Developing a Speech Activity Detection System for the DARPA RATS Program

Tim Ng[1], *Bing Zhang*[1], *Long Nguyen*[1], *Spyros Matsoukas*[1], *Xinhui Zhou*[2], *Nima Mesgarani*[2], *Karel Veselý*[3], *and Pavel Matějka*[3]

[1] Raytheon BBN Technologies, Cambridge, MA, USA
[2] University of Maryland, College Park, MD, USA
[3] Brno University of Technology, Brno, Czech Republic
tng@bbn.com, zxinhui@umd.edu, iveselyk@fit.vutbr.cz

Abstract

This paper describes the speech activity detection (SAD) system developed by the Patrol team for the first phase of the DARPA RATS (Robust Automatic Transcription of Speech) program, which seeks to advance state of the art detection capabilities on audio from highly degraded communication channels. We present two approaches to SAD, one based on Gaussian mixture models, and one based on multi-layer perceptrons. We show that significant gains in SAD accuracy can be obtained by careful design of acoustic front end, feature normalization, incorporation of long span features via data-driven dimensionality reducing transforms, and channel dependent modeling. We also present a novel technique for normalizing detection scores from different systems for the purpose of system combination.
Index Terms: speech activity detection, noisy speech

1. Introduction

Speech Activity Detection (SAD) is the task of detecting when human speech occurs in an audio signal. SAD is an important component of most speech processing applications. For example, silence audio signal can be avoided from being coded and transmitted in telecommunication to save computation and bandwidth. SAD can also be a handy tool for an analyst who looks for speech regions in a long audio signal in which the majority of data is non-speech.

Typically, SAD is not evaluated on its own, but rather as part of a more complex speech processing pipeline such as speaker or language recognition. Recently, however, there has been a renewed interest in evaluating SAD as a standalone task, thanks to the DARPA RATS (Robust Automatic Transcription of Speech) project, which is concerned with advancing the state of the art in processing of speech from multiple languages carried over degraded radio communication channels.

Several techniques have been proposed for SAD, including: energy-based thresholding; frame-level speech/non-speech classification based on multi-layer perceptrons (MLP) (e.g., [1]) or Gaussian mixture models (GMM); speech/non-speech hidden Markov models (HMMs) (e.g., [2]); phoneme recognition based on MLPs [3] or HMMs [4]; segmental models using support vector machines [5].

In this paper, we describe two SAD systems developed by the Patrol team for the first phase of the DARPA RATS evaluation. Section 2 gives an overview of the data and evaluation protocol. In Section 3, we present the first system, developed by BBN, which is based on Gaussian mixture models (GMMs) and makes use of noise-robust cortical features provided by University of Maryland. The second system, developed by Brno University of Technology (BUT), is based on MLPs and is described in Section 4. A novel technique for combining the two SAD systems is presented in Section 5. The paper concludes with a discussion of future work in Section 6.

This paper is based upon work supported by the DARPA RATS Program. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. Approved for Public Release, Distribution Unlimited.

2. Data and Scoring

2.1. Training and Test Data

The Linguistic Data Consortium (LDC) provides the training and test data for the RATS participants. The audio recordings annotated were selected from existing speech corpora, such as the Fisher English and Arabic Levantine conversational telephone speech (CTS) training collections, as well as new collections specific for RATS. The latter includes telephone conversations in Arabic Levantine, Pashto, and Urdu.

These recordings were retransmitted through 8 different communication channels, labeled by the letters A through H [6]. A "push-to-talk" (PPT) transmission protocol was used in all of the channels except G. PPT states produce some regions where two or more non-transmission (NT) segments may occur. Speech (S), non-speech (NS), and NT regions are marked in the LDC-provided annotations. Note, however, that NT annotations are automatic and therefore not always accurate.

During the development period for the phase one evaluation, LDC delivered three incremental data releases for training and test. The 1st release contains about 655 hours audio data for training, the 2nd 725 hours, and the 3rd 517 hours. Only the 1st release was used as the training set throughout most of the development experiments reported in this paper, as the other two releases were made available just prior to the evaluation.

