JHU Sys1 Description

1 Abstract

JHU Sys1 served as the baseline system for the JHU team. This is essentially an out-of-the-box system built through previous research using the Callhome corpus, and as a result, all i-vector processing occurs at 8kHz. Only the SAD algorithm is new, while the rest is held over from previous research.

2 Data Resources

The SAD algorithm used audio from European Parliament videos. The clean, 16kHz microphone recordings were subsequently augmented with noise and music from the MUSAN corpus, and reverberated with impulse responses from the AIRS corpus.

The UBM and T matrix for i-vector extraction were trained with data from SRE NIST evaluations 2004, 2005, 2006, and 2008. PLDA was trained with data only from NIST SRE 2004 and 2005.

3 Algorithm Details

3.1 SAD

The SAD algorithm was a 5-layer TDNN with a final sigmoid layer and used log-compressed magnitudes from 35 mel filters as input. The first layer aggregated 5 frames (50ms) of context, and each subsequent layer doubled the width of a 3 tap filter, resulting, in the end, in +/- 640 ms of context at the final sigmoid layer. ReLU non-linearities were used between all layers (except the final sigmoid non-linearity). While the system was initially trained on the European Parliament audio (with a cross-entropy metric), the final layer was retrained using the DIHARD dev data. Final output probabilities were smoothed with a 500ms median filter prior to thresholding at 0.50.

3.2 Segment Clustering

Speech was segmented into approximately 2 second windows with 1 second hops, 20 MFCCs were extracted every 10ms after resampling audio to 8kHz, and a 64-dimensional i-vector was extracted for each segment using a 1024-component UBM. These segments i-vectors were scored with PLDA (trained with segments labeled for speaker and channel instead of just speaker), and clustered using AHC (average distance at merge). The stopping threshold used was the same as the threshold previously used for Callhome, though it was also confirmed to be a reasonable choice by sweeping the dev data.

4 System Requirements

- Hyperparameters for the i-vector system were trained using CPUs only. The processors themselves were selected according to a grid scheduler, and so a variety were used (and the specific choice was not important).
- The SAD system was built in PyTorch and trained on a single GeForce GTX 1080 GPU card with 12GB of available memory.
- All elements were run on CPUs only at test time (again determined by scheduler with no concern for CPU specifics).
- **To process 10-minute recording**: 30 seconds (12 seconds for SAD marks, 18 seconds for clustering)