Safety, Security, and Convenience: The Benefits of Voice Recognition Technology

KOSHINAKA Takafumi, LEE Kong Aik

Abstract

We use our voices all the time to communicate with one another. Talking is the simplest and easiest means of information transmission. It could also be one of the simplest means of identification. Recent advances in computing technology have made voice recognition — a biometric technology based on the unique characteristics specific to an individual's voice — more convenient, safer, and more secure than ever. In this paper, we review the current state of voice recognition technology and show how deep learning — the core of contemporary AI technology — is providing the key to unlock the power of biometrics. We will also look in some detail at NEC's work in the field of voice recognition technology, which is at the forefront of worldwide efforts to make this technology accessible and reliable. Finally, we discuss potential industrial applications for voice recognition technology such as public safety solutions.

Keywords

speaker verification, speaker identification, speaker recognition, deep learning, speech recognition

1. Introduction

Communication is an integral part of our daily lives and we use many different means to communicate with one another. However, none is more important than the human voice. Speaking and listening are fundamental, the basis for all other forms of communication. To speak and listen, you don't need an electronic device; you don't even need paper and pen. All you need is your voice. No other means of communication is simpler or easier.

The voice — which is the medium for speaking and listening communication — is a type of human biometrics. Because each person's voice has characteristics that are unique and peculiar to that voice alone, the voice can be used for biometric identification. Voice recognition offers an easy and simple means of individual identification for users. Moreover, this type of authentications requires no special equipment; conventional microphones and telephones can be used, and no expensive, special sensor device is required. Setting up a voice recognition system is easy and relatively inexpensive.

What is voice recognition technology, how does it work, and what is its connection to deep learning — one of today's core AI technologies? These are some of the questions we will try to answer in this paper. We will also take a look at the work NEC has been doing in this area, examine the world-class voice recognition technology the company has developed, and consider potential industrial applications for voice recognition technology such as public safety solutions.

2. Voice Recognition Technology

Everyone is used to guessing whose voice they are listening to even when they cannot see the speaker. As long as you know the person, you're pretty likely to guess correctly. The unique characteristics of each individual's voice are dictated by various physical features such as the shapes of the vocal chords and oral cavity (physical characteristics), as well as by speech habits particular to each of us (behavioral characteristics). Voice recognition technology identifies the speaker by extracting and analyzing the features that relate to these individual physical and behavioral characteristics.

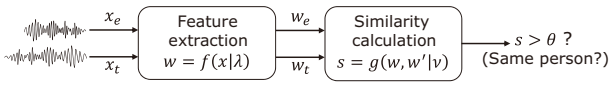

Fig. 1 Basic configuration of voice recognition system (one-to-one comparison): Models (λ, v) to extract features and calculate similarity are determined by data through learning.

2.1 Technological components of voice recognition

Technically speaking, voice recognition is called speaker recognition or speaker verification. In many cases, it refers to a technology that uses one-to-one processing to compare two voices to determine if they are the same person. Speaker identification, on the other hand, which seeks to identify an unknown individual by their voice, performs one-to-many processing. But even this ultimately boils down to multiple repetitions of one-to-one comparisons. Thus, the basic unit of processing is one-to-one processing, as shown in **Fig. 1**.

Today's most popular framework for feature extraction is a framework called i-vector[1]. Using a standard model of phonemes comprised of many speakers' voices (various vowels and consonants), the i-vector extracts the differences between the standard model and input voice as a feature. However, if all the differences are extracted, the feature will be enormous, with potentially hundreds of thousands of dimensions. To avoid this problem, i-vector compresses such an enormous feature to around 400 dimensions using factor analysis. To calculate similarity, a model called probabilistic linear discriminant analysis (PLDA)[2] is often used. The PLDA stochastically reformulates equations using linear discriminant analysis (LDA) — a traditional method for machine learning — and automatically selects the feature best suited for identification of the speaker based on the 400-dimension feature of the i-vector. Once the data has been analyzed, the similarity is calculated as a likelihood ratio.

