Acoustic and Data-driven Features for Robust Speech Activity Detection

Samuel Thomas[1], *Sri Harish Mallidi*[1], *Thomas Janu*[1], *Hynek Hermansky*[1],
Nima Mesgarani[2], *Xinhui Zhou*[2], *Shihab Shamma*[2],
Tim Ng[3], *Bing Zhang*[3], *Long Nguyen*[3] *and Spyros Matsoukas*[3]

[1] The Johns Hopkins University, Baltimore, MD, USA {samuel,mallidi,tjanu,hynek}@jhu.edu
[2] University of Maryland, College Park, MD, USA {mnima,zxihui,sas}@umd.edu
[3] Raytheon BBN Technologies, Cambridge, MA, USA {tng,bzhang,ln,smatsouk}@bbn.com

Abstract

In this paper we evaluate different features for speech activity detection (SAD). Several signal processing techniