The QUT-NOISE-TIMIT Corpus for the Evaluation of Voice Activity Detection Algorithms

David Dean, Sridha Sridharan, Robert Vogt, Michael Mason

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

ddean@ieee.org, {s.sridharan, r.vogt, m.mason}@qut.edu.au

Abstract

The QUT-NOISE-TIMIT corpus consists of 600 hours of noisy speech sequences designed to enable a thorough evaluation of voice activity detection (VAD) algorithms across a wide variety of common background noise scenarios. In order to construct the final mixed-speech database, a collection of over 10 hours of background noise was conducted across 10 unique locations covering 5 common noise scenarios, to create the QUT-NOISE corpus. This background noise corpus was then mixed with speech events chosen from the TIMIT clean speech corpus over a wide variety of noise lengths, signal-to-noise ratios (SNRs) and active speech proportions to form the mixed-speech QUT-NOISE-TIMIT corpus. The evaluation of five baseline VAD systems on the QUT-NOISE-TIMIT corpus is conducted to validate the data and show that the variety of noise available will allow for better evaluation of VAD systems than existing approaches in the literature.

Index Terms: voice activity detection, speech databases, evaluation protocols

1. Introduction

Voice activity detection (VAD) is the process of detecting which portions of a given audio file contain conversational speech. The VAD process acts as a crucial front-end stage on many related speech processing algorithms, such as speech enhancement, speech coding, and automatic speech and speaker recognition.

One of the main unsolved issues in evaluating VAD 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 VAD systems, many approaches have mixed existing clean speech databases with background noise data collected separately at the required noise level. However, while the large speech corpora available to researchers through this approach allow a wide variety of speakers to be evaluated for VAD, the short recordings, typically less than 5 minutes, of existing popular noise datasets such as NOISEX-92 [1] or AURORA-2 [2] have limited the ability to adequately test VAD algorithms in a wide range of background noise conditions.

To attempt to alleviate this shortcoming in the existing VAD literature, we have collected a large corpus consisting of over 10 hours of background noise in a wide variety of locations covering common noise scenarios. In addition, the reverberant response of the environment was also collected in some locations in order to allow equivalent reverberation to be applied to the inserted speech. By combining our noise recordings with the clean-speech TIMIT [3] database, we created 600 hours of noisy recordings, covering 24,000 individual files over a wide range of noise levels and speech utilisation percentages. By using the large volume of noisy speech available in the resulting QUT-NOISE-TIMIT corpus[1], VAD algorithms can be readily tested against a wide range of common noise scenarios.

2. The QUT-NOISE background noise corpus

This section will outline the construction of the QUT-NOISE background noise corpus, which will be used to provide the background noise for the construction of the mixed speech QUT-NOISE-TIMIT corpus in Section 3.

2.1. Scenarios

In order to provide a simulation of noisy speech in a wide variety of typical background noise conditions, we conducted a collection 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 separate common background noise scenarios.

2.1.1. CAFE

The two locations of the CAFE scenario were a typical outdoor cafe environment (CAFE-CAFE) and a typical indoor shopping centre food-court (CAFE-FOODCOURTB). These recordings are typified by medium to high levels of background speech babble, and kitchen noises from the cafe environment.

2.1.2. HOME

The two locations for the HOME scenario were in the kitchen (HOME-KITCHEN) and living-room (HOME-LIVINGB) of the primary author during typical home activities. The kitchen recordings consist of sections of relative silence interrupted occasionally by typical kitchen noises. The living room recordings consist of children singing, talking and playing alongside television or music noise.

2.1.3. STREET

The two locations for the STREET scenario were at the roadside near typical inner-city (STREET-CITY) and outer-city (STREET-KG) traffic-light controlled intersections. Both recordings largely consist of road traffic noise, with the inner-city recordings also having significant pedestrian traffic as well

[1] For further information regarding the database contact Sridha Sridharan at s.sridharan@qut.edu.au.

as bird noise from a nearby park, while the outer-city recordings mostly consisting of cycles of traffic noise as the traffic lights changed.

2.1.4. CAR

As only one car was available for the CAR scenario, in lieu of two separate locations, the scenario was divided into driving with the windows down (CAR-WINDOWNB) or with the windows up (CAR-WINUPB). Because the car used was only available for a short time, all recordings were conducted on a single day. For both 'locations' the first session was recorded as highway driving, and the second was recorded based upon driving in city and suburban areas. All recordings are characterised by road (and wind for CAR-WINDOWNB) noise and typical car-interior noises (such as indicator, key or luggage-movement noise) but with no radio or speech noise.

