INVESTIGATING DEEP NEURAL NETWORKS FOR SPEAKER DIARIZATION IN THE DIHARD CHALLENGE

Ivan Himawan, Md Hafizur Rahman, Sridha Sridharan, Clinton Fookes, Ahilan Kanagasundaram

Speech Research Lab, SAIVT, Queensland University of Technology, Australia
{i.himawan, m20.rahman, s.sridharan, c.fookes}@qut.edu.au, ahilan.eng@gmail.com

ABSTRACT

We investigate the use of deep neural networks (DNNs) for the speaker diarization task to improve performance under domain mismatched conditions. Three unsupervised domain adaptation techniques, namely inter-dataset variability compensation (IDVC), domain-invariant covariance normalization (DICN), and domain mismatch modeling (DMM), are applied on DNN based speaker embeddings to compensate for the mismatch in the embedding subspace. We present results conducted on the DIHARD data, which was released for the 2018 diarization challenge. Collected from a diverse set of domains, this data provides very challenging domain mismatched conditions for the diarization task. Our results provide insights into how the performance of our proposed system could be further improved.

Index Terms— DIHARD challenge, speaker diarization, deep neural network, domain adaptation, data augmentation

1. INTRODUCTION

Speaker diarization aims to automatically provide annotation labels to each speaker-homogeneous region in a recording that corresponds to the unique speaker's identity. The ultimate goal for this task is to answer '*Who Spoke When*' in a multi-speaker environment. There are many potential applications that can be benefitted from speaker diarization systems such as ASR, forensics applications, biometric security and information recovery. For example, since ASR typically assume one speaker is speaking in an utterance, having information regarding speakers and their speaking boundaries could potentially improve ASR accuracy [1].

Advances in speaker diarization are influenced by the techniques proposed for speaker verification [2]. The speaker diarization task is used to determine whether or not the segment containing speech in an utterance was spoken by the same person as in the previous segments. This process can be repeated for all spoken segments in the conversation to determine if they were spoken by the same individual. Speaker diarization can be divided into two main tasks: segmentation and clustering. Speaker segmentation is the process of detecting speaker change in an audio stream based on speaker identities, background noise and channel conditions. The aim of this phase is to detect the longest possible speaker-homogeneous segments in the audio stream. The purity of extracted segments directly affects the overall diarization performance, as speaker model comparison and clustering depends on a pure segmentation of the audio stream. A popular segmentation approach is Bayesian Information Criterion (BIC) [3, 4]. In this approach, the marginal probability is calculated between initial segments to form a BIC matrix, which is later used to determine whether speech segments are likely to be modelled using single or double multivariate Gaussian distributions. Recently, HMM/Viterbi segmentation [5, 6] is used to find the speaker changing points in the audio stream using ergodic HMM states, where any state of the HMM can be attained from another state within a single step.

Researchers have discussed the latest speaker diarization problems at the Jelinek Summer Workshop on Speech and Language Technology (JSALT) 2017. It was found that existing state-of-the-art diarization systems performed poorly when recordings had a wide range of acoustic conditions, a large amount of noise, and high levels of speaker overlap. The first DIHARD challenge has been proposed to develop a state-of-the-art diarization system to address this issue [7, 8, 9].

The application of deep learning for speaker clustering (and diarization) is a relatively new area where substantial gains in performance can still be achieved using challenging data [10]. Feature vectors that were extracted from DNN trained to recognize speaker identities, called speaker embeddings, have been shown to be effective in speaker verification tasks [11], and they are also applicable to the speaker diarization problem [12]. It is suggested in [13] that for short utterance conditions, the use of embeddings is superior compared to i-vectors. Since we are dealing with very short speech in speaker diarization, the use of embeddings is expected to improve performance. In contrast, it is still common to use Gaussian Mixture Model (GMM) modeling techniques and MFCC features in mainstream diarization systems. To deal with different acoustic conditions between the training data and DIHARD data, we also explore previously developed domain mismatch compensation techniques for i-vectors, and applied the same techniques to compensate the mismatch in the embedding subspace.