We make use of three development sets. The official dev1 consists of 11 hours, and was selected by SAIC (the RATS Evaluation Team) from the 1st and 2nd incremental releases. dev1_v2 consists of 71 hours, which is the development data defined in the 1st incremental release. dev1_p3 consists of 25 hours and was randomly selected by BBN from the 3rd incremental release training data. The annotated NT regions were excluded from the definitions for these development sets. The dev1_v2 and official dev1 sets were used as the main test sets during the system development, while dev1_p3 was only used after adding the 2nd and 3rd incremental releases to training.

2.2. Scoring

Our SAD systems output the start and end time of the hypothesized speech segments. These hypotheses are scored against a reference of manually annotated speech start/end times of the test audio. In the RATS program, SAD scoring is carried out using software provided by SAIC. The software computes the following two types of errors: (a) False Rejection (FR), i.e., misclassification of speech as non-speech, and (b) False Alarm (FA), i.e., misclassification of non-speech as speech.

Forgiveness collars of 500ms and 200ms are applied to the non-speech and speech sides, respectively, of each annotated non-speech/speech boundary in the reference. In other words, a region of 700ms centered at each of the annotated boundaries is excluded from the scoring. The remaining speech/non-speech regions are used to compute the FR and FA rates as follows: $P_{FR} = \frac{D_{FR}}{D_S}$ and $P_{FA} = \frac{D_{FA}}{D_{NS}}$, where D_{FR} is the total duration of the falsely rejected speech, D_S is the total duration of scored speech, D_{FA} is the total duration of falsely accepted non-speech, and D_{NS} is the total duration of scored non-speech in the audio. In this work, systems are evaluated in terms of Equal Error Rate (EER), which is the operating point at which P_{FR} is equal to P_{FA}.[1]

3. BBN SAD System

3.1. Acoustic Features

The baseline BBN SAD system makes use of standard Perceptual Linear Prediction (PLP) [7] front-ends. The audio is analyzed with a 25 ms sliding window and a 10 ms shift, extracting 14 cepstral coefficients and normalized energy from each frame. The cepstral coefficients are then normalized to zero mean and identity covariance on an audio file basis. These features, along with their first and second time derivatives, form a 45-dimensional feature vector that is subsequently used to train our baseline classifiers.

3.2. Model Estimation

The SAD system employed in this study is a 2-class classifier. Unless stated otherwise, each of the speech and non-speech classes is modeled with a 512-component diagonal covariance Gaussian mixture model (GMM). The GMM was initialized by running the K-Means algorithm with binary splitting on a random subset of the training data. The GMM was then refined with Maximum Likelihood (ML) estimation by running a few iterations of the Expectation-Maximization (EM) algorithm on all of the data.

3.3. Segmentation

GMM-based SAD is essentially a frame-level classification problem. The simplest solution is to compare the log likelihood ratio (LLR) between speech and non-speech models for each frame to a threshold, and make a decision for each frame independently. However, since the LLR can be very noisy, this simple method will unavoidably lead to high segmentation error. In our systems, we applied smoothing for better performance.

We tried two algorithms. In the first approach, likelihood smoothing, we replaced the LLR of a frame with the averaged LLR of multiple frames within a window before we compared it to a threshold. In the second method, median filter smoothing, we first made frame-level decisions by comparing LLR to a threshold and then applied a median filter to the decisions, which are binary numbers. In both algorithms the threshold and the window size were tuned in order to minimize the sum of the false alarm and false reject scores. We found that the likelihood smoothing method works consistently better. The optimal window size was determined to be around 81 frames in both cases.

[1]This operating point is typically obtained by adjusting a certain parameter of the SAD system (e.g., likelihood ratio threshold) on the test set.

