R&D Spirits

Achieving More Natural and Enjoyable Speech Communications

Akitoshi Kataoka
Group Leader
Acoustic Information Group
Media Processing Project
NTT Cyber Space Laboratories

As broadband transmission systems continue to progress, the trend in speech and audio signal processing is shifting from efficient spectrum usage to high-quality communications. In parallel with this trend, the Acoustic Information Group at NTT Cyber Space Laboratories is focusing its R&D efforts on next-generation speech communications. With a history originating with the telephone, what does the future hold for speech communications? We put our questions to group leader Akitoshi Kataoka, a researcher who has been deeply involved in the international standardization of the CS-ACELP speech coder.

Research of Sound: Where Human Senses and Psychology Play an Important Role

—Dr. Kataoka, please give us an overview of current research efforts in this field.

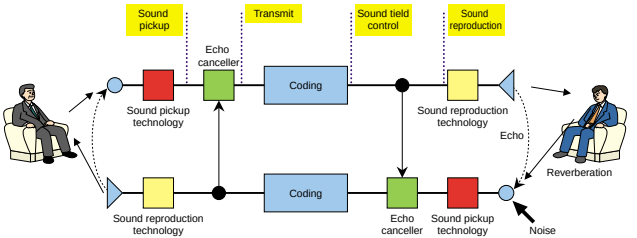
In our group, our research of speech and audio signal processing covers a wide range of technologies from sound pickup and transmission to control methods and sound reproduction. Some key examples are noise and reverberation suppression and spot pickup in the area of sound pickup; speech and audio compression in transmission; and sound image and spot sound reproduction in the area of sound reproduction. We are also researching echo cancellation and other advanced functions (Fig. 1). I myself researched speech-coding systems before becoming group leader of the Acoustic Information Group last year.

—What are the group's research objectives?

When people hear about research involving sound, the most common thing that comes to mind is audio-related technologies. As part of NTT Laboratories, however, we are aiming our research at bidirectional communications as in videoconferencing and videophones, and we are more concerned with sound of a practical nature than sound simply for appreciation. In fact, videoconferencing and other advanced communications systems are already coming into use, and the aim of our research is to achieve "speech communications" that can be used in a more natural manner with a greater sense of presence. Specifically, we aim to achieve a natural and enjoyable sound image by picking up all neighboring sounds and playing them back in stereo to make it feel as if the person you are talking to is right in front of you. Here, we want the user to be able to talk freely without having to worry about speaking into a receiver as in today's telephones. As a service, this will also involve video, and we have therefore begun collaborating with groups researching video.

—In what way is this research special?

This research must take human characteristics into account. In the end, it is people who are sensitive to sound, and it is people who will evaluate new sound technologies. In telephone communications, for example, the prime requirement is to suppress noise so that the other person's voice can be clearly heard. If, however, that means transforming noise into a



Sound pickup technology: noise, reverberation suppression; spot pickup
Sound reproduction technology: sound-image; spot sound reproduction
Echo canceller: echo suppression
Coding: compression of speech and audio information

Fig. 1. Speech and audio signal processing technology.

sound unfamiliar to the human ear, the result will be an unpleasant feeling. A case in point is communication by cellular phones. Current cellular systems adopt a scheme that models the human voice. Such a scheme, though, cannot easily model noise from cars and elsewhere, so sounds that cannot be heard in the natural world come to be generated and transmitted. The other party then hears sounds that he or she has never heard before, and these sounds can produce an uncomfortable, unnerving feeling. Against this background, speech-coding research, which was originally obsessed with reducing the bit transfer rate by efficient spectrum usage, is now attaching more importance to quality by raising the bit transfer rate as advances in wireless technologies make more bandwidth available.

—From a technology perspective, what makes this research stand out?

