The QUT-NOISE-SRE Protocol for the Evaluation of Noisy Speaker Recognition

David Dean, Ahilan Kanagasundaram, Houman Ghaemmaghami, Md Hafizur Rahman, Sridha Sridharan

Speech and Audio Laboratory, Queensland University of Technology, Brisbane, QLD, Australia

ddean@ieee.org, {a.kanagasundaram, houman.ghaemmaghami, m20.rahman, s.sridharan}@qut.edu.au

Abstract

The QUT-NOISE-SRE protocol is designed to mix the large QUT-NOISE database, consisting of over 10 hours of background noise, collected across 10 unique locations covering 5 common noise scenarios, with commonly used speaker recognition datasets such as Switchboard, Mixer and the speaker recognition evaluation (SRE) datasets provided by NIST. By allowing common, clean, speech corpora to be mixed with a wide variety of noise conditions, environmental reverberant responses, and signal-to-noise ratios, this protocol provides a solid basis for the development, evaluation and benchmarking of robust speaker recognition algorithms, and is freely available to download alongside the QUT-NOISE database. In this work, we use the QUT-NOISE-SRE protocol to evaluate a state-of-the-art PLDA i-vector speaker recognition system, demonstrating the importance of designing voice-activity-detection front-ends specifically for speaker recognition, rather than aiming for perfect coherence with the true speech/non-speech boundaries.

Index Terms: noisy speaker verification, speech databases, evaluation protocols

1. Introduction

While research in the field of speaker recognition has been ongoing for decades, the greatest cause of errors still remains the same: the issue of mismatch caused by intersession variability. The term intersession variability encompasses a number of phenomena contributing to this mismatch, including different recording devices, speech coding, transmisison channels, and variability introduced by the speaker (such as linguistic content). One particular area of intersession variability that has recently received renewed interest is the effect of noise and reverberation in the recording environment.

The recent development of the PRISM evaluation set [1], and the inclusion of background noise in the evaluation of speaker recognition systems in the 2012 National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) [2] provoked a significant increase in the investigation of modern i-vector speaker recognition techniques in the presence of (often simulated) environmental noise [3, 4, 5, 6, 7].

However, one of the main unsolved issues in evaluating speaker recognition algorithms is the lack of a suitable large corpus of noisy speech available covering many speakers in a variety of noisy environments, and with a wide range of noise levels. In order to begin to approach the volume required to properly evaluate speaker recognition systems, most approaches have mixed existing clean speech databases with relatively short background noise data collected separately at the required noise level [3, 6, 7]. However, while the large speech corpora available to researchers through this approach allow a wide variety of speakers to be evaluated for VAD, the limited conditions, and short recordings–typically less than 5 minutes, of existing popular noise datasets such as NOISEX-92 [8], AURORA-2 [9] or freesound.org [1], have limited the ability to adequately test speaker recognition algorithms in a wide range of background noise conditions, especially when the length of the speech recordings exceed the noise.

Having noticed an earlier, similar, shortcoming in the voice-activity-detection literature, We previously [10] collected the comparatively large QUT-NOISE corpus, consisting of 20 recordings of at least 30 minutes of background noise and reverberant responses in a wide variety of locations covering common noise scenarios. This data was then combined with the clean-speech TIMIT [11] database, to create the QUT-NOISE-TIMIT database of over 600 hours of noisy recordings, and was demonstrated as a useful resource for the evaluation of voice-activity detection in noisy conditions.

In this paper, we will take the QUT-NOISE data originally collected for the construction of the QUT-NOISE-TIMIT database, and design the freely available QUT-NOISE-SRE protocol[1] that will allow for the comprehensive evaluation and benchmarking of modern speaker recognition approaches across a wide variety of background noise scenarios, noise levels and reverberation conditions. The paper concludes with a short evaluation of the front-end effect of voice activity detection on state-of-the-art speaker recognition systems.

2. The QUT-NOISE background noise corpus

This section will briefly reintroduce the QUT-NOISE background noise corpus from [10], which will be used to provide the background noise for the construction of the mixed-speech QUT-NOISE-SRE protocol outlined later in this paper.

2.1. Scenarios

In order to provide a simulation of noisy speech in a wide variety of typical background noise conditions, the corpus consists of 20 noise sessions of at least 30 minutes duration each. Two separate noise recordings, separated by at least one day in all but the CAR scenario, were conducted in 10 separate locations over 5 commonly encountered background noise scenarios.

CAFE: The two locations of the CAFE scenario were a typical outdoor cafe environment (CAFE-CAFE) and

This research was funded by the Australian Research Council (ARC) Linkage Grant No: LP130100110.

[1] Visit https://qut.edu.au/research/saivt to download the database and protocol information.