Both i-vector and PLDA are formulated using probabilistic models based on the assumption of Gaussian distribution (normal distribution). The de facto standards for voice recognition, i-vector and PLDA incorporate various machine learning techniques. Capable of automated learning, they can generate optimal model parameters from a large amount of data.

2.2 Incorporation of deep learning

Recently, researchers in the fields of image and speech recognition have sought to improve accuracy by applying deep learning. Voice recognition is no exception to this trend; research into deep learning got underway in 2014, and a paradigm shift is now taking place in this field, a shift that promises to bring voice recognition into the mainstream.

This shift is marked by the emergence of a system called deep speaker embedding, or x-vector, which exponentially increases the accuracy of voice recognition. Researchers in the field have eagerly seized on x-vector as a new feature extractor with the potential to replace the conventional i-vector system[3]. **Fig. 2** shows the concept of the x-vector system. First, a deep neural network (DNN) composed of a feature extractor and discriminator is trained to correctly deduce speakers from their voices. The feature extractor of the DNN has been designed in such a way that it pulls only the information suitable for speaker identification from their voices.

Because speech is time-series data with variable length, the amount of data input to the neural network is also variable. This very fact makes it more difficult to handle voices than images. However, it is possible with the x-vector to output a feature in a fixed number of dimensions by inserting a pooling layer — which aggregates the data in a temporal direction — in the end of the feature extractor.

Trials in introducing deep learning to voice recognition do not stop at feature extraction and range widely from the front-end (speech/non-speech recognition and speech enhancement under noisy conditions) to the back-end (similarity calculation). An end-to-end system has also emerged that performs learning of the entire system by replacing all the technical components with the neural network[4]. This trend is likely to continue in the future.

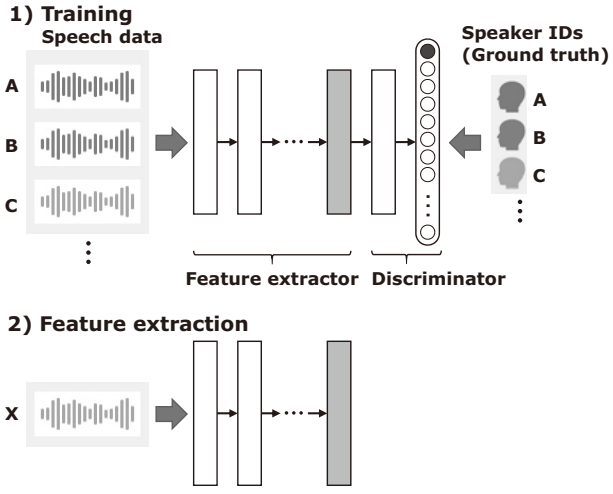

Fig. 2 Concept for feature extraction based on deep learning (x-vector).

2.3 NEC's Work in This Field Shows Promise

At NEC, we see voice recognition technology as one of the leading next-generation biometric modalities, closely following fingerprint and face recognition. Consequently, we have been working hard to develop this technology for practical use and have achieved results that have made us the world's leader in this field.

We were the first to see the potential of deep learning and the first to start research in this promising area. That early research has paid off with the development of powerful unique technologies; these include a sophisticated filter to accurately detect voice activity by distinguishing speech from non-speech in noisy environments[5], a noise reduction system to eliminate noise components from the features of noisy speech[6], and technology to infuse a short-duration utterance with the same quantity of features as can be drawn from long-duration utterance, as this makes it easier to obtain information pertaining to individual characteristics[7].

Besides, NEC has been actively participating in the Speaker Recognition Evaluation (SRE) series — evaluations conducted by the U.S. National Institute of Standards and Technology (NIST)[8]. The SRE series are a competition in which more than 60 teams (in SRE18) from industry-academia-government organizations around the world participate and compete against one another to test the speaker recognition accuracy of their systems using the same data set. We have repeatedly demonstrated our technological superiority in these competitions.

