologies with the knowledge obtained. Today, "communication" is coming to include not only communication between people but also communication between people and things and between things themselves. In this regard, we might ask ourselves how society could be improved for both people and the environment and how communication technology might be used to this end. Finding answers to these questions is the mission of NTT Communication Science Laboratories. We undertake this research with great determination based not on existing concepts but on this definition of communication science as an evolving field of study.

—What is your personal research theme?

I have been researching pattern recognition from the start with the ultimate goal of giving computers the ability to hear and comprehend on the same level as human beings. To be more precise, I have been researching acoustic models to convert speech to text. In this process, the sound picked up by a computer's microphone is compared and judged against an acoustic model in a database and converted to text. But since the accuracy of this conversion is only as good as that of the acoustic model, the problem here is how to prepare an accurate model. I have been involved in the construction and development of techniques for this purpose.

—How will such research benefit society in the future?

For human beings, speech is a very easy-to-use medium. If the speech recognition performance of computers can be improved, we can expect various ripple effects. Take, for example, an NTT call center that receives telephone inquiries from customers. If the primary processing of this call center can be mechanized, we can expect significant cost savings to result. We can also envision a speech recognition system that monitors the handling of customers and collects material for the creation of call-center guidelines. An advanced version of such a system could issue a warning in real time whenever an inappropriate response is made to a customer. In addition, techniques for comparing and judging entered informa-

tion against a model are not limited to speech recognition—they have much in common with various kinds of pattern recognition. Accordingly, if an effective technique for speech recognition can be established, it should be possible to apply it to image and text recognition as well. That would mark the beginning of intelligent computers and the coming of dramatic changes to all social systems.

—What are some technical points of interest in the pattern recognition research that you have been involved with?

The most interesting aspect of this research would probably be the generalized probabilistic descent (GPD) method, which is one type of discriminative learning method. This method is somewhat complex, but I would like to give you an outline of it here.

Assume that the following statement has been made: "This is a writing tool. Is it a pen or pencil?" To answer this question, we could create a model of a pen and one of a pencil, and then compare and judge the object against these models using various functions. There are various ways of going about this, but the most common approach is to create representative models of a pen and pencil and measure the likelihood that the input pattern agrees with each of these models as a basis for making a decision. This is called the "maximum likelihood estimation" method. But the most important thing in pattern recognition is not models *per se*, but rather the classes that individual objects belong to, and in particular, where the boundaries of those classes are located. In our example, we would need to accurately determine where the class of a pen ends and where the class of a pencil ends. The more accurately such boundaries can be determined, the less we have to be concerned with the problem of whether a model adequately represents a certain class of things. We therefore abandon this idea of maximum likelihood estimation that assumes the modeling of average patterns for different classes. Instead, we apply a learning process to input patterns so that models can grow and more accurate boundaries can be drawn, resulting in a minimization of classification errors. This is the idea behind our discriminative learning method.

I first became engaged in the research of discriminative learning methods at the end of the 1980s. At that time, discriminative learning methods could not be directly applied to the recognition of speech signals, which are actually time signals. This prompted us to look for a method that could overcome this problem. In the end, we extended a classical learning method called "probabilistic descent" developed in the 1960s so that it could be directly applied to hidden Markov models (HMMs), which are suitable for representing time signals, and developed a design method that could minimize pattern classification error. This is our GPD method. In other words, despite the fact that speech signals are very difficult to process since their patterns have different temporal lengths, this GPD method can handle them directly. It also provides a theoretical guarantee that a model can be trained, or designed, in principle to minimize recognition error. And though computation is heavy, the recognition accuracy that the GPD method can achieve is dramatically higher than that of past techniques. I think the IEEE award and the fellowship that I had the honor of receiving for this achievement reflect the value of this work.

Misclassification measure

$$d_k(x;\Lambda) = -g_k(x;\Lambda) + \left[\frac{1}{M-1}\sum_{j,j\neq k}\{g_j(x;\Lambda)\}^{\mu}\right]^{1/\mu}$$