2.1.5. REVERB

The two locations for the REVERB scenario were an enclosed indoor pool (REVERB-POOL) and an partially enclosed carpark (REVERB-CARPARK). Both locations were chosen as environments that were expected to produce a large reverberant response. In addition to the large levels of reverberation, the pool environment is characterised by splashing and running noise, while the carpark environment is characterised by nearby road noise and occasional carpark vehicular noise.

2.2. Recording setup

2.2.1. Equipment

The background noise corpus recording was accomplished with a prosumer-quality Zoom H2 handheld stereo microphone recorder. This device was chosen as the quality of the background noise recordings should be higher than typical recording scenarios, allowing any expected recording quality to be easily synthesised.

In order to calculate the room response in the reverberant environments (CAR and REVERB), a single studio-quality KRK RP5 studio monitor was used to play an number of frequency sweeps. This particular studio monitor was chosen as it had a good linear frequency response allowing for the best reproduction of the frequency sweeps used for calculating the room response. Further detail on calculating the room reverberant response is provided in Section 2.3.

2.2.2. Noise recording

Each of the 20 noise sessions were recorded with the Zoom H2 set to record raw stereo WAV output with a sampling rate of 48 kHz, and 16 bits per sample. The recordings were conducted using the rear microphone pair of the Zoom H2, as the greater microphone angular separation (when compared to the front microphone pair) could potentially allow for more useful comparisons to be made between the two channels in future research.

In order to calculate the room response in the reverberant CAR and REVERB scenarios, 10 second frequency sweeps were played with the studio monitor positioned several metres away from the microphone. Each reverberant session contained 12 frequency sweeps, with 6 before the main 30+ minute recording session and 6 after.

Each of the noise sessions collected was manually labeled with the boundaries of the main 30+ minute recording session, as well as the rough locations of each individual frequency sweep in the reverberant sessions. In addition, the locations of any bad portions of data (such as microphone failure) were labeled to allow them to be avoided.

2.3. Calculating the reverberant response

Based on the work of Farina [4], the multiple frequency sweeps used to measure the environment's reverberant response were constructed using a sine wave with the instantaneous frequency varying exponentially from 100 Hz to 20 kHz over a period of 10 seconds. An exponential amplitude modulation term is also added to compensate for the differing energy generated between the low and high frequencies of the raw sweep.

As the six recorded sweeps had been roughly labeled alongside the original noise recordings, the recorded sweeps could then be deconvolved with a clean sweep to arrive at an estimate of the environment's reverberant response. As the response is located through a deconvolution process, precise labeling of the end point of the frequency sweeps was not required. All six impulse response were then averaged in order to attenuate environmental noise captured alongside the sweeps to arrive at the final estimate of the environmental reverberant response, which was then saved alongside the noise recording for future use.

3. The QUT-NOISE-TIMIT mixed speech corpus

This section will outline the construction of the QUT-NOISE-TIMIT mixed speech corpus by mixing background noise sessions chosen from the QUT-NOISE corpus outlined in Section 2 with clean speech chosen from the TIMIT corpus [3].

An overview of the speech sequences available in the final constructed database is shown in Figure 1. In total 600 hours of noisy speech sequences were created over 24,000 files, consisting of 100 files for each of 6 SNRs by 2 noise lengths for each of the 20 recording sessions in the QUT-NOISE corpus.

In order to allow for repeatable division of the QUT-NOISE-TIMIT corpus, two location groups have additionally been defined covering one location for each scenario. These location groups are labelled in Figure 1 as groups A and B.

3.1. Construction

3.1.1. Background noise

Given a particular noise sessions from the QUT-NOISE corpus, for each specified noise length and SNR, 100 background audio scenes were extracted from the recorded noise session. The starting point of each scene was randomly chosen from the labeled main portion of the recording session, excluding the first five minutes[2], and restricted in such a way as to avoid any portion of the recording session labeled as bad data.

Once the location of the 60 or 120 second background audio scene was chosen from the noise session, the left-hand channel was taken and low-pass filtered and down-sampled from the original 48 kHz to the desired 16 kHz sample rate of the final speech sequence files.

3.1.2. Speech events

Once 100 audio scenes had been selected for a particular noise session, length and SNR, speech events were randomly chosen from the TIMIT corpus such that 25 sessions contained less

[2]The first five minutes were excluded to allow for the possibility of training models on noise not used in the final noisy speech corpus.

