2023 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC 2023). Ixtapa, Mexico

Combining Deep Learning with Domain Adaptation and Filtering Techniques for Speech Recognition in Noisy Environments

Emmanuel de J. Velásquez-Martínez[1],
Aldonso Becerra-Sánchez[2],
José I. de la Rosa-Vargas[3],
Efrén González-Ramírez[4],
Armando Rodarte-Rodríguez[5],
Gustavo Zepeda-Valles[6]
Unidad Académica de Ingeniería Eléctrica
Universidad Autónma de Zacatecas
Zacatecas, México
[1]iemmanuelvm@gmail.com, [2]a7donso@uaz.edu.mx,
[3]ismaelrv@ieee.org, [4]gonzalezefren@uaz.edu.mx,
[5]armandorodarte19@gmail.com,
[6]gzepeda@uaz.edu.mx

Nivia I. Escalante-García[7],
J. Ernesto Olvera-González[8]
Laboratorio de Iluminación Artificial
Tecnológico Nacional de México, Campus Pabellón de Arteaga
Aguascalientes, México
[7]aivineg82@gmail.com, [8]e.olvera.Itp@gmail.com

***Abstract*—Speech recognition is a common task in various everyday user systems; however, its effectiveness is limited in noisy environments such as moving vehicles, homes with ambient noise, mobile phones, among others. This work proposes to combine deep learning techniques with domain adaptation and filtering based on Wavelet Transform to eliminate both stationary and non-stationary noise in speech signals in automatic speech recognition (ASR) and speaker identification tasks. It demonstrates how a deep neural network model with domain adaptation, using Optimal Transport, can be trained to mitigate different types of noise. Evaluations were conducted based on Short-Term Objective Intelligibility (STOI) and Perceptual Evaluation of Speech Quality (PESQ). The Wavelet Transform (WT) was applied as a filtering technique to perform a second processing on the speech signal enhanced by the deep neural network, resulting in an average improvement of 20% in STOI and 9% in PESQ compared to the noisy signal. The process was evaluated on a pre-trained ASR system, achieving a general decrease in WER of 14.24%, while an average 99% accuracy in speaker identification. Thus, the proposed approach provides a significant improvement in speech recognition performance by addressing the problem of noisy speech.**

I. Introduction

Speech recognition technology can be divided into two main subareas: speech recognition, which analyzes the content of words, and speaker identification, which identifies the speaker of those words [1]. Currently, speech recognition technology has various applications, such as in voice communication systems, hearing devices, and automatic speech recognition [2], [3], [4]. These technologies have been affected by an external factor, which is noise. Noisy signals are generated by different environments and can be naturally mixed with the speech signal. This causes noise interference to severely decrease the perceptual quality and intelligibility in speech recognition tasks [5]. Studies in this field aim to improve speech in noisy environments by enhancing the quality and intelligibility of speech corrupted by noise [6], [7]. Therefore, achieving automatic speech robustness in real-world noisy scenarios is essential [8]. It is crucial to seek, apply, and/or combine effective techniques to enhance the signal quality, thus improving the accuracy of speech recognition systems in different noisy environments. Currently, deep learning models and filtering techniques are applicable to these scenarios [5], but the current challenge lies in the fact that, given different acoustic environments, the implementation of these models may behave differently compared to experimentation, and unforeseen noises can significantly degrade the quality of the processed signal [9]. This mismatch is generally referred to as a domain mismatch problem. Therefore, an effective solution is required to perform domain adaptation in order to align them with a more accurate transformation formulation from noisy speech to clean speech, making the approach applicable in various acoustic environments [10].

The objective of this work is to increase the noise absorption efficiency for different signal-to-noise ratio (SNR) levels in order to reduce the risk of incorrect speech recognition. This is achieved by proposing the implementation of generative deep learning, domain adaptation and filtering techniques based on Wavelet transform. The obtained results demonstrated improved speech intelligibility and perception. These results were reflected in ASR, with a general decrease in word error rate (WER) of 14.24% and an average increase in speaker identification metrics by 55.0% compared to noisy speech.

The rest of the paper is organized as follows. Section 2

979-8-3503-3688-7/23/$31.00