INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

The SRI System for the NIST OpenSAD 2015 Speech Activity Detection Evaluation

Martin Graciarena[1], *Luciana Ferrer*[2], *Vikramjit Mitra*[1]

[1] SRI International, California, USA
[2] Departamento de Computación, FCEyN, Universidad de Buenos Aires and CONICET, Argentina
martin@speech.sri.com

Abstract

In this paper, we present the SRI system submission to the NIST OpenSAD 2015 speech activity detection (SAD) evaluation. We present results on three different development databases that we created from the provided data. We present system-development results for feature normalization; for feature fusion with acoustic, voicing, and channel bottleneck features; and finally for SAD bottleneck-feature fusion. We present a novel technique called test adaptive calibration, which is designed to improve decision-threshold selection for each test waveform. We present unsupervised test adaptation of the fusion component and describe its tight synergy to the test adaptive calibration component. Finally, we present results on the evaluation test data and show how the proposed techniques lead to significant gains on channels unseen during training.

Index Terms: speech activity detection, noise robustness, channel degradation

1. Introduction

Speech activity detection (SAD) is an essential pre-processing step in most speech-processing tasks, such as speech recognition, speaker and language identification, and emotion detection. For those tasks, the goal of SAD is filtering out non-speech regions, keeping only the regions containing speech, the information relevant for the task. SAD is also a task in its own right, in which large amounts of audio are collected and searched for speech, which can be later provided to a human analyst. But SAD can be difficult when the input audio is degraded by noise at low signal-to-noise (SNR) ratios. Therefore, designing SAD systems that are robust to noise and background distortions is important.

This paper describes the development of a SAD system for the OpenSAD 2015 evaluation. It describes the OpenSAD evaluation in section 2, the system overview in section 3, the SRI development databases in section 4, the experimental setup in section 5, the experimental results in section 6, the evaluation results in section 7, and finally conclusions in section 8.

2. OpenSAD 2015 Evaluation

The OpenSAD 2015 was an open evaluation organized by the National Institute of Standards (NIST). The evaluation was intended to provide SAD system developers with an independent performance evaluation on a variety of audio data. The intention of this evaluation was to advance technology that both can select audio file regions of speech for a human user to examine and, provided some configuration changes, can be used for downstream automatic processing by technologies, such as speech recognition, speaker identification, language identification, or machine translation.

The NIST OpenSAD evaluation has ties to the DARPA Robust Automatic Transcription of Speech (RATS) program. The RATS program was designed to advance the current state of the art in identifying speech activity regions, keywords, languages and speakers in signals from distorted, degraded, weak, and/or noisy communication channels. The OpenSAD data used a sequestered RATS evaluation set (from LDC) [1]. Some of that data included noises and a variety of transmitter/receiver radio-link channels. The source speech data for transmission originated as telephone speech (landline or cell) over public telephone networks.

3. System Overview

Figure 1 below shows the complete SRI SAD system that was submitted to the OpenSAD 2015 evaluation. It was composed of five subsystems, each with a three-way feature fusion of acoustic, voicing, and channel bottleneck (BN) features. These subsystems were combined by fusing SAD BN features. The fusion output was the tertiary submission. When we added a novel system called test adaptive calibration (TAC), described below, to the tertiary system, this became the secondary system. Finally, when we added a test unsupervised adaptation module to the secondary system, this became the primary system

The proposed system included multiple features. Some were acoustic features modeling spectral characteristics; others were voicing features aimed at capturing spectral pitch harmonics; and finally, some were channel BN features aimed at modeling the channel characteristics.

The acoustic features included Mel-Cepstral (MFCC) [2]; Perceptual Linear Prediction (PLP) [3], Power Normalized Cepstral Coefficient (PNCC) [4], Normalized Modulation Cepstral Coefficient (NMCC) [5], and Time Domain Gammatone Cepstral Coefficient (TDGCC) features. The voicing features included the Combo [6] feature from UTD and the Kaldi toolkit pitch and voicing features [7]. Finally, we also used a channel bottleneck (CBN) [8] feature computed from a deep neural network (DNN) trained to classify the channels in the OpenSAD training data.

http://dx.doi.org/10.21437/Interspeech.2016-550

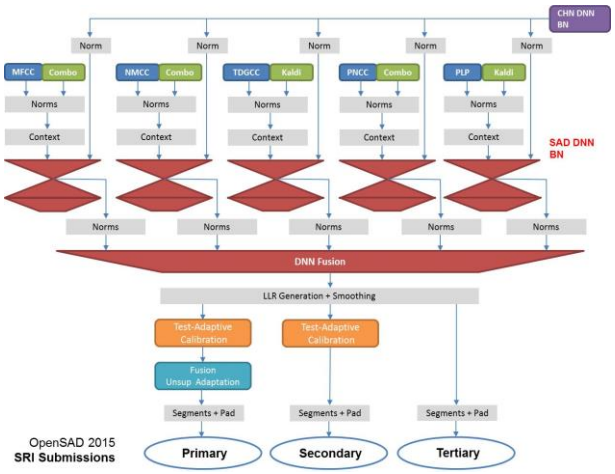

Figure 1: *Schematic diagram of SRI system for OpenSAD 2015 evaluation.*

4. SRI Development Database

Using the provided OpenSAD data listed in Section 2, we created the three development conditions shown below in Table 1. The training data is always restricted to a subset of the original RATS channels (source, B, D, E, F, G, H) by the evaluation plan. For testing, we create a matched condition using channels seen in training (except source) and an unseen condition using novel channels. Finally, we also create a mismatched condition using the original RATS channels to test the generalization of our SAD systems on these channels. This mismatched condition was setup by leaving two channels out at a time for training, testing on the held-out channels and pooling results from those three experiments.

Table 1. *SRI Development Database. Source is abbreviated as S.*

Condition	Train Channels	Development channels
Matched	S, B, D, E, F, G, H	B, D, E, F, G, H
Novel		XA, XH, XI, XK, XN
Mismatch	S, D, E, F, G	B, H
	S, B, E, F, H	D, G
	S, B, D, G, H	E, F

5. Experimental Setup

Currently, deep neural network-based SAD systems are the state of the art [9, 8, 10, 11]. As a model, these systems use a deep neural network (DNN) trained to predict the posterior of the speech and non-speech classes at the output layer. The posteriors are converted into log-likelihood ratios (LLRs) by using Bayes rule, assuming equal priors for both classes. In a final step, these LLRs are smoothed by averaging their values over a rolling window typically 31 to 71 frames long. The final SAD decisions are made by thresholding these LLRs, with a threshold chosen based on the desired operating point. For some applications, the resulting speech regions (any contiguous frames for which the LLR value was above the threshold) are padded with a certain number of frames on each side. This padding reduces the amount of missed speech near the detected speech regions while potentially increasing the false alarm rate.

In this paper, the SAD model is a DNN with three hidden layers. Each layer included five hundred neurons. In some experiments, the middle hidden layer is a bottleneck layer of reduced size. The input to the DNN was a concatenation of cepstral, voicing and channel BN features, as described in Section 4, over a window of 31 frames.

Two types of error can be computed for SAD: (1) the miss rate (the proportion of speech frames labeled as non-speech), and (2) the false alarm rate (the proportion of non-speech frames labeled as speech). In Phase 4 of the RATS program, a "forgiveness" collar of 2.0 seconds was used around all annotated speech regions. The false alarm errors over those regions were disregarded. Here, we used this same collar for our results and the actual decision cost function (DCF) as the metric. To obtain the DCFs, we post-processed the LLRs from each of the systems as described before in this section, by using an average filter of 41 frames, thresholded them, padded each resulting speech region with 0.3 seconds for development, for the primary system and 0.4 seconds for the other two systems, on each side, and finally computed the false alarm (Pfa) and miss (Pmiss) rates for each channel to calculate the DCF. The official DCF for the OpenSAD 2015 evaluation is given by:

$$DCF = 0.75 * Pmiss + 0.25 * Pfa$$

The optimal theoretical threshold for this DCF is given by -1.10. This is the threshold that minimizes the DCF for the specified weights when LLRs are well-calibrated (without considering padding and smoothing). This is the threshold we use for our experiments.