As I mentioned before, our research covers all aspects of speech and audio signal processing in communications. Here, if even one thing in the process performs poorly, then everything else will suffer, no matter how good they may technically be. In short, it is important to achieve a "total balance." Of course, there is superior technology for each and every aspect of communications, but our research is not really about making epoch-making breakthroughs through any one kind of amazing technology. When researching speech and audio processing from a "total" perspective in the way that we do, there is really no other way to go about it. Perhaps this in itself is a feature of our research approach, which I often compare to baseball. It certainly would be nice to be able to hit home runs all the time, but in reality, you often end up with nothing but strikeouts and no score at all! It would be better to aim for hits and get the occasional home run. This is why I tell young researchers that "even a bunt will get you on base." Of course, going all out for a home run is sometimes good as well.

—What form do you see this research taking in the future?

In terms of specific services, our research will initially be focused on introducing videoconferencing systems to corporate users. The corporate market, though, is limited, and we will eventually focus our research efforts on the home market. To get to that point, however, a variety of issues must first be addressed, and it will take a number of years before actual services can be provided. Honestly speaking, I think it will take about ten years to be able to provide high-reality speech communications for that market.

—Can you give us some examples of these future issues?

Technically speaking, sound pickup performance

has not yet reached a satisfactory level. Making a compact microphone array is also a major issue. Initially, a single microphone was usually used to pick up sound, and we thought that we could process the sound so obtained to create a feeling of presence. However, this also has its limits, and we found that multiple microphones are essential. The human voice has frequency components ranging from 50 Hz to 7 kHz, which means that a microphone array covering this range would have to be as long as 1.7 meters in front of the speaker. This is impractical, so there is a need for new technology that can make the microphone array small enough to sit on top of a television. Another issue concerns multi-channel echo cancellers for high-reality communication. At the current stage, they require an excessive amount of computation to work. Furthermore, in addition to purely technical issues, it is vital that we lower the cost. A system with a total price tag of about 1 million yen (about $9,000) will not find widespread use. Developing technology that can lower the price to a level acceptable to even general households is one issue that we must address in the years to come.

A Comprehensive Research System: NTT's Strong Point

—How did the Media Processing Project begin?

Well, I think it was inevitable given the history of our research activities. As I've pointed out earlier, our research covers a wide range of speech and audio fields, and the feeling was that we should try to make this comprehensive approach an NTT feature. As for myself, I had been researching sound pickup and coding since entering the company, and I was thought to have a broad knowledge of these areas. All in all, I think my superiors made the right decision in making me group leader.

—What specific topics have you studied up to now?

As an undergraduate, I belonged to a research laboratory involved with ultrasonic impulses. Then, on entering graduate school, I thought that I should also learn about electric circuits, and I took up the research of high-frequency inverters. Later, on looking for employment, I chose NTT since it provided the best environment for testing my capabilities as opposed to continuing with the same studies of my student days. Consequently, I was not especially obsessed about what I would be researching at first. My first research topic turned out to be in the audio field, and microphone arrays in particular, which corresponds to sound pickup in our current research. I spent about four years in this area and then turned to speech coding in 1990. This change came about as the digitization of cellular telephones began to progress and as NTT fell behind Motorola in the initial full-rate system. To regain our position here, NTT needed people with a background in signal processing and electric circuits, and I guess I was a natural choice what with my studies of electric circuits in my graduate-school days. It was initially agreed that I would spend four years in speech coding, but my activities here came to include a proposal for an international standard of the 8-kbit/s CS-ACELP (Conjugate-Structure Algebraic Code Excited Linear Prediction) speech coder (Fig. 2). In the end, I needed a total of seven years to complete my work here. I then spent about three and a half years in business-related activities, but returned to NTT Cyber Space Laboratories in October 2000.

—Can you tell us something about international research trends in speech and audio signal processing?

Besides NTT, France Telecom is another research institution that takes a comprehensive approach to research in this area. In the past, catching up with AT&T in this area was one of our main objectives, and they were also one of our competitors. At present, our main competitor in videoconferencing systems is Polycom, Inc. (formerly PictureTel Corporation) in the USA. In terms of individual research areas, NEC has been working on echo cancellers, and more recently, Microsoft has been incorporating echo cancellation in its Windows Messenger function in Windows XP. This is also a competitive area for us.