Smooth misclassification loss

$$l_k(x;\Lambda) = \frac{1}{1+\exp(-\alpha d_k(x;\Lambda)+\beta)}$$

Probabilistic optimization mechanism

$$\delta\Lambda(x(t),C_k,\Lambda) = -\varepsilon\nabla l_k(x(t);\Lambda) \qquad \Lambda(t+1) = \Lambda(t) + \delta\Lambda(x(t),C_k,\Lambda)$$

$$L(\dot{\Lambda}) = \sum_k \int_{\Omega} p(x,C_k) l_k(x;\dot{\Lambda}) 1(x\in C_k) dx$$

$$\simeq \sum_k \int_{\Omega} p(x,C_k) 1(x\in C_k) 1(p_{\dot{\Lambda}}(C_k|x) \neq \max_j p_{\dot{\Lambda}}(C_k|x)) dx$$

$$\mathcal{E} = \sum_k \int_{\Omega_k} p_{\dot{\Lambda}}(x,C_k) 1(x\in C_k) dx \qquad \Omega_k = \left\{x\in\Omega \middle| p_{\dot{\Lambda}}(C_k|x) \neq \max_j p_{\dot{\Lambda}}(C_j|x)\right\}$$

Fig. 1. Essence of the formalization of the generalized probabilistic descent (GPD) method.

—How is this research progressing?

Well, we are researching various aspects of speech recognition, but we are devoting particular effort to a new formulation of the GPD method. Our past approach was to achieve a learning process that could minimize classification error within a framework like the HMM that can handle time signals, but this was not very applicable to the control of design tolerance, that is to say, of statistical stability. To solve this problem, we are attempting to redefine the GPD method with a new parameter space of fixed dimensions. The completion of this undertaking should make it easier to use the GPD method and to make it a more general-purpose discriminative learning method.

And though this is more the work of my colleagues than my own work, we have been developing a framework for representing models using a finite-state transducer (FST) with the aim of achieving advanced speech recognition functions. This framework is evolving into a mechanism that can recognize about 2 million words. The current level of large-vocabulary speech recognition systems in the world ranges from 200,000 to 300,000 words, which makes NTT the only organization with a system that can recognize 2 million words. Considering that "Kojien," Japan's standard dictionary, contains about 800,000 words, the recognition of up to 2 million words would enable the handling of proper nouns such as personal names and place names that in the past were often treated as "unknown words." A model representation on such a huge scale as this one should be able to express various types of knowledge in an integrated fashion. We believe that our research should not only improve the accuracy of speech recognition but also bring about major innovations in computer-based communications.

International interest in an originally developed GPD method

—Dr. Katagiri, please tell us about trends in this research field both here in Japan and overseas.

Research in this field is active overseas, particularly in the United States. The International Conference on Acoustics, Speech, and Signal Processing (ICASSP) sponsored by IEEE includes an annual session on discriminative learning methods, and research that follows in the footsteps of our work must certainly be presented there. As for Japan, I must confess that research in this area has dropped off. The last half of the 1980s and the first half of the 1990s was a boom period for our research and for neural network research, which are closely related, and a good number of sessions on discriminative learning methods were held during that time. At present, it appears that the one organization making steady progress in this area is NTT.

By the way, there is a great difference in the way that discriminative learning methods have been approached in Japan and overseas. I don't know the real reason for this, but I have a feeling that differences in computation power lie at the root of it. In the United States, connecting personal computers (PCs) together to perform parallel processing has become commonplace, while in Japan, the conditions for doing that have not yet been established. To be sure, PCs have dropped in price, but Japan is still weak at putting together such configurations. Perhaps methods like discriminative learning are being avoided because of the heavy computation that they require. At any rate, I would speculate that some kind of revival in the research of discriminative learning methods is needed in Japan.

—What kind of response have you received in Japan and overseas to your research?

Since I received the Signal Processing Society Senior Award from IEEE in 1994, my research has been receiving a certain amount of attention. However, the path to this recognition was never easy. The first time I presented a paper on GPD was in 1990 as a joint work with researchers at Bell Laboratories. My coauthors and I were very confident about our work, and we submitted it to ICASSP, the most prestigious forum in this research field. At that time, however, papers were initially evaluated based on a summary of only a few hundred words, which was hardly enough to convey the importance and novelty of our research to the reviewers. In the end, the paper was completely rejected, and the only presentation that we could make of the work that year was at the Acousti-