6. Experimental Results

In this section, we show results for different normalization methods, fusion results, test-adaptive calibration and adaptation.

6.1. Baseline DNN SAD and Feature Normalization

We explored multiple feature-normalization approaches: *raw* - no normalization; *mvn* – waveform-level normalization; *2smvn* – mean and variance normalization over a window of 2 seconds located around each frame; *mvn+2smvn* – feature-level fusion of the mvn and 2smvn normalizations.

Table 2. *DCF of baseline system and different feature-normalization approaches.*

Feature	Norm	Match	Mismatch	Novel
MFCC	raw	2.50	8.37	9.54
	mvn	2.60	4.57	5.81
	mvn + 2smvn	2.11	4.52	5.30

Table 2 shows results on our development dataset for a system based only on MFCC features, normalized in these four ways. Results show that waveform-level normalization produces high error reduction in mismatched conditions compared to no normalization. Clear gain was achieved from the dual normalization in all conditions. In all further experiments presented here, we use dual normalization.

6.2. Feature-Level Fusion

We explored feature-level fusion by first investigating two-way fusion of the MFCC acoustic feature with different voicing and channel BN features, and then three-way feature fusion.

In almost all cases, Table 3 shows a gain from feature fusion over MFCC. The three-way fusion is best on the matched condition and second-best on the mismatched and novel conditions. Two-way fusion with Combo is best on the mismatched condition, while two-way fusion with Channel BN is best on the novel condition. Overall, the best trade-off across conditions seems to be given by the three-way fusion.

Table 3. *DCF of feature-level fusion systems with MFCC, voicing and channel BN features (CBN).*

System	Match	Mismatch	Novel
MFCC	2.11	4.52	5.30
MFCC + CBN	2.02	4.49	4.86
MFCC + Combo	2.06	4.04	6.13
MFCC + Kaldi	2.05	4.37	5.67
MFCC+Combo+CBN	1.93	4.32	5.11

6.3. BN-Level Fusion

We present DCF results for the five different systems and the BN level fusion as well. Each of these systems was a three-way feature fusion of three different features: one acoustic feature, one voicing feature, and the channel BN features.

For the individual results of the five systems, we used the DNN topology with three hidden layers, each of size 500. For the BN-level fusion experiment, we used a different DNN with three hidden layers of 500, 13 and 500 neurons. For this network, we first trained it on the training set. Next, we ran the network and extracted the outputs of the 13-dimensional bottleneck layer. Those features were then fed to the BN fusion network.

We also show results for a five-way fusion system. This fusion network was a DNN with two hidden layers of 500 and 100 neurons each, taking the raw 13-dimensional BN features from each of the 5 systems as input, appended with the waveform-level mvn normalized ones to get a 26-dimensional vector.

From Table 4, we conclude that on the matched condition, the five systems perform similarly. On the mismatched condition, the best model uses the PNCC feature, while on the novel condition, the PLP feature performs best. The BN fusion system improves greatly on the matched and mismatched condition over the best single system in each case. However, on the novel condition, it is close to the second-best performance.

6.4. Test Adaptive Calibration

We proposed a novel approach to improve calibration, which we call test adaptive calibration (TAC). The proposed approach is done independently for each test waveform and has two steps: (1) model fitting, followed by (2) a threshold correction step (i.e., LLR distribution shifting). The goal of this step is to allow for a single threshold to be optimal across all test waveforms.

Prior background includes the Sadjadi paper [6], where the threshold is also estimated on the test waveform. The main difference between our approach and that described in [6] is that [6] uses a two Gaussian model, whereas in this paper, we estimate the model size. Additionally, here, we use a tied-variance GMM model.