In SRE18, testing was conducted with two tasks: one to find a specific individual from telephone conversations marred by background noise and poor line conditions; and the other to find a specific individual from multiple individuals who appear in video segments on the Internet such as YouTube. Both tasks featured technically severe conditions with a high level of difficulty. In the telephone conversations for example, the degree of accuracy for the baseline presented by the NIST was only 88.8% (11.2% crossover error rate). This does not by any means suggest that the technical level of the NIST's baseline system was low. In fact, this baseline system was the latest state-of-the-art system equipped with the above-mentioned x-vector feature extractor. Taking all this into account, NEC's system achieved accuracy of 95.0% (5.0% crossover error rate) — which was an error rate less than half of what the newest cutting-edge system could achieve.

When you are developing this kind of system, you have to push the quality and performance of every component to the limit. The remarkable improvement in accuracy that we were able to achieve as accomplished by developing an original feature extraction system by adding an auxiliary network called an attention mechanism to the x-vector. This new mechanism automatically selects those parts of the recording where individual voice characteristics are more prominent[9]. We modified the deep learning process as well to enable effective learning without the massive amounts of training data usually required. Instead, we developed a new method for augmenting data by converting limited voice data to multiply the apparent number of speakers several times.

3. Industrial Applications

Finally, let's consider the potential benefits to society of voice recognition technology (**Fig. 3**).

E-commerce: Signatures for small purchases with credit cards are rarely required any more. This lowers the barrier to purchasing for both buyers and sellers by streamlining and speeding up the payment process. Nowadays, convenience is just as important to consumers as security. Using voice recognition meets both these needs. The voice is a simple medium people use for everyday communication, so biometric authentication using voice provides users with a handy and easy means of individual identification. Voice recognition is an identification method ideal for individual identification in

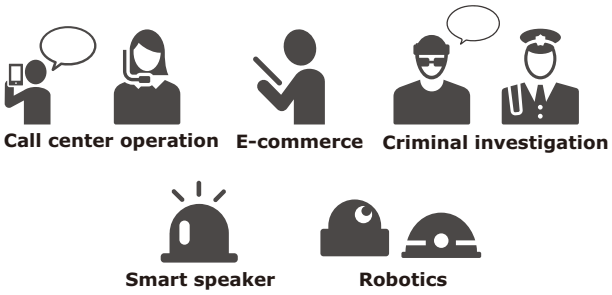

Fig. 3 A wide range of scenarios where voice recognition can play a role.

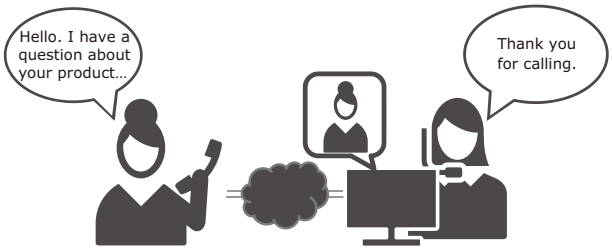

Fig. 4 Call center support: Quick confirmation of customer identification.

commercial transactions such as e-commerce and Internet banking.

Call center operations: As more companies take customer-oriented approaches, they are continually striving to improve their services at contact points with customers such as call centers. Some of the issues that have arisen include simplification of individual identification procedure for important customers who make phone calls frequently (**Fig. 4**) and early identification of problem customers such as chronic claimers. Because voice recognition is the only biometric that can be used on the telephone where participants are not visible to one another, it's ideal for call center operations as it makes it possible to identify customers in the course of a natural conversation.

Criminal investigation: Telephone-based fraud is a sophisticated and constantly evolving criminal enterprise, always adapting to the various measures taken to combat it. Voice recognition may prove helpful in investigating these crimes, providing an analytical tool to support tracking of perpetrator. It can also support surveillance of organized crime on telephone and the Internet. Voice analysis can also be used proactively to suppress crime as it may be capable of picking up information pointing to potential criminal activity — more so even than the surveillance cameras that have become so commonplace on our streets in recent years.

Other: Voice-based individual identification is likely to be the biometric of choice for hearables such as smart speakers and smart earbuds as well as for user-friendly interfaces such as robots[10].

4. Conclusion

Voice recognition is clearly one of the easiest biometrics to implement and use. Now, thanks to the incorporation of deep learning in voice recognition systems, this technology is much more reliable and secure. NEC has established itself as a world leader in this field with superior technology that is setting the standard for accuracy and performance. Ideally suited for a broad range of applications such as e-commerce, call centers, and criminal investigation, voice recognition offers user-friendly convenience and high accuracy. NEC is committed to bringing the benefits of this technology to society and to enhancing and refining that technology.