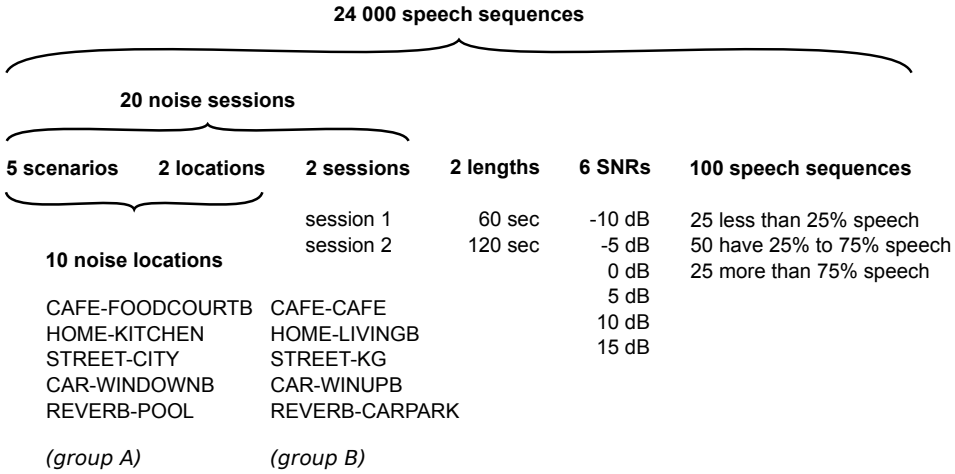

Figure 1: An overview of the speech sequences available in the QUT-NOISE-TIMIT corpus.

than 25% speech (including no speech), 50 sessions had between 25% and 75%, and the remaining 25 sessions had more than 75% speech. In order to ensure that the speech utterances could remain different during training and testing of VAD systems, the full list of TIMIT speech files (covering both the train and test portions of the corpus) were divided into 4 equal sized lists, which were chosen based on which location group (A or B) and session (first or second) each noise recording session belonged to.

In order to allow for longer speech events than the 2-3 seconds of a typical TIMIT utterance, each speech event was randomly combined with it's previous partner with a probability of 50%. In order to emulate the effects of co-talking, if two utterances were combined, they were also overlapped by a time period chosen from a uniform random distribution between 0 and 1 second. Finally, the silence-lengths between the different combined speech events were randomly distributed between 0 and 1 second, then scaled up to fill the remaining space in the background noise sequence.

In order to simulate the reverberant environmental response, speech events intended for combination with the reverberant CAR and REVERB background noise sessions were convolved with the saved reverberant responses of the environment in which they were to be inserted.

3.1.3. Combining speech with background noise

Before the clean TIMIT speech files and selected background noise sequence could be combined to form the final noisy speech sequences in the QUT-NOISE-TIMIT corpus, the speech events and background noise were scaled to match the desired SNR. This process was performed by first scaling the individual speech events to a ITU-T standard (P830) reference signal level of -26 dBov, and then scaling the background noise energy in relation to the reference speech level according to the desired SNR. This approach ensured that all background noise events had a well defined signal level, even in files in which no speech events had been inserted.

Once the background noise levels and speech events had been scaled appropriately, the final noisy speech sequences were obtained by a sample-by-sample summing of the speech and background noise sessions where they overlapped. While this approach has resulted in some clipping at low SNR levels due to high noise energy, it was deemed more important to maintain a consistent reference energy level in similar SNR sequences.

3.2. Metadata

Alongside each of the constructed audio files created for the QUT-NOISE-TIMIT corpus, two label files are also provided indicating both the locations of the original TIMIT speech files, and the location of the distinct non-overlapping speech events contained within. The latter event labeling method was determined using by removing the outside-silence periods from the TIMIT speech files (using the word-level labeling provided with TIMIT) and combining overlapping speech into a single labeled 'speech'. Remaining non-speech events were then labeled as 'nonspeech'. These event-label files will serve as the ground truth for speech detection experiments.

4. Voice activity detection experiments

In order to provide a common reference to facilitate simple results comparison and also to validate the collected data, five baseline VAD systems have been evaluated using the QUT-NOISE-TIMIT corpus.

4.1. Baseline systems

Five baseline VAD systems were chosen for evaluation using the QUT-NOISE-TIMIT corpus. These VAD systems were

- ITU G729 Annex B [5] (G729B),
- ETSI Advanced Front-end VAD [6] (ETSI),
- Ramirez's long-term spectral divergence [7] (LTSD),
- Sohn's model-based likelihood ratio [8] (Sohn), and
- A GMM based learning approach using MFCC features (GMM-MFCC).

The G729 and ETSI VAD systems are performed directly using the publicly available VAD code released as part of those standards and were run as-is without any training. The LTSD and Sohn approaches use frame-by-frame speech-likelihood scores calculated using similar approaches to the published algorithms in [7] and [8]. Finally, the GMM-MFCC system uses the ground truth event labeling in the training divisions to train *speech* and

nonspeech models on MFCC-based speech features. The frame-by-frame speech-likelihood scores for the GMM-MFCC system are then given as the difference between the log-likelihoods of the speech and nonspeech GMM models.

