JHU-CLSP SYSTEM DESCRIPTION FOR OPENSAT SAD EVALUATION

Vimal Manohar[1]*, Jesús Villalba*[1]*, Raghavendra Reddy Pappagary*[1]*, Jan Trmal*[1]*,*
Jonas Borgstrom[2]*, Pedro Torres-Carrasquillo*[2]*, Doug Sturim*[2]*,*
Najim Dehak[1]*, Daniel Povey*[1]*, Sanjeev Khudanpur*[1]

[1]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA.
[2]MIT Lincoln Laboratory, Lexington, MA, USA
{vmanoha1,jvillal7}@jhu.edu

ABSTRACT

The JHU-CLSP submission to the OpenSAT2017 speech activity detection evaluation consists of a fusion of several neural network based VADs. These systems combine TDNN and LSTM layers. Two of the VADs were trained on Babel and Fisher data augmented with reverberation, noise and music. Another system was trained using only the OpenSAT VAST development data. A fourth system was trained on TIMIT and CMU Child data also augmented with noise and reverberation.

Index Terms— SAD, VAD, TDNN, LSTM

1. INTRODUCTION

The JHU-CLSP submission to the OpenSAT2017 speech activity detection evaluation consists of a fusion of several neural network based VADs. The speech activity detection networks contain TDNN layers followed by LSTM or statistics pooling [1] for long-context. The neural network computes the speech/non-speech posterior probabilities. Afterwards, we use those posteriors as emission probabilities of a two-states HMM with duration constraints and obtain the speech activity labels by Viterby decoding.

For the VAST condition, we submitted five systems:

- Primary: majority voting fusion of contrastives 1–4.
- Constrastive 1: TDNN with statistics pooling trained on 10 Babel languages [2] and Fisher English [3] perturbed with various room impulse responses, additive noise and music (JHU set).
- Contrastive 2: TDNN-LSTM trained on the same data as contrastive 1.
- Contrastive 3: TDNN with stats pooling trained on OpenSAT2017 dev set.
- Contrastive 4: TDNN with stats pooling trained on OpenSAT2017 dev set; and TIMIT and CMU Child with noise and reverberation (MITLL set).

For the Babel condition, we submitted one primary system.

2. TRAINING DATA

We trained the neural networks on three different datasets. The first one was prepared by the JHU team while the second one was prepared by the MITLL team. The third dataset was the OpenSAT development set.

2.1. JHU training set

As training data for the neural network, we use data from 10 different Babel languages that were publicly available - Assamese, Bengali, Cantonese, Georgian, Haitian, Pashto, Swahili, Tagalog, Turkish, Vietnamese, and Fisher English. We select about 10 hours from each corpus. To improve the robustness of the speech activity detector to unseen data, we add various perturbations:

- Reverberation: We use synthetic room impulse responses (RIRs) [4]. For each recording, we create 10 copies each reverberated with a randomly selected RIR.
- Additive noise: Many different kinds of noise recordings, both foreground and background noises, from the MUSAN corpus [5] were added. For each recording from the base corpus, several noise recordings were randomly selected, perturbed using the synthetic RIRs and added at different points in the recording. For 5 copies, we added noise from MUSAN corpus scaled between 20dB and -5dB; for the other 5 copies, we added music from MUSAN corpus scaled between 5dB and -5dB.
- Speed and volume perturbation: We apply speed perturbation [6] by randomly choosing one of the 3 speeds – 0.9, 1.0 and 1.1. We additionally scale the volume randomly by a scale between 0.03125 and 2. This is to help the neural network be less sensitive to the energy

of the input signal. This is important for robustness because the input features to the neural networks do not have cepstral mean subtraction or any other normalization.

The datasets that we used don't include speech activity detection labels. To infer those labels we performed force alignment using HMM-GMM systems trained on the respective corpora. For the Babel languages, we used the data in the Full LP condition to train the HMM-GMMs. For the Fisher English, we used a 100 hour subset to train the HMM-GMMs.

We align the manual segments of each corpus with the corresponding transcription using a speaker-adaptively trained HMM-GMM system to create alignments. These are converted into phone labels, which are deterministically mapped to 2 classes – speech and silence.

We separately decode both the manual segments and the regions outside the manual segments using a speaker-independent LDA+MLLT HMM-GMM system to get best path alignments. For outside the manual segments, we keep only frames that are silence. For the manual segments, we keep frames for which the transcription-constrained alignments and the decoded hypothesis are both the same class (silence or speech).

2.2. MITLL training set

The multicondition data was simulated by concatenating utterances from the TIMIT and CMU Child corpora. For each simulated file, 10 utterances were chosen from the two corpora, with randomly set intervals and gains. Reverberation was introduced by applying a room impulse response from the Voice Home corpus. Finally, additive noise was introduced from the noise and music subsets of the MUSAN corpus, and mixed at an SNR chosen between 20 dB and -3dB.