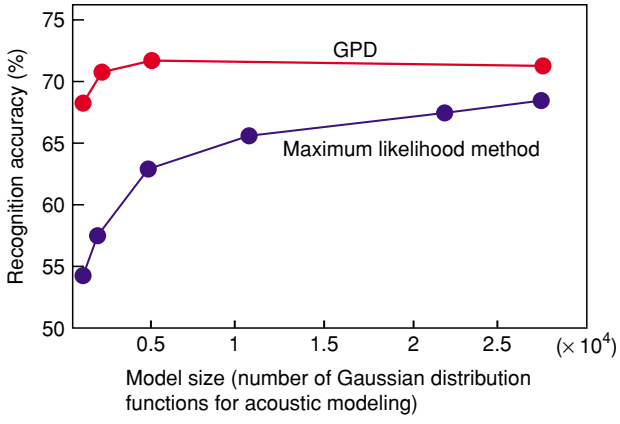

Fig. 2. This figure illustrates that GPD consistently produces higher recognition rates than its traditional counterpart, the maximum likelihood method, in an example word recognition task.

cal Society of Japan. But even there, other papers presented at the same time received most of the attention and there was little response to our GPD method. After that, we continued to receive little recognition, so the period was somewhat depressing. Fortunately, things began to change after the IEEE Award.

—Are you involved in any collaboration with outside institutions?

One great feature of NTT Laboratories is to be as open as possible to the outside, and I can say that we have been involved with much collaboration. For example, we have a very close cooperative relationship with MIT and Stanford University. We have also invited a number of renowned researchers to serve as research professors here. These include Professor Fumitada Itakura of Meijo University, my former senior at NTT Laboratories and recipient of the Asahi Prize in January of this year, Professor Nobuo Masataka of Kyoto University Primate Research Institute and author of "Monkeys with Cell Phones" (Chukoshinsha), and Professor Fred Juang of Georgia Institute of Technology. In addition, many of our members have presented lectures at various universities. As for myself, I served as a guest professor at the Graduate School of Kyoto University up until the year before last, and I have been a lecturer at Doshisha University since last year.

The excitement of understanding: the motivation behind research

—Dr. Katagiri, what was your major at university?

Well, my undergraduate degree was in electrical engineering, but in graduate school, I was a member of Professor Ken'ichi Kido's research laboratory that was working on speech-information processing. I think I became interested in human-oriented research as opposed to simply electrical technology because of Professor Kido's course related to human-machine interaction. Many students wanted to join his laboratory, but as enrollment was limited, attendees were decided by playing "rock paper scissors," which is an old tradition at the university for situations like this. If I had lost, perhaps my life thereafter would have been different.

—Please tell us something about your research history after entering NTT.

On entering NTT, I was first assigned to Section No. 4, Basic Research Division at Musashino Electrical Communications Laboratories, which was a speech-research group that NTT took great pride in