The LTSD, Sohn and GMM systems were further smoothed by a 1-second median filter to attenuate short-term variation, and then thresholded using thresholds tuned using the training-divisions outlined in the evaluation protocol.

4.2. Evaluation protocol

The QUT-NOISE-TIMIT corpus is suitable for division for the training and testing of VAD systems over a large range of operating conditions. For the particular set of VAD experiments conducted here, we have chosen to assume knowledge of the broad SNR level of the target environment, but no knowledge of the actual location or even the scenario of the target environment. Accordingly, for each of the three broad SNR levels chosen (*low*: 15, 10 dB; *medium*: 0, 5 dB; *high*: -10, -5 dB) the VAD systems were trained on one location group (A or B) and tested on the other (B or A) and vice-versa. In this way the VAD systems were trained and tested over all scenarios at a particular broad noise level, but with no prior knowledge of the actual locations they were tested in.

Performance of the final VAD segmentation results were measured by comparing the segmentation results to the ground truth event-label files created alongside the QUT-NOISE-TIMIT corpus. VAD segmentation results were measured according to the miss rate (MR) measuring the proportion of true-speech frames not detected as speech, and the false-alarm rate (FAR) measuring the proportion of non-speech frames incorrectly detected at speech. These two error rates were combined in the half-total-error rate (HTER), being the average of the MR and FAR. The segmentation-tuning performed for the LTSD, Sohn and GMM-MFCC systems were also based on minimising the HTER in the training partitions.

4.3. Experimental results

The performance of the five VAD systems is shown according to the HTER in Figure 2 for each of the three broad noise levels. Each of the results is shown as a stacked bar indicating the proportions of the MR (bottom, darker) and the FAR (top, lighter) to the total HTER. Accordingly, the height of the MR and FAR sections is half of the actual levels of that error type.

From an analysis of the results shown in Figure 2, it can be seen that the GMM-MFCC, LTSD and Sohn systems typically outperform the standards-based ETSI and G729B VAD systems. The large volume of noisy speech available in the QUT-NOISE-TIMIT over a wide range of scenarios has allowed for the choice of relatively robust segmentation thresholds for these systems in comparison to the built-in threshold approaches of the standards-based systems. In particular, the ability of the GMM-MFCC VAD system to learn the cross-scenario characteristics of speech and nonspeech provided the best performance across all noise levels.

5. Conclusion

Within this paper, we have outlined the development of the QUT-NOISE-TIMIT corpus, an extensive noisy speech corpus for the evaluation of automatic VAD systems. This corpus consists of 600 hours of noisy speech sequences constructed from over 20 hours of background noise collected over a wide variety of typical VAD scenarios, including environment reverberant responses, combined with the TIMIT clean speech corpus. We have also demonstrated the use of the QUT-NOISE-TIMIT corpus for the evaluation of a number of baseline VAD algorithms.

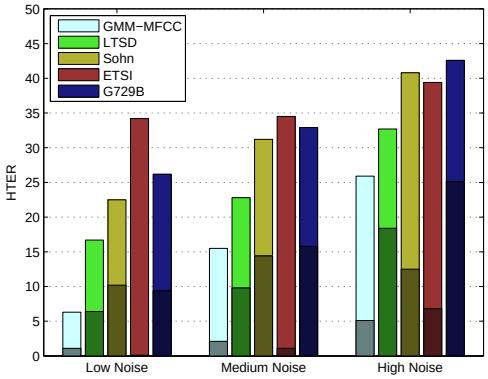

Figure 2: HTER performance of all VAD systems over the three broad noise levels. Each bar is divided according to the contribution of the MR (bottom, darker) and the FAR (top, lighter) to the overall HTER.

We believe that this database will form a solid basis for the development and evaluation of robust VAD algorithms that can operate across a wide variety of noise scenarios. For further information regarding the database contact Sridha Sridharan at `s.sridharan@qut.edu.au`.

6. References

[1] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[2] D. Pearce and H. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Sixth International Conference on Spoken Language Processing*, Citeseer, 2000.

[3] W. Fisher, G. Doddington, and K. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. DARPA Workshop on Speech Recognition*, pp. 93–99, 1986.

[4] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *108th Audio Engineering Society Convention*, Citeseer, 2000.

[5] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lambline, and J.-P. Petit, "ITU-T Rec. G.729 Annex B: A silence compression scheme for G.729 optimized for V.70 digital simultaneous voice and data applications," tech. rep., ITU, 1996.

[6] "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms, v 1.1.4.." ETSI ES 202 050, 2005.

[7] J. Ramírez, J. Segura, C. Benítez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, pp. 271–287, 2004.

[8] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, pp. 1–3, January 1999.