Table 4. *DCF of three-way feature-level fusion systems and of their BN-level fusion system.*

System	Match	Mismatch	Novel
MFCC+Combo+CBN	1.93	4.32	5.11
PNCC+Combo+CBN	1.93	3.92	6.17
TDGCC+Kaldi+CBN	1.92	4.49	6.33
NMCC+Combo+CBN	1.93	4.25	6.24
PLP+Kaldi+CBN	1.98	5.65	4.87
5-way BN Fusion	1.84	3.45	5.29

6.4.1. Model Fitting

In our TAC approach, the first step is fitting a Gaussian mixture model to the LLR distribution. We used a tied-variance Gaussian mixture model; therefore, all the Gaussian models had the same variance. This was essential for good performance. For the model size (i.e., the number of Gaussians), we chose between two and three Gaussians. We found the best number of Gaussians by using the Bayesian Information Criterion (BIC) [12].

Figure 2 shows the model fit to one RATS waveform from channel G. The true LLR, obtained with an MFCC-only system trained on mismatched channels, is shown in green, superimposed by the true distributions (extracted from the annotations) of speech (S) in blue and the non-speech (NS) in red. The model approximates reasonably well the LLR distribution; however, it fails in modeling the speech distribution, which is clearly non-Gaussian.

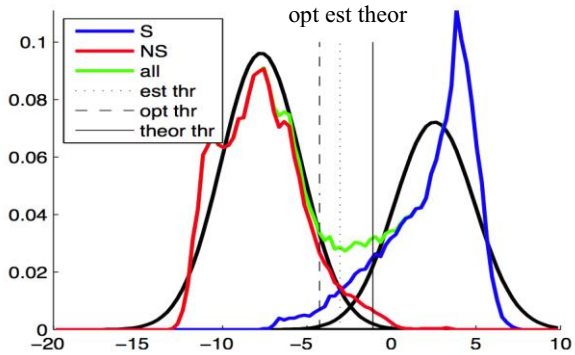

Figure 2: *Example of LLR distribution on one sample from the RATS data, Gaussian model fit, and thresholds (estimated (est), optimal (opt), and theoretical (theor)). The optimal threshold is computed as the threshold that minimizes the DCF for this particular waveform. The theoretical threshold is -1.10 (see Section 5). The estimated threshold is computed as explained in Section 6.4.2.*

6.4.2. LLR Distribution Fitting

The second step of our TAC approach is threshold correction or LLR distribution shifting. Given the GMM estimated in the previous step, we assume that the right Gaussian corresponds to the speech distribution and the GMM composed by the rest

of the Gaussians to the non-speech distribution (after renormalizing the weights to add to one). Given these assumptions, we calculate the threshold that minimizes the DCF for these distributions. In order to mitigate extreme values in the estimated threshold caused by inaccuracies in the estimated GMM, we finally interpolate this estimated threshold with the theoretically (globally) optimal, which, for the openSAD DCF function is given by -1.1. The interpolation weights were optimized in the development data. Finally, the LLR distribution is shifted by the difference between this interpolated value and -1.1. Then the resulting LLRs can be thresholded, as in the other systems, using the -1.1 threshold.

Figure 2 shows three thresholds: (1) the theoretical one at the -1.1 value with a solid line; (2) the estimated one before interpolation with the theoretical with a dotted line; and (3) the optimal one (computed to minimize the DCF given the speech/non-speech empirical LLR distributions for this waveform) with a dashed line. It is clear that, for this sample, the estimated threshold from the proposed model is closer to the optimal one than the theoretical one.

6.5. Unsupervised Fusion Adaptation

The OpenSAD evaluation was designed such that channel information was provided and included a mandate in the evaluation plan indicating that adaptation should be done within each channel.

We performed unsupervised adaptation per channel directly on the DNN fusion model and the steps are detailed below:

- Generate adaptation hypothesis from LLRs with an un-adapted model, labeling as speech all frames with LLR above -1.1+0.5 and as non-speech all frames with LLR below -1.1-0.5.
- Perform model adaptation on the BN-level fusion DNN to these hypotheses within each channel with model regularization. Model regularization is a penalty term on the change in the model parameters.
- Use the channel-adapted fusion model on each channel to generate new LLRs.

Table 5 shows the DCF results from using the BN fusion system, followed by using unsupervised fusion adaptation, next by using only TAC on the fusion output, and finally by combining TAC followed by unsupervised fusion adaptation.