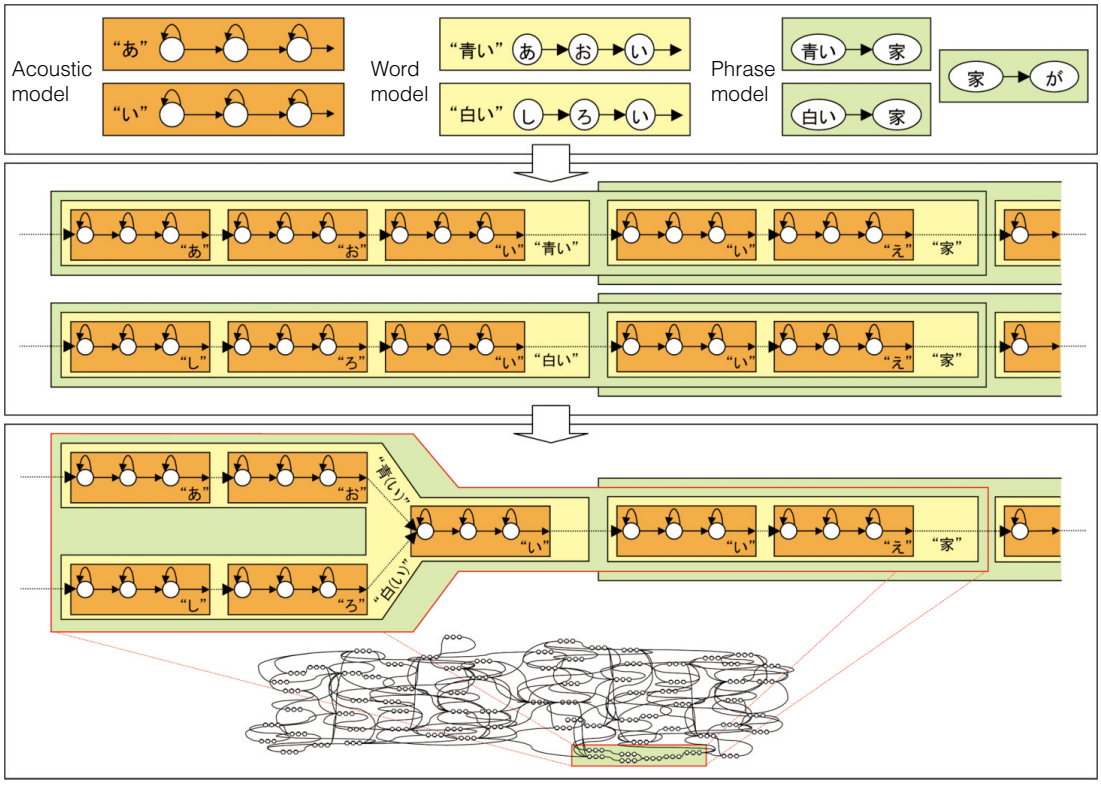

Fig. 3. This figure illustrates the behavior of the finite-state transducer that is currently studied at NTT Communication Science Laboratories for large vocabulary speech recognition.

for its achievements in speech coding and telephony standardization. The head of the group at that time was Professor Itakura, who I mentioned earlier. Next, when the Advanced Telecommunications Research Institute International, or ATR, was established, I was assigned to hearing research in ATR Auditory and Visual Perception Research Laboratories. There, I conducted research as the only member of a section set up through the good offices of Yoh'ichi Tohkura, who is now the Deputy Director General of the National Institute of Informatics (NII). After that assignment, I worked at Bell Laboratories for a nine-month period beginning in 1989, and then returned to NTT, where I was able to devote myself to research for several years. But since becoming an NTT R&D Fellow, I have worked only in the Research Planning Section, and I am currently serving as director of Communication Science Laboratories. Accordingly, I can't really say that I have had many periods in which I could devote myself to pure research. I sometimes wonder whether I deserved being named an NTT R&D Fellow.

—What do you feel is so interesting about research work?

The excitement of finally understanding a difficult problem—when you finally say "I got it!"—is what makes research a very interesting endeavor. I have experienced that wonderful spine-tingling feeling at the point of enlightenment any number of times during my research career. For example, I had that exact sensation when I was thinking about the GPD method and I suddenly realized that I could turn it into a methodology if I approached it in a certain way. That was immensely interesting, and experiences like that provided me with the motivation to continue my research work.

—How do you think the research of speech recognition will develop from here on?

To begin with, we would like to make the GPD method easier to use. To be more specific, we would like to reduce the amount of computation that it now requires as much as possible and would like to make it so that anyone can use it without the need for extensive "tuning" know-how. In speech recognition, however, it is not sufficient to simply improve the performance of an acoustic model no matter how far that model can be improved. In actuality, human beings understand the meaning of sounds that they hear by taking into account various limitations possessed by language. Japanese people, for example, listen to and understand each other while unconsciously holding to the rule that a consonant cannot be used by itself in the Japanese language. In addition, connections based on context are extremely important in speech recognition. I therefore think that design work must also consider how a language model should be created and how tasks should be divided. A part of our GPD method has been based on this approach and the plan is to expand this line of thinking across the entire method. If this can be achieved, I think that our speech recognition technology cannot help but be incorporated in the outside world.

NTT Laboratories: the last treasure of the 20th century

—Dr. Katagiri, what are your plans for the future?

For the near future, I think we should be able to solve some of the crises that we now face using telecommunications. That might sound like an exaggerated claim, but it's a sincere feeling. For example, it should be possible to provide an effective response to today's serious energy and environmental problems through the use of videophones and the Internet. But despite the fact that a broadband infrastructure is spreading rapidly, business trips for the purpose of attending meetings are still fairly frequent, and the reality is that videoconferencing is not functioning well. Nevertheless, if telecommunications were to take root in the real world, there would be no need for people to move from one location to another for meetings. This would make a substantial contribution to alleviating the energy problem. To this end, it is essential that we develop telecommunications to the point where people can share each other's space.

Telecommunication technology should also be useful in the revival of Japan's industrial sector. Given that the economies of neighboring countries are capable of rapid growth as we have seen, an industrial structure that depends on cheap overseas labor will reach its limit sooner or later. To overcome this problem, we must pursue a knowledge-intensive type of production. I believe that telecommunication technologies involving pattern recognition, statistical learning, and knowledge processing provide a foundation for knowledge-intensive technologies and hold the key to the success or failure of new-generation production.

