Int. j. inf. tecnol.
https://doi.org/10.1007/s41870-024-01877-z

ORIGINAL RESEARCH

A speaker identification-verification approach for noise-corrupted and improved speech using fusion features and a convolutional neural network

Rohun Nisa[1] · Asifa Mehraj Baba[1]

Received: 6 January 2024 / Accepted: 9 April 2024

Abstract The degraded quality of the speech input signal has a negative impact on speaker recognition techniques. We address the issues of speaker recognition from noise-corrupted audio signals in the presence of four noise variants, including factory noise, car noise, street traffic noise, and voice babble noise, as well as noise-suppressed enhanced speech. The goal of this research is to create a speaker recognition algorithm that is resistant to a diverse spectrum of speech capture quality, background scenarios, and interferences. In this work, three distinct features, including Mel Frequency Cepstral Coefficients (MFCC), Normalized Pitch Frequency (NPF), and Normalized Phase Cepstral Coefficients (NPCC) are combined. The analysis that MFCC, NPF, and NPCC illustrate distinct features of speech underlies our observation. A Convolutional Neural Network (CNN) is used in our speaker recognition strategy to learn speaker-dependent attributes from fragments of Mel features, normalized pitch features, and phase cepstral features of clean speech, corrupted speech, and enhanced speech. The performance is measured using the ITU-T test signals and compared to previous algorithms at different Signal-to-Noise-Ratios of 0 dB, 5 dB, 10 dB, and 15 dB. For enhanced speech, all three features, MFCC, NPF, and NPCC, provided productive speaker identification and verification performance.

Keywords Convolutional neural network · Feature extraction · Feature fusion · Speaker identification · Speaker verification · Speech enhancement

1 Introduction

Speaker recognition, a biometric method, utilizes speech features to authenticate a user's uniqueness through automated analysis of voice signals. Over recent decades, Automatic Speaker Recognition (ASR) systems have advanced significantly, finding applications in forensics, banking, and security. These systems comprise preprocessing, feature extraction, and speaker modeling components. Preprocessing involves refining input signals by eliminating non-speech elements and performing tasks like pre-emphasis and end-point detection [1] [2]. Feature extraction, termed "front end preprocessing," transforms voice signals into numerical characteristics essential for training and testing speaker recognition systems. Speaker modeling constructs methods for speaker feature matching, crucial in the recognition stage for identification or verification purposes. Thus, speaker recognition systems serve vital roles across various domains, ensuring efficient and secure user authentication [3] (Fig. 1).

Speaker recognition systems often struggle in challenging acoustic environments due to factors like low audio SNR, diverse accents, and ambient noise, such as babble noise in crowded places. Conventional methods heavily rely on short-term spectral features like MFCC and Linear Prediction Cepstral Coefficients (LPCC), limiting their effectiveness in the presence of acoustic degradations. To address this, our research proposes a deep learning-based method called 1D-Frame Level-Feature Fusion-CNN. By combining MFCC with normalized pitch and phase features, this approach enhances recognition

✉ Rohun Nisa
rohunnisa@islamicuniversity.edu.in

Asifa Mehraj Baba
asifa.baba@islamicuniversity.edu.in

[1] Department of Electronics and Communication Engineering, Islamic University of Science and Technology, Awantipora, Jammu & Kashmir, India 192122

Published online: 19 May 2024