2.3. OpenSAT development set

We also trained a VAD on the development set in order to have an upper bound of the performance that we could achieve by training and testing on the same dataset.

3. NEURAL NETWORK DESCRIPTION

We use a time-delay neural network [7] with either LSTMP layers [8] or statistics pooling [1] layers interleaved with TDNN layers to provide a long-context. We use 4 TDNN layers and 2 layers of LSTMP or Statistics pooling. The overall context of the network is set to be around 1s, with around 0.8s of left context and 0.2s of right context. The network is trained with cross-entropy objective to predict the speech/non-speech labels.

3.1. Input features

As input features to the neural networks, we use 28-dimensional MFCC features extracted from the bandwidth 330Hz to 3000Hz, which corresponds to the telephone bandwidth. This makes the neural network usable for both telephone speech and microphone speech.

3.2. Auxiliary task outputs

The neural network is trained to predict some auxiliary features in addition to speech/non-speech labels. These act as regularizers to increase the robustness of the neural networks. At test time, these outputs are not computed, and only the main output for speech/non-speech detection is evaluated. We use two auxiliary features:

- Ideal-ratio-mask (IRM) – This is the ratio of Mel filterbank coefficients of the clean signal to the sum of Mel filterbank coefficients of clean and noise signals. This is converted to log-domain so that they represent log-probability of the sub-band containing speech (clean) energy.

$$\text{IRM}(f) = \frac{S(f)}{S(f) + N(f)}, \quad f = 1, 2, \ldots, K \quad (1)$$

  where $S(f)$ and $N(f)$ are the Mel filterbank coefficients of the clean and noise signals respectively for sub-band f.

- Music labels – For the recordings, where we add music during perturbation, we add auxiliary labels that specify music/non-music.

The auxiliary tasks are trained simultaneously with the main speech/non-speech task in a multi-task framework. For predicting music/non-music labels, we use a cross-entropy objective. For predicting sub-band log-IRM features, we use cross-entropy objective.

4. DECODING

Given the test data, its MFCC features are propagated through the neural network to get speech/non-speech log-posteriors, which are converted to log-likelihoods by subtracting the log-priors. The log-likelihoods are fed to a HMM Viterby decoder. The HMM has a minimum duration constraint of 10 frames for silence and 30 frames for speech. At the end of the minimum duration, there is a self-loop with probability 0.99 for speech and 0.9 for silence. To bias the decoding towards, speech we allow a transition from the end of silence class to the start of the speech class with a probability of 0.009 and a transition to start of silence class with a probability of 0.0009. We force a transition to the start of speech class at the end of

Table 2. Total CPU time (s) on VAST eval data

Process	Statistics pooling	LSTM
Nnet propagation	2056	20990
Decoding	254	466
Post-processing	25	17
Total	2335	21473

silence class with a probability of 0.009. The remaining probability mass is for transition to end state. The best path from the HMM decoder is post-processed to add a 0.2s padding on the edges of the speech segments, followed by merging small silence regions that are less than 0.3s long into the adjacent speech regions.

5. RESULTS

Table 1 shows the results of or systems on the VAST development set.

Table 1. VAD Results on the VAST dev data

System	Avg Cost	Avg P_{miss} (%)	Avg P_{fa} (%)
Primary (Fusion)	0.083	5.06	18.03
Contrast 1 (Stats JHU)	0.088	5.56	18.66
Contrast 2 (LSTM JHU)	0.090	3.22	26.6
Contrast 3 (Stats OpenSAT dev)	0.077	6.82	10.41
Contrast 4 (Stats MITLL + OpenSAT dev)	0.119	12.7	9.42

6. TIMING

Table 2 shows CPU time for the two neural architectures that we used.

7. REFERENCES

[1] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, Acoustic modelling from the signal domain using cnns. in INTERSPEECH, 2016, pp. 34343438.

[2] M. Harper, IARPA Babel Program. 2014.

[3] C. Cieri, D. Miller, and K. Walker, The fisher corpus: A resource for the next generations of speech-to-text. in LREC, 2004, vol. 4, pp. 6971.

[4] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, A study on data augmentation of reverberant speech for robust speech recognition, in Acoustics, speech and signal processing (icassp), 2017 ieee international conference on, 2017, pp. 52205224.

[5] D. Snyder, G. Chen, and D. Povey, Musan: A music, speech, and noise corpus, arXiv preprint arXiv:1510.08484, 2015.

[6] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, Audio augmentation for speech recognition. in INTERSPEECH, 2015, pp. 35863589.

[7] V. Peddinti, D. Povey, and S. Khudanpur, A time delay neural network architecture for efficient modeling of long temporal contexts. in INTERSPEECH, 2015, pp. 32143218.

[8] H. Sak, A. Senior, and F. Beaufays, Long short-term memory recurrent neural network architectures for large scale acoustic modeling, in Fifteenth annual conference of the international speech communication association, 2014.