Proj. Scheme	Context	official dev1	dev1_v2
Derivatives	7	3.50	4.69
HLDA	11	2.91	3.90
	15	2.74	3.69
	21	2.93	3.60
	31	3.05	3.86

Table 1: EER (%) for using different feature projection schemes and varying context

3.4. Long Span Features

Besides computing derivatives of PLP features, we experimented with frame concatenation, in which the base energy and cepstra are concatenated across consecutive frames to form long-span features that capture long-term information.

The long span features described above cannot be modeled directly due to their high dimensionality, and so we investigated options for dimensionality reduction. Among the standard techniques, both Linear Discriminant Analysis (LDA) and Heteroscedastic Discriminant Analysis (HDA) [8] are not suitable for our 2-way classification problem because they rely on the rank-1 between-class scatter matrix and hence can only project down to a single dimension.

We therefore turned to Heteroscedastic LDA (HLDA) [9], which attempts to find the maximum-likelihood full-space linear transformation based on a model that consists of class-specific diagonal covariance GMMs for the upper p (useful) dimensions of the transformed feature vectors, and a single diagonal covariance gaussian for the lower $n - p$ (nuisance) dimensions. Because of this joint modeling of useful and nuisance dimensions, HLDA can produce effective p-dimensional feature projections, with $p > 1$, even in the 2-class SAD problem. Note, however, that HLDA, being a maximum likelihood technique, does not actually find the optimal projection from a classification perspective. In fact, it can produce degenerate results when the input features contain dimensions that are linearly dependent, although we did not observe this problem in the experiments reported in this paper.

As shown in Table 1, besides conventional derivatives we have experimented with long span features of different context spans. The results show that long span features with 15-frame concatenation provided the best SAD EER. It is also shown that about 21% relative reduction in EER was obtained by using long span features with HLDA as compared to using derivatives. In all HLDA experiments, the output dimensionality was set to 45 so as to have a fair comparison to the derivatives baseline.

Cortical features are high-dimensional, multiscale spectro-temporal modulation features that are extracted from a window of 0.5 seconds and have been shown to be very robust to noise on a speech detection task [5]. To make the features tractable for use with a GMM based acoustic modeling, the dimensionality is reduced to 140 using tensor principal component analysis

Acoustic Features	official dev1	dev1_v2
15-frame PLP	2.74	3.69
+Cortical Features	2.45	2.72

Table 2: EER (%) for using cortical features

Covariance	#Components	official dev1	dev1_v2
diagonal	512	2.45	2.72
	2048	2.19	2.37
full	256	1.98	2.29

Table 3: EER (%) for using diagonal or full Covariance GMM

(PCA) based on higher-order singular-value decomposition, as described in [5].

To incorporate the cortical features into our SAD system, we append the 140-dimensional cortical features to the 225-dimensional long-span features (15-frame PLPs). HLDA is then used to reduce the dimensionality of the resulting features from 365 to 45. As shown in Table 2, we obtained 11% to 26% relative reduction in EER from combining the cortical features with the long span PLP features[2]. We therefore use cortical features in all experiments presented in the rest of this paper.

3.5. Model Complexity

Although HLDA is used to reduce the dimensionality of the acoustic features, the resulting features may not be fully uncorrelated. Therefore, the use of full covariances in GMM should help by capturing the correlation among the features in different dimensions. As shown in Table 3, we obtained 16% to 19% relative reduction in EER by using full covariance GMMs with 256 components when compared to 512-component diagonal GMMs. The improvement is reduced to 3% - 10% relative when compared to a 2048-component diagonal GMM system.

Given that the audio data is retransmitted through different transmitters/receivers with varying acoustic characteristics, we investigated channel-dependent models. As shown in Table 4, by using channel-dependent GMMs, we obtained 8% to 13% relative reduction EER. Automatic channel classification is done on each audio file to determine which channel-dependent GMMs are to be used for SAD. The classification is done as follows:

$$L = \arg\max_{c \in \{A-H\}} \prod_{t=1}^{T} (S_c(o_t) + N_c(o_t))$$

where L is the automatic channel classification label, T total number of frames for the audio file, o_t acoustic feature at time t, and $S_c(.)$ and $N_c(.)$ are the likelihoods for the speech and non-speech GMMs for channel c, respectively. The channel classification error is 0 for both dev1_v2 and official_dev1, and 1.9% for dev1_p3. Channel-dependent models are used in the rest of the paper.

