Speech/Acoustic Analysis Technology - Its Application in Support of Public Solutions

KOSHINAKA Takafumi, HOSHUYAMA Osamu, ONISHI Yoshifumi, ISOTANI Ryosuke, TANI Masahiro

Abstract

The advent of the age of big data has further raised interest in the need to extract useful information from the huge amount of data that accumulates in the course of our everyday lives. This may be facilitated by high speed and low cost data analysis solutions. These technologies that process the speech/acoustic information that forms the critical component of real-world information are also becoming more important for understanding the context of the analyzed data. They are expected to be employed for public solutions that will support the safety, security, efficiency and equality of society. This paper introduces an innovative technology designed to extract meaningful information from speech/acoustic media and goes on to discuss its application in public solutions.

Keywords

speech recognition, speaker recognition, emotion recognition, noise elimination, acoustic event detection, noise suppressor, beamforming, big data

1. Introduction

Technologies for the speech processing of human voices and the acoustic processing that handles general sounds are promising media processing technologies that together with image/video processing are expected to contribute to the solution of various social issues. The age of big data is now here and cloud computing enables the fast, low-priced processing of a huge amount of multimedia information. The speech and acoustic processing technologies are attracting much attention as a means of data analysis for obtaining valuable information from the large-scale data production of the real world.

From the physical viewpoint, both speech and acoustic information are just waves that are transmitted as vibrations of a medium such as air. For humans, however, these waves contain information that is significant to them. For example, the variety of information revealed in speech includes not only the words that are expressed in various languages but also the sex, age and emotions of the speaker. Even general sounds other than speech, such as the tweets of birds and the wind blowing through a wood, may reveal what is present and/or happening at a location. In the world of computers, the information as described above can be extracted by analyzing the data that is transmitted as physical waves.

In this paper we introduce the latest speech and acoustic processing technologies from the viewpoint of data analysis. These tools are capable of extracting significant information including speech and sound media. We also review proposals for the application of these technologies in public solutions.

2. Speech Analysis

2.1 Text-based Analysis

The human voice, or speech, transmits various information, and the most abundant and informative content is in the “words” that can be expressed as text. The technological development of speech recognition for extracting text from speech was started as early as the 1950’s[1]. The common understanding in the academic world is that the current technical level of speech recognition is still far from that of the human listening ability. In the industrial sphere, however, speech recognition has recently been spreading rapidly as a user interface of smartphones, thereby making this technology very familiar to the general consumer.

At NEC, we have been advancing the R&D of speech rec-

Fig. 1 Speech recognition: From human-vs-machine to human-vs-machine.

ognition from a very early stage. With the vision of “from human-versus-machine to human-versus-human,” the targets of recent R&D and commercialization ventures have shifted from the development of a means of enabling inputting words or certain phrases into a computer. The focus now is on the development of computers that can understand conversations between humans (**Fig. 1**).

A typical development is the VoiceGraphy minute compilation solution[2]. This solution transcribes the conversations between humans into text. Conversations held in a governmental assembly or at a conference for governmental agencies or enterprises can be automatically converted into text in order to reduce significantly the time and cost required for minute compilations. Our technology is now put to practical use in situations previously considered to be unsuitable for speech recognition such as in court interrogations[2].

2.2 Non-text-based Analysis

Human speech contains various information other than the “words” that can be expressed in text. This section deals with the speaker and emotion recognition technologies that have recently entered the practical stage.

Speaker recognition is a technology for identifying an individual from his/her speech. This new and handy means of speech-based biometrics following on the fingerprints, face and veins industrial applications is expected to continue to progress in the future. The speaker recognition can be applied over a wide range, and the fields in which it is regarded to be most promising are the solutions related to safety and security in society. For example in criminal investigations for the identification of a kidnapper from a ransom request call or for the surveillance of crimes and accidents at public locations. The applications in the consumer fields are also expected for use in the identification of bank transactions via phone (telephone banking) or for improved response quality, by identifying a client at a call center (**Fig. 2**).

We have long been challenging the R&D and commercialization of this technology, and we have eventually succeeded in developing a large-scale database search system based on the high-speed and high-accuracy speech matching system that has been developed over long years of speech research. The system is being put to practical use in governmental agencies.

