JAIST Repository
https://dspace.jaist.ac.jp/

Title	マッチング追跡と音声指紋のスパースコーディングを利用した聴覚表現
Author(s)	TRAN, KIM DUNG
Citation	
Issue Date	2022-09
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/18141
Rights	
Description	Supervisor:鵜木 祐史, 先端科学技術研究科, 博士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Abstract

Speech is one of the natural ways human beings communicate; however, natural speech is impossible for long-distance communication. Nowadays, global real-time communication over the internet has become a vital part of modern civilization. The internet is not only a telecommunication technology, it has become a world. Consequently, the risks related to speech in the virtual world become more complicated than just wiring or eavesdropping telephone lines. Actions should be taken to ensure that we do not become victims of this world. We need to deal with security issues when speech is used as a tool to control automated systems; spoofing and privacy when speech is used in communication; and properties protection and management when speech is used as commercial products.

One way to deal with these issues is using speech fingerprint. Speech signals are believed to convey unique features that can be used as a biometric security measure along with iris, fingerprint, and facial recognitions. The purpose of a speech fingerprint technique is extracting distinguishable features related to speaker individuality and linguistic content from speech signals and combining the features to create unique speech fingerprints. The speech fingerprints can then be used for speaker verification in security, anti-spoofing in communication, and properties protection of commercial products.

Current speech fingerprint techniques produce speech fingerprints in three basic steps. In step one, speech signals in time domain are projected into a time-frequency domain. In step two, patterns analysis methods are used to obtain important spatiotemporal features. In step three, hashing methods are used to combine the obtained features to create speech fingerprints. Step one is a challenging issue in speech coding. Various speech coding techniques and speech representation models have been proposed such as spectrograms using Fourier and wavelet transforms, auditory spectrograms using auditory filterbanks, spikegrams using matching pursuit algorithms, and auditory sparse representations using perceptual matching pursuit algorithms. The purpose of step one is emphasizing important acoustical features on a representation model. Step two is another challenging issue in pattern recognition. Speech signals are the natural carrier of information about speaker individuality, language, emotion, and so on. Obtaining unique and distinguishable features for speech fingerprints is the purpose of step two. Step three is also a challenging issue. Reducing the dimensions of features causes loss of information and thus, speech fingerprints become less distinguishable. Keeping high dimensions of features causes problems in storage, transferring, and searching. The purpose of step three is producing speech fingerprints that are accurate in matching and convenient to use.

Although the current speech fingerprint techniques in the literature can achieve high performance in various application. However, there is critical drawback is that they are driven by practical results; thus, actual speech fingerprints are either lesser important or not the main focus in their applications. The present study assumes that speech fingerprints are highly related to speaker individuality, and they are a part of the neural activity patterns of the auditory nerves. Based on this assumption, the purpose of the present study is extracting biologically accurate speech fingerprints. In pursuing this purpose, the first task aims to mimic the neural activity patterns to obtain speech fingerprints. Then in the second task, the uniqueness of the proposed speech fingerprints is verified. Finally in the third task, a speech fingerprint identification method is used to apply speech fingerprints in practical applications.

Emphasizing significant features of a speech signal in a signal representation model, e.g., spectrogram, spikegram, and auditory representation, is an essential task of a speech fingerprint technique. Previous studies have revealed that by mimicking the neural activity patterns (NAP) of the auditory periphery

to obtain perceptual features of speech signals, the resultant auditory representation is beneficial to speech-coding and pattern-analysis applications in comparison with spectrogram and spikegram representations. This study proposes to use auditory representations in the process of creating speech fingerprints.

Many efforts have been spent on applying psychoacoustics to concentrate perceptual features on auditory representations to mimic the neural activity patterns generated by the auditory periphery to reproduce the amazing abilities of our hearing system. However, several limitations—using the Bark scale and gammatone basis—remain in the methods used for creating auditory representations. This study found that by mimicking: (1) the sparseness of NAP with a sparse coding technique, i.e., a matching pursuit algorithm, (2) the characteristic frequency of basilar membrane motion with an equivalent rectangular bandwidth scale, (3) the impulse response measured at the basilar membrane with a gammachirp function, and (4) auditory masking with a masking model, perceptual features in auditory representations could achieve similar perceptual evaluation scores, e.g., PEMP-Q and PESQ, while requiring the lowest number of non-zero elements in comparison with features in spectrograms and spikegrams.

Our hearing system has the ability to identify who is speaking, understand spoken language, recognize emotions, etc. simultaneously in very noisy conditions. This miracle is still a mystery to science. Contemporary knowledge divides our hearing system into the auditory periphery and the central auditory cortex. The auditory periphery is responsible for converting speech as sound pressures into NAPs at the auditory nerve and the central auditory cortex is responsible for cognitive functions. At the present time, due to the lack of equipment to obtain the real NAPs at the auditory nerve. Therefore, current speech analysis techniques can only be evaluated by using perceptual evaluation scores and pattern analysis methods. Because of these reasons, in the second task, the present study hypothesizes that there must be unique patterns that help the central auditory cortex identify who is speaking. Therefore, a landmark-based pattern analysis technique is used to combine the features on auditory representations. This technique is used to create a graph-like structure of perceptual features to mimic the neural activity patterns. Then, a uint32 function is used to convert the perceptual structures into hash sequences for fast indexing. Experimental results show that the perceptual structures of auditory representations are the most effective in identifying speakers.

In the last task, a deep hashing technique is used as a speech fingerprint identification algorithm. At first, the proposed speech fingerprint method is used to extract speech fingerprints from speech signals. Then, the extracted speech fingerprints are used as input features of a supervised deep learning algorithm. Then, the deep learning algorithm converted the speech fingerprints into binary hash sequences in the Hamming space. Finally, speaker identification and retrieval experiments are conducted to evaluate the effectiveness of the speech fingerprints and the identification algorithm. Experimental results show that the proposed method can achieve very high identification performance that is competitive to other contemporary state-of-the-art methods.

Keywords: auditory filterbank, gammatone/gammachirp, masking effect, perceptual features, spikegram