3.6. Adding New Training Data

We re-investigated the effect of different context lengths for PLP frame concatenation after incorporating cortical features and channel-dependent models. As shown in Table 5, slightly better results are obtained with 31-frame concatenation; it is therefore used in our final evaluation system and the experiments reported in the rest of this paper.

[2]The cortical features by themselves perform about equally to long-span PLP features.

Chn-Dep. Model	official dev1	dev1_v2
No	1.98	2.29
Yes	1.82	2.00

Table 4: EER (%) for using channel independent/dependent models

Context for PLP	official dev1	dev1_v2
15	1.82	2.00
31	1.79	2.00

Table 5: EER (%) for using cortical features with different PLP context length

LDC made the 2nd and 3rd incremental releases available shortly before the RATS phase 1 evaluation. As shown in Table 6, by adding these new releases into the training, no significant change for official dev1 and dev1_v2 is observed, while 37% relative reduction in EER is obtained on dev1_p3. This is due to the fact that dev1_p3 is from the 3rd release which is different from the the other two data releases.

3.7. Speech Padding

We discovered that small regions of speech are consistently missed right after non-transmission (NT) regions in channel F. The NT regions in channel F are different from those in other channels, in that they exhibit impulses that take some time to decay to zero, affecting the neighboring speech quite significantly. To alleviate this problem, we extended the boundaries of each detected speech segment by 0.1s. As shown in Table 7, by padding the speech segments, we obtained 12% to 27% relative reduction in EER for channel F, and 7% to 9% relative in overall EER.

4. BUT SAD System

The BUT SAD system uses a multi-layer perceptron (MLP) to map long-temporal spectral features to speech/non-speech posterior probabilities. These log-posteriors are then used as emission probabilities in two ergodic HMMs, the speech/non-speech models. The final segmentation is obtained by a Viterbi decoder, which finds a smooth path through the posteriors. More details about this system are provided below.

Features and MLP configuration. The acoustic features are 15 log-mel-filterbank outputs, with 61-frame context. Each band is scaled by a Hamming window and reduced by discrete cosine transform (DCT) to 31 dimensions, forming a vector of 465-dimensional features per frame (15 bands times 31 DCT coefficients). The MLP has two hidden layers, each with 200 sigmoidal neurons. Different configurations were explored for the output layer, and the best results were obtained by having a speech/non-speech output for each of the 9 channels (channels A-H plus the source channel -

Training Data	official dev1	dev1_v2	dev1_p3
1st release	1.79	2.00	2.34
+2nd & 3rd releases	1.71	2.06	1.48

Table 6: EER (%) for adding new training data

	System	official dev1	dev1_v2
channel F	baseline	2.28	2.93
	+padding	1.66	2.58
overall	baseline	1.71	2.06
	+padding	1.56	1.92

Table 7: Equal Error Rate for before and after speech padding

Normalization	Chn. A excluded from training	Chn. A included in training
None	17.30	2.95
CMN (1-pass)	5.48	2.90
CMVN (2-pass)	3.79	2.71

Table 8: EER (%) on dev1_v2 channel A, showing the effect of 2-pass feature normalization. CMN: global mean normalization, single-pass SAD. CMVN: mean and variance normalization based on two-pass SAD

prior to retransmission). The 18-dimensional output posteriors sum up to 1 due to the softmax function, thus each node can be interpreted as joint probability of channel and speech/non-speech classes. This channel-dependent model outperformed the channel-independent model, since channels A-H have very different characteristics.