* YouTube is a trademark or registered trademark of Google LLC.
* All other company names and product names that appear in this paper are trademarks or registered trademarks of their respective companies.

Reference

1) Najim Dehak et al., "Front-End Factor Analysis for Speaker Verification," IEEE Transactions on Audio, Speech, and Language Processing, Vol.19, pp.788-798, May 2011.
2) Simon J. D. Prince et al., "Probabilistic Models for Inference about Identity," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.34, Jan. 2012.
3) David Snyder et al., "X-vectors: Robust DNN Embeddings for Speaker Recognition," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Apr. 2018.
4) Georg Heigold et al., "End-to-end Text-dependent Speaker Verification," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Mar. 2016.
5) Hitoshi Yamamoto et al., "Robust i-vector extraction tightly coupled with voice activity detection using deep neural networks," Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Dec. 2017.
6) Shivangi Mahto et al., "I-vector Transformation Using Novel Discriminative Denoising Autoencoder for Noise-Robust Speaker Recognition," INTERSPEECH, Aug. 2017.
7) Hitoshi Yamamoto et al., "Denoising Autoencoder-Based Speaker Feature Restoration for Utterances of Short Duration," INTERSPEECH, Sep. 2015.
8) Speaker Recognition, National Institute of Standards and Technology (NIST) https://www.nist.gov/itl/iad/mig/speaker-recognition
9) Koji Okabe et al., "Attentive Statistics Pooling for Deep Speaker Embedding," INTERSPEECH, Sep. 2018.
10) T. Arakawa, "Ear Acoustic Authentication Technology: Using Sound to Identify the Distinctive Shape of the Ear Canal", NEC Technical Journal, Vol. 13, No.2, Apr. 2019.

Authors' Profiles

KOSHINAKA Takafumi
Ph.D.
Senior Principal Researcher
Biometrics Research Laboratories

LEE Kong Aik
Ph.D.
Senior Principal Researcher
Biometrics Research Laboratories

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.13 No.2 Social Value Creation Using Biometrics

Vol.13 No.2
April 2019

Special Issue TOP

Remarks for Special Issue on Social Value Creation Using Biometrics
Committed to Supporting Social Values via Biometrics

Papers for Special Issue

Commitment to Biometrics NEC Is Promoting
Bio-IDiom — NEC's Biometric Authentication Brand
The Future Evolution and Development of Biometrics Studies
Privacy Measures of Biometrics Businesses

Services and Solutions That Leverage Biometrics
The Western Identification Network: Identification as a Service in a Federated Architecture
Use of Face Authentication Systems Associated with the "My Number Card"
Face Recognition Cloud Service "NeoFace Cloud"
NEC Enhanced Video Analytics Provides Advanced Solutions for Video Analytics
New In-Store Biometric Solutions Are Shaping the Future of Retail Services
ID Service Providing Instantaneous Availability of User's Desired Financial Services
Biometrics-Based Approach to Improve Experience from Non-routine Lifestyle Fields
Construction Site Personnel Entrance/Exit Management Service Based on Face Recognition and Location Info
The Importance of Personal Identification in the Fields of Next-Generation Fabrication (Monozukuri)

Core Technologies and Advanced Technologies to Support Biometrics
How Face Recognition Technology and Person Re-identification Technology Can Help Make Our World Safer and More Secure
Advanced Iris Recognition Using Fusion Techniques
Advanced New Technology Uses New Feature Amount to Improve Accuracy of Latent Fingerprint Matching
Safety, Security, and Convenience: The Benefits of Voice Recognition Technology
Ear Acoustic Authentication Technology: Using Sound to Identify the Distinctive Shape of the Ear Canal
Automatic Classification of Behavior Patterns for High-Precision Detection of Suspicious Individuals in Video Images
Facial-Video-Based Drowsiness Estimation Technology for Operation on Low-End IoT Devices

NEC Information

NEWS
2018 C&C Prize Ceremony