From Table 5, we conclude that unsupervised adaptation produces a small degradation on the matched condition, a small gain on the mismatched condition, and approximately the same result on the novel condition. TAC produces a small degradation on the matched condition, similar gains on the mismatched condition as provided by unsupervised adaptation, and a large gain on the novel condition. The combination of both techniques achieves a similar small degradation on the matched condition and a gain on the mismatched and novel conditions over the best result in each condition. This shows clear synergy when using these two techniques together. The reason for this finding is that TAC improves the calibration in each waveform, improving the hypothesis generation for unsupervised adaptation over all waveforms within a channel. The combination of both techniques therefore achieves better results than the best technique alone.

Table 5. *DCF for the 5-way BN-level fusion baseline, with unsupervised adaptation (Adapt) applied, with TAC applied, and with both techniques applied.*

System	Match	Mismatch	Novel
BN Fusion	1.84	3.45	5.29
BN Fusion + Adapt.	1.93	3.10	5.30
BN Fusion + TAC	1.89	3.20	4.49
BN Fusion + Adapt. + TAC	1.93	3.00	4.26

7. Evaluation Results

In Table 6, we present the SRI system results on the NIST OpenSAD 2015 evaluation data. The key for the submission descriptions is shown in Figure 1. The results are shown in three channel groups: (1) the average over the RATS channels, (2) the average over the novel channels, and (3) the average over all the channels.

Table 6. *DCF of SRI submitted systems on the OpenSAD 2015 evaluation data.*

System	RATS	Novel	All
SRI Primary	4.99	1.48	3.64
SRI Secondary	4.71	1.74	3.57
SRI Tertiary	4.85	2.19	3.82

From Table 6, we conclude that on the RATS channels, which include six matched channels and two unseen channels, the three submissions perform quite similarly. The degradation pattern on the matched condition is similar between Tables 5 and 6. On the novel channels, the secondary submission, with only TAC applied significantly reduces the error on the novel condition. The primary submission, which adds unsupervised adaptation, provides an additional gain compared to the secondary submission. These findings coincide with what was observed on the development data. Finally, on all channels, the DCF is lowered by the secondary submission, but is increased by the primary submission. The DCF difference between RATS and Novel is likely due to different label correction procedures.

8. Conclusions

We presented the SRI system for the NIST OpenSAD 2015 evaluation. We showed that feature normalization significantly impacted performance. Feature fusion provided error reductions on most conditions. BN-level fusion gave major error reductions compared to the best model on the matched and mismatched conditions. The proposed test adaptive calibration combined with unsupervised adaptation of the BN fusion model gave important error reductions, especially on the mismatched condition.

9. Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-15-C-0037. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. Distribution Statement A: Approved for Public Release, Distribution Unlimited.

10. References

[1] K. Walker and S. Strassel, "The RATS radio traffic collection system," *Odyssey 2012: The Speaker and Language Recognition Workshop*, 2012.

[2] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Acoustical Society of America Journal*, vol. 87, pp.1738–1752, Apr. 1990.

[4] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring", *Proc. ICASSP*, pp. 4574–4577, 2010.

[5] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," *Proc. ICASSP 2012*, March 2012, pp. 4117–4120.

[6] S. O. Sadjadi and J. H. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, pp. 197–200, 2013.

[7] Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal,and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," *Proceedings of ICASSP 2014.*

[8] J. Ma, "Improving the speech activity detection for the DARPA RATS phase-3 evaluation," *Interspeech 2014.*

[9] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on YouTube using deep neural networks," *Interspeech*, Lyon, France, Aug. 2013.

[10] S. Thomas, G. Saon, M. Van Segbroeck, and S. S. Narayanan, "Improvements to the IBM speech activity detection system for the DARPA RATS program," *Proc. ICASSP*, Brisbane, Australia, May 2015.

[11] L. Ferrer, M. Graciarena, and V. Mitra, "A phonetically aware system for speech activity detection," *Proc. ICASSP*, Shanghai, China, March 2016.

[12] G.Shwarz, "Estimation the dimension of a model," *The Annals of Statistics*, Vol.6, pp. 461–464, 1978.