Segmentation. The MLP posteriors are post-processed by a Viterbi decoder with two models, speech and non-speech, each represented as an ergodic HMM with 9 states, where each state corresponds to an MLP output node. The HMM state loop transition probabilities are set to 0.5, and the rest of the probability mass is equally distributed to all other transitions. During decoding, the HMM can switch from one channel state to another, thus the most convenient channel is dynamically selected. An additional penalty is applied to transitions from speech states to non-speech states, in order to discourage the generation of speech or non-speech segments that are too short.

Two-pass feature normalization. The MLP SAD system, as described above, achieves an overall EER of 2.04 on the dev1_v2 set, which is very close to BBN's GMM SAD dev1_v2 result reported in Table 7. Both systems perform well on channels seen in the training data, but how well do they generalize on new channels? The system robustness can be improved significantly by using following two-pass feature normalization method: *Pass 1)* global mean subtraction is applied to the log-mel-filterbank outputs, and the speech regions are detected. *Pass 2)* the mean- and variance-normalization statistics are accumulated on speech regions from the first pass, a second pass SAD is performed on renormalized features. For both passes, the normalization is done on a per audio file basis.

The effect of this hierarchical feature normalization is shown in Table 8, which shows EER results on dev1_v2 channel A, with and without data from channel A in training.

5. System Combination

Neural Network is used in the BUT SAD system while GMM is used in the BBN system. Improvement should be obtained by combining the two systems as they are different in nature.

To combine the SAD outputs from the BBN and BUT SAD systems, we need to normalize each system's frame-level scores so that they can be comparable to each other. We achieve this by mapping the scores to their corresponding estimated $1 - pFA$ values. This non-parametric mapping is derived from the system scores on dev1_v2. For example, to normalize the BBN frame-level smoothed GMM likelihood ratios, we first sort them and then measure the *pFA* that results from using each distinct likelihood ratio as a detection threshold for speech/non-speech classification. That provides a mapping of that likelihood ratio to $1 - pFA$. The same normalization is applied to the frame posteriors from the BUT system.

System	official dev1	dev1_v2
BBN	1.56	1.92
BUT	2.09	2.04
BBN+BUT	1.42	1.82

Table 9: EER (%) for system combination

The normalized scores from the two systems are combined as follows: $S_{comb} = \alpha S_{BBN} + (1 - \alpha) S_{BUT}$
where S_{comb} is the combined score, $\alpha = 0.7$, S_{BBN} is the normalized score from the BBN system, and S_{BUT} is the normalized score from the BUT system.

As shown in Table 9, through system combination, we obtained 5% to 9% relative reduction in EER as compared to the BBN system.

6. Conclusions

We have presented the SAD system developed by the Patrol team for the DARPA RATS phase 1 evaluation. The system achieves high accuracy on audio from noisy radio communication channels, due to its use of noise robust long-span temporal features, data-driven dimensionality reduction, channel dependent modeling, and system combination. Future work will explore more advanced acoustic models, including HMM-based self-organized units and model-based noise-compensation methods.

7. References

[1] J. Dines, J. Vepa, and T. Hain, "The segmentation of multi-channel meeting recording for automatic speech recognition," in *Proc. of ICSLP*, Pittsburgh, USA, September 2006.

[2] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, "Robust speech recognition in noisy environments: The 2001 IBM spine evaluation system," in *Proc. of ICASSP*, 2002.

[3] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical sturctures of neural networks for phoneme recognition," in *Proc. of ICASSP*, Toulouse, France, May 2006.

[4] D. Liu and F. Kubala, "Fast speaker change detection for broadcast news transcription and indexing," in *Proc. of Eurospeech*, 1999.

[5] N. Mesgarani, M. Slaney, and S. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, pp. 920–930, May 2006.

[6] K. Walker and S. Strassel, "The rats radio traffic collection system," in *Proc. of ISCA Odyssey*, 2012.

[7] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87(4), pp. 1738–1752, 1990.

[8] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *Proc. of ICASSP*, vol. II, June 2000, pp. 129–132.

[9] N. Kumar and A. Andreou, "A generalization of linear discriminant analysis in maximum likelihood framework," Johns Hopkins University, Tech. Rep., 1996.