Chapter 6
Speech-Based Speaker Gender Identification Using Deep Learning

G. Aishwarya Laxmi
Avinashilingam Institute for Home Science and Higher Education for Women, India

Judith Justin
https://orcid.org/0000-0001-6339-0841
Avinashilingam Institute for Home Science and Higher Education for Women, India

R. Vanithamani
https://orcid.org/0009-0006-2767-3629
Avinashilingam Institute for Home Science and Higher Education for Women, India

Pavithra Suchindran
https://orcid.org/0000-0003-3878-8473
Avinashilingam Institute for Home Science and Higher Education for Women, India

ABSTRACT

Speech recognition is a key area in Artificial Intelligence that involves extracting meaningful information from speech signals, such as gender, age, or emotion. While humans can easily identify a speaker's gender through conversation, computers require advanced models to perform this task effectively. This study proposes a deep learning-based approach for gender classification using features extracted from speech samples. Two feature extraction methods: Mel Frequency Cepstral

DOI: 10.4018/979-8-3373-3048-8.ch006

Coefficients (MFCC) and Linear Prediction Cepstral Coefficients (LPCC) were evaluated using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models. MFCC with CNN achieved 89.1% accuracy, and with LSTM, 90.5%. In contrast, LPCC with CNN and LSTM produced lower accuracies of 84% and 85.2%, respectively. Results show that MFCC features significantly improve the performance of both models in classifying speaker gender.

INTRODUCTION

Speech builds an individual connection between human beings; it is the most significant component of the communication system. Speech in humans is produced through a coordinated process that involves several systems. The flow of air, which is necessary for the production of sound, is initiated by the lungs. Phonation is the process by which the vocal cords vibrate when air passes through the larynx, creating voiced sounds. The pharynx, mouth, and nasal cavities are among the resonant regions through which this sound passes, influencing the tone and quality of the voice. The next step is articulation, in which unique speech sounds are produced by the tongue, lips, and other tissues. The brain controls the entire system, maintaining accurate and fluent spoken language expression (Livieris et al., 2019). For the objective to imitate human-like speech in machines, speech processing technologies have advanced to include voice recognition systems, which are inspired by the natural speech generation process. In addition to linguistics and health, an extensive understanding of this mechanism is essential for improving voice-driven AI systems, speech therapy, and communication assistance for those with speech disorders.

The human voice is versatile and varies in range, which makes it capable of producing various sounds due to the vibration of the vocal cords. Each individual has unique linguistic features. The location, shape, and mobility of a person's vocal tract can affect the linguistic characteristics of speech samples. Speech samples may convey information about an individual's age, gender, accent, and emotional aspects (Singhal & Sharma, 2023a). The two main characteristics of speech samples are pitch and tone, which may differ between people. In speech, pitch refers to the highness or lowness of a tone as recognized by the auditory system, which is determined by the vocal cords cyclic production of vibrations per second. Typically, the pitch of human speech ranges between 100 to 400 Hz. The key acoustic component of tone is pitch. Tone mainly refers to dialects where word differentiation is based on pitch (Ertam, 2019). Speech samples play a vital role in speech recognition, as a speaker's distinct voice characteristics are influenced by their vocal traits, such as the vocal cords physical structure and the vocal tract's resonating chambers. It is attainable to employ these traits for security.