—Is your group involved in any international joint activities?

We have exchanged information with France Telecom and other institutions at academic forums. In addition, information is sometimes exchanged between researchers on a personal level and guest researchers are occasionally accepted. At present, however, we are not engaged in any formal joint research. For it to be worthwhile, I believe joint research must be mutually beneficial. In other words, joint research is not very meaningful unless both parties can absorb some technology they don't already

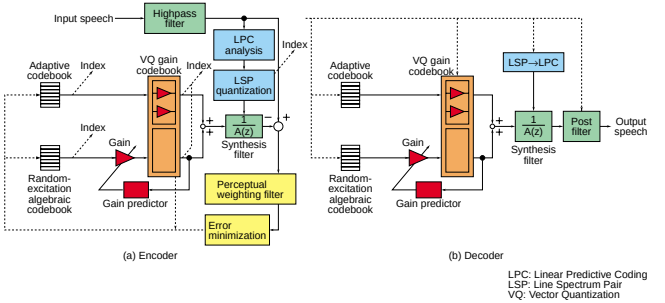


Fig. 2. CS-ACELP speech coder.

have. Of course, NTT hardly intends to come out on top of all other research institutions, but neither does it want to be on the losing end. All in all, I think that discussions and information exchange at academic forums and through standardization activities form the basis of international cooperation.

International Standardization of CS-ACELP: Turning Point of Research Activities

—Can you name a major event in your group's research?

Yes, the adoption of the 8-kbit/s CS-ACELP speech coder that we developed in 1996 as ITU international standard G.729. By chance, our only competitor at that time was France Telecom, but we were able to get the coder into one form by both competing and negotiating at the same time. This coder is being used in almost all of NTT DoCoMo's cellular phones. It is also being used as a basic element of the scalable broadband speech coder that has recently been reported in research papers. The CS-ACELP speech coder has been recognized internationally as effective technology, and as a researcher, I can only say that I'm delighted to see it being used throughout the world.

—Was there anything particularly difficult in these standardization activities?

As I mentioned earlier, it's people who become the final evaluators of sound, and we found that there were significant differences in the perception of sound by people in different countries. It turned out to be extremely difficult to carry on discussions using evaluation data that differed from country to country. In evaluating sound, we used general users and not researchers, but on a 1-to-5 scale, it was rare for Japanese subjects to give even high-quality sound a top score of 5; scores of 2 or 3 were more common. In contrast, American subjects tended to give high scores to sound that Japanese found rather noisy. Of course, these results reflect national characteristics. Initially, when asked why such differences appeared, we replied to the effect that Japanese people are more severe in evaluating sound, and this eventually came to be understood by overseas parties. In the end, it was realized that there was not much that could be done about the low Japanese scores.

—What kind of effect do you think past international activities have had?

As might be expected, their greatest effect has been international standardization. Furthermore, in conjunction with the patented technology included in the CS-ACELP speech coder, we were fortunate to receive the "National Commendation for Invention, Japan Patent Office Commissioner's Award" from the Japan Institute of Invention and Innovation just this

June. His Royal Highness, Prince Hitachinomiya, was present at the award ceremony, and we also received a medal from him. Needless to say, receiving this kind of formal recognition was a very joyous occasion. At the same time, we feel in no way that this marks the end of our work. This technology is still needed in the broadband world, and we feel that there is still room for improvement. An approach more suitable to that world, however, will no doubt be needed. As for myself, I would like to move on to a new research phase while keeping an eye on the growth of this technology that has "left the nest," so to speak.

Passing the Research Baton: One Task for the Future

—Based on your overseas experiences, how do you think NTT is viewed by others?

As a carrier's research laboratory, I believe that international institutions like ITU have a lot of trust in NTT. In recent years, a variety of companies and organizations, including those related to terminal systems, have been joining ITU, and not a few have a profit incentive for doing so. If our only objective were standardization, however, we would be in trouble. As a private enterprise, NTT is concerned with profit, but it also attaches considerable importance to contributing to society. NTT's stance in this regard has been highly evaluated by others and some institutions have offered their support.