—In your opinion, what is the significance of NTT Laboratories?

As a researcher, I think it's the last treasure handed down from the 20th century. Let me explain. Bell Laboratories and NTT Laboratories were the finest research institutions in the 20th-century world of telecommunications. And Bell Laboratories can be called an American treasure of the 20th century. But

as time passed, Bell Laboratories went through major changes, and the only research treasure left became NTT Laboratories. Even if we broaden our viewpoint here to include research institutions that are holding on to the traditions of the 20th century, I can only think of the IBM Watson Research Center. Perhaps researchers who are constantly trying to innovate will find this emphasis on "tradition" strange, but if we exchange it with the words "know-how," its importance suddenly becomes clear. Fortunately, NTT continues to place importance on basic research as a corporate policy. It supports basic research in a variety of ways, starting with the provision of research funds. From my position as an on-site manager, I consider the preservation of NTT tradition to be very important work.

—Dr. Katagiri, could you leave us a message for young researchers?

I'd be happy to. Taking my research as an example, it is more important in research to search out boundaries rather than models that usually represent central regions in place of boundaries. Many seeds of new research can be found in boundary regions rather than in the provision of models. With this in mind, I wrote "Please look around you" in an e-mail that I sent out to all staff members of the Communication Science Laboratories on my first day as director. What I meant by this was to look for the seeds of innovation in the boundaries of your own region and in the boundaries of your colleagues' regions. NTT Laboratories has a wide spectrum of talented people, and I want young researchers to look at the people around them and consider how their work might benefit their own work. If such a frame of mind can be established, I think it will have great value in forming teams made up of professional researchers.

Interviewee profile

■ Career highlights

Shigeru Katagiri received the B.E. and M.E. degrees in electrical engineering and Dr. Eng. degree in information engineering from Tohoku University, Sendai, Japan in 1977, 1979, and 1982. From 1982 to 1986, he worked at the Musashino Electrical Communication Laboratories, Nippon Telegraph and Telephone Public Corporation (now NTT), Tokyo, Japan, where he was engaged in speech recognition research. From 1986 to 1998, he was with the Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan. At ATR, he worked on various speech-related research issues that included speech recognition, auditory scene analysis, and spoken language acquisition. From 1997 to 1998, he headed the Hearing and Speech Processing Research Department at ATR Human Information Processing Research Laboratories. Since 1999, he has been with NTT Communication Science Laboratories (CS Labs), Kyoto, Japan, where he has been engaged in a wide range of machine-learning research. There, he served as Supervisor of the Research Planning Section for two years. Currently, he occupies the position of Director at NTT CS Labs. From 1989 to 1990, he was a visiting researcher at the Speech Research Department, AT&T Bell Laboratories, Murray Hill, NJ, USA. He has also served as an Adjunct Professor at the Graduate School of Kyoto University (1998-2004), and as a lecturer at both the Graduate School of Nagoya University (2001-2003) and Doshisha University (2004-present). He is an IEEE Fellow and an NTT R&D Fellow.

■ Awards

1. 22nd Sato Paper Award of the Acoustical Society of Japan (ASJ) (1982)
2. 27th Sato Paper Award of the ASJ (1987)
3. ATR R&D Award (1991)
4. ATR R&D Award Special Prize (1992)
5. IEEE Signal Processing Society 1993 Senior Award (1994)
6. Best Presentation Award of The Japanese Society for AI (2002)

■ Major academic society functions

1. Associate Editor of the IEEE Transactions on Signal Processing (1994-1997)
2. Chair (1999-2000) of Technical Committee on Neural Networks for Signal Processing of the IEEE Signal Processing Society (IEEE-SPS)
3. Vice-Chair (1997-1998) of Tokyo Chapter of the IEEE-SPS
4. Action Editor of Neural Networks (2000-)
5. Associate Editor of the Transactions of IEICE D-II (1997-2000)
6. Senior Editor of EURASIP Journal on Applied Signal Processing (2001-2002)
7. Member of IEEE Neural Networks Society Administration Board (1998-2002)
8. Member of Technical Committee on Multimedia Signal Processing of the IEEE-SPS (1998-2001)
9. IEEE-SPS Board of Governors, Member-at-Large (2003-2005)
10. Member of IEEE Frank Rosenblatt Award Committee (2005)
11. Chair of Kansai Section, Acoustical Society of Japan (2005)