Next, we describe the development of emotion recognition technology and its applications in call centers. For enterprises, the call centers are important contact points for collecting raw voice of customers. As many major cities and public institutions have also established call centers, they are also attracting attention as suitable points for studying the voices of citizens.

We are conducting R&D into the analysis of conversations that is capable of identifying the complaints and requests that contained in the conversations made at call centers. The complaint call detection system (**Fig. 3**) extracts complaint calls by recognizing the anger emotions of a customer, as well as the apologetic expressions of an operator. Since the operator uses apologetic expressions like “I am sorry but could you tell us your address?” in the formal context, the system detects only the apologies spoken with sincere emotion and not such superficially used words. This makes it possible to identify those parts of conversations that refer to the causes of the complaints. These are often concentrated where customer anger and operator apologies appear together.

By employing our studies accumulated from call center operations, we have also developed a technology for judging whether or not the customer who made a call to the call center has knowledge of the product. This technology can be applied to the analysis of customer classifications or to a method of

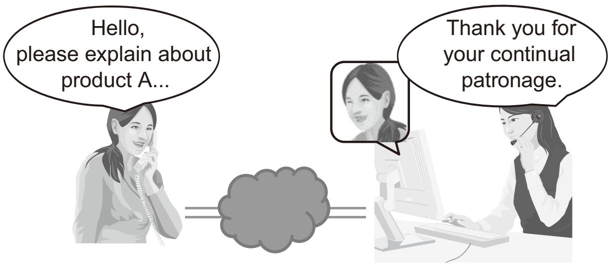
Hello, please explain about product A...
Thank you for your continual patronage.

Fig. 2 Recognition technology of a speaking individual from speech.

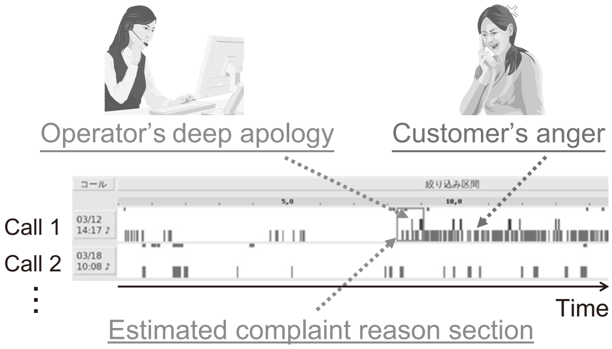

Fig. 3 Complaint call detection system.

how to respond suitably to the caller's knowledge level. As the targets of the emotion recognition research have been expanding recently, it is anticipated that the technology will be capable in the future of employing speech analysis to determine the personality, physical build or health anomalies of a targeted individual.

2.3 Noise Elimination

In both the text-based and non-text-based analyses, any noise superimposed on speech poses a problem for extracting information from such speech. Erroneous recognitions occur frequently during voice input to a smartphone in a noisy environment.

NEC has long been investigating this problem and has commercialized the VoiceDo speech recognition system, which thanks to the noise cancellation technology that uses two microphones can be used in noisy environments such as in factories or warehouses. This technology has been advanced even further to a new technology that combines the "two-microphone noise cancellation" and "model-based speech enhancement" technologies (**Fig. 4**).

The new technology applies single-microphone speech enhancement[5] using the speech models (knowledge) assumable in quiet environments, in addition to the latest 2-microphone noise cancellation system. This procedure aims to reduce speech distortion due to noise elimination, and to enable recognition of speech that is spoken at a distance from the microphone, even in more noisy environments. This technology has already been applied to the voice input of car navigation systems and has made possible speech recognition in environments with five times higher noise than in the previous environments. R&D is currently advancing aimed at applications for various text-based and non-text-based analyses solutions, and in various environments including in the public areas.

3. Acoustic Analysis

3.1 Acoustic Event Detection

Various sounds other than speech exist in the real world. Attempts have begun to analyze general sound backgrounds by computer and to identify "when," "where" and "what" is happening.