—What direction would you like to see this research take in the future?

Actually, I think there are two directions. First, I would like to see research that aims to master the field of sound to create world-recognized value centered on sound-related technologies. Second, considering that total communication performance cannot be improved on the basis of sound only, I think there is also a need for active cooperation with other media including video. Which is better, "maximizing sound" or "integrating sound," I really can't say at present. For the future, moreover, I believe that bringing specialists together as we have done in our group can lead to great things, although I don't know if this would work at the very forefront of research activities. While this is only my personal vision, I believe that developing a holographic telephone*1 would be an excellent goal for the next stage, where research currently in progress spreads to general households. That is my dream.

—Dr. Kataoka, what is life like in NTT Laboratories?

NTT Laboratories is a place where you can explore many possibilities. It is sometimes said that the reason for this is that NTT is blessed with many facilities and research funds. That may be true, but I also think that NTT excels at passing on tradition. Why is it that people like myself can suddenly change to the study of speech coding after research in the audio field and then progress as far as proposing an international standard in only one or two years? It is because there are ample research assets left by people who came before me. When you are creating a program at NTT, for example, there is already a software library, so you don't have to begin from scratch. There are also various people who are more than willing to give advice. In short, successful research means "passing the baton" not only in terms of technology and knowledge but also in terms of research methods, directions, etc. Finally, in addition to being a research institute with a long history, NTT also stands out in the way that it assembles researchers of every generation. This is because research tradition cannot be instantly passed from a veteran to a newcomer—it takes time. Perhaps my work in the coming years will be to develop effective ways of passing this research baton on to those who come after me.

*1 A holographic telephone would enable you to talk with someone as if that person were right in front of you. It would create a life-size holographic image of the called party so that you could listen and talk to that person while watching his or her facial expression.

Interviewee profile

■ Career highlights

Akitoshi Kataoka received the B.E., M.E., and Ph.D. degrees in electrical engineering from Doshisha University of Kyoto in 1984, 1986, and 1999 respectively. Since joining NTT Laboratories in 1986, he has been engaged in research on noise reduction, acoustic arrays, speech-signal processing, and 8-kbit/s speech coding and wideband coding algorithms for ITU-T standards. He is currently the Manager of the Acoustic Information Group in the Media Processing Laboratory, NTT Cyber Space Laboratories, Tokyo, Japan. Dr. Kataoka is a member of the Acoustical Society of Japan, the Institute of Electronics, Information and Communication Engineers (IEICE), and IEEE.

■ Major awards

1996 Technology Development Award (Acoustical Society of Japan)
TELECOM System Technology Award (Telecommunications Advancement Foundation)
NTT President's Award
2003 The Prize of Commissioner of the Japan Patent Office (Japan Institute of Invention and Innovation)

■ Publications

A. Kataoka and Y. Ichinose, "A microphone-array configuration for AMNOR (Adaptive microphone-array system for noise reduction)," J. Acoust. Soc. Jpn. (E) 11, 6, pp. 317-325, 1990.

A. Kataoka, S. Kurihara, S. Hayashi, and T. Moriya, "Improved CELP-based Coding in a Noisy Environment using a Trained Conjugate Sparse Codebook," IEICE Trans. INF.& SYST., Vol. E79-D, No. 2, pp. 123-129, 1996.

A. Kataoka, T. Moriya, and S. Hayashi, "An 8-kb/s Conjugate Structure CELP (CS-CELP) Speech Coder," IEEE Trans. Speech and Audio Processing, Vol. 4, No. 6, pp. 401-411, 1996.

A. Kataoka, S. Kurihara, and S. Hayashi, "6.4-kbit/s variable-bit-rate extension to the G.729 (CS-ACELP) speech coder," IEICE Trans. INF.&SYST., Vol. E80-D, No. 12, pp. 1183-1189, 1997.