We have also been carrying out R&D of acoustic analysis for extracting useful information from general background sounds and, as part of these efforts, we have developed the acoustic event detection technology that detects sounds associated with specific events. This technology is capable of detecting and notifying abnormal sounds such as screams or glass-breaking sounds in real time from the environmental sounds captured by microphones installed for example in a public area. It can thereby contribute to the early discovery and early solution of a crime or an accident (**Fig. 5**). The system based on the developed technology not only learns the target sounds to be detected but also learns sounds that would hinder the detection, so that it can detect target sounds accurately from various sounds in the real world by identifying the differences between the target sounds and other sounds. Its effectiveness has been proven after several months of experimental demonstrations held in public areas in which various sounds were mixed.

3.2 Remote Sound Extraction

The validity of acoustic event detection is dependent on the clarity of the input sounds. When the target sounds are general sounds, they are not always generated near the microphone, unlike the case of speech detection, and they often suffer from interference from external disturbances due to ambient noise and reverberation. In order to improve detection performance

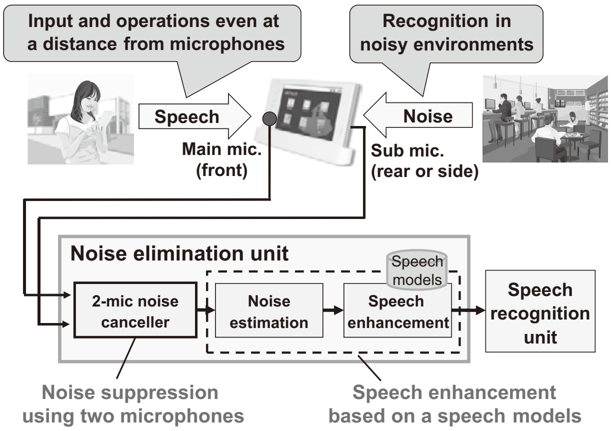
Input and operations even at a distance from microphones
Recognition in noisy environments
Speech
Noise
Main mic. (front)
Sub mic. (rear or side)
Noise elimination unit
Speech models
2-mic noise canceller
Noise estimation
Speech enhancement
Speech recognition unit
Noise suppression using two microphones
Speech enhancement based on a speech models

Fig. 4 Noise elimination technology for speech recognition.

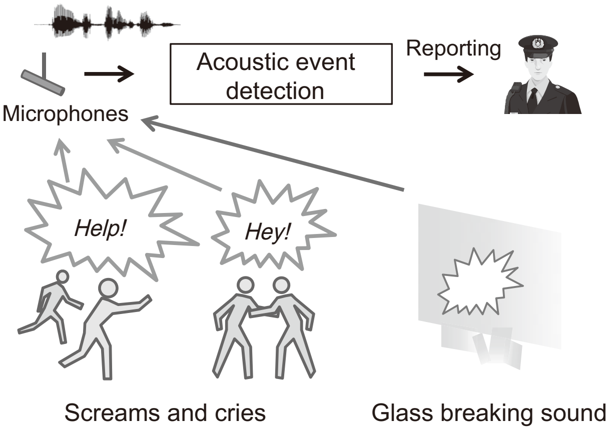

Fig. 5 Application scenario of acoustic event detection.

by eliminating external disturbances, we are currently conducting R&D into noise suppression and beamforming.

As the name indicates the noise suppressor is a technology for suppressing noise. It distinguishes the differences in the properties of the target sounds and background noise and extracts the target sound by suppressing the noise exclusively. Significantly, the suppression of ambient noise in phone calls via mobile phones has already been achieved and has recently even succeeded in suppressing wind noise that has previously been difficult to eliminate[3].

Beamforming is a technique for isolating and extracting only the sounds arriving from a specific direction by using several microphones[4]. It not only acquires clear sounds but also contributes to audio analyses as useful information indicating the direction from which the target sounds are input. For example, if a glass-breaking sound is observed in the street, it may potentially indicate a house-breaking if it originates from the sidewalk and the potential of a vehicle accident or break-in if it originates from the street.

In the solutions related to safety and security as described above, the acoustic analysis often presents synergistic effects when it is combined with video analysis. Consequently, multimodal analysis combining various media is currently under study.

4. Conclusion

In the above, we overviewed the speech and acoustic analysis technologies and their applications in the domain of public solutions. We expect that the speech and acoustic analysis technologies will be applied more and more in public solutions in the future and will contribute thereby to the solution of social issues over a wide range.

Reference

1) S. Furui, "Toward the 4th Generation Automatic Speech Recognition Technology," The Journal of the Institute of Electronics, Information and Communication Engineers, Vol. 95, No. 5, pp. 422-426, May. 2012.

2) Paul Wang, et al., "e-Evidence, Information Governance, user authorization, Inter-Agency Collaboration, device integrity," NEC Technical Journal, Vol.9, No.1, 2014.12

3) M. Kato, & A. K. Sugiyama: A wind-noise suppressor based on wind - onset detection and spectral gain modification, Int'l Workshop on Acoustic Signal Enhancement, Sep., 2014.

4) O. Hoshuyama, A. K. Sugiyama: Robust adaptive beamforming, Book chapter of "Microphone Arrays," Editors: M. Brandstein & D. Ward, Springer, 2001.

Authors' Profiles

KOSHINAKA Takafumi
Doctor of Engineering
Senior Principal Researcher
Information and Media Processing Laboratories

HOSHUYAMA Osamu
Doctor of Engineering
Principal Researcher
Information and Media Processing Laboratories

ONISHI Yoshifumi
Doctor of Science
Principal Researcher
Information and Media Processing Laboratories

ISOTANI Ryosuke
Principal Researcher
Information and Media Processing Laboratories

TANI Masahiro
Assistant Manager
NEC Laboratories Singapore
NEC Asia Pacific Pte. Ltd.

Information about the NEC Technical Journal

Thank you for reading the paper.
If you are interested in the NEC Technical Journal, you can also read other papers on our website.

Link to NEC Technical Journal website

Japanese

English

Vol.9 No.1 Special Issue on Solutions for Society - Creating a Safer and More Secure Society

Remarks for Special Issue on Solutions for Society - Creating a Safer and More Secure Society
NEC's Vision for Public Solutions
NEC's Public Safety Initiative

For a life of efficiency and equality
New Services Realized with the "My Number" System
"NEC Stadium Solutions" Played a Critical Role in Construction of the World Cup
Deployment of Eye-Catching, Visually Appealing Flight Information Systems
NEC SDN Solutions Accelerate New Service Implementations for Railway Stations
Cloud-Based Interpreting Service Using a Videoconference Telephone Compatible with Multiple Devices
Easy-to-Use, Smartphone-Oriented Internet Banking, featuring Color Universal Design
The World's Best Face Recognition System to Achieve Safety and Security in Our Society
Product Line-up for Face Recognition Solutions and its Social Applications

For a safer and more secure life
Healthcare challenge with ICT (Information and Communication Technologies)
Information Governance
Safety Awareness Network
Building a Safer City in Singapore
Securing the Future in Tigre
New Congestion Estimation System Based On the "Crowd Behavior Analysis Technology"
Speech/Acoustic Analysis Technology - Its Application in Support of Public Solutions
High-Sensitivity Camera for Round-the-Clock Surveillance
Imaging Solutions for Search & Rescue Operations
Emergency Mobile Radio Network based on Software-Defined Radio

For the security and safety of critical infrastructure
Centralized Information Control System Supporting Safe and Stable Shinkansen Transportation
Smart Water Management Technology with Intelligent Sensing and ICT for the Integrated Water Systems
A Water Leak Detection Service Based on Sensors and ICT Solutions
Harbor Monitoring Network System for Detecting Suspicious Objects Approaching Critical Facilities in Coastal Areas
Failure Sign Monitoring System for Large-scale Plants Applying System Invariant Analysis Technology (SIAT)
Infrared Camera Image Processing Technology and Examples of Applications
Cyber Security Factory - Our Commitment to Help Developing More Effective Methods of Coping with Today's Increasingly Sophisticated Cyber Threat

Advanced technologies for a Safer and More Secure Society
Technologies for Improving the Speed and Accuracy of Fingerprint Identification Systems in Support of Public Bodies
Compression Technologies Supporting Next Generation Broadcasting Services - Ultra-HD Digital Video Compression Technology and Real Time HEVC Compression Unit Corresponding to 4K HD Images

NEC
NEC Technical Journal
Special Issue on Solutions for Society - Creating a Safer and More Secure Society

Vol.9 No.1

January, 2015

Special Issue TOP

NEC Information

NEWS
NEC Starts Operation of Satellite Integration Center
Development of Water Purification System Type2 Reverse Osmosis (WPS RO2) for Japan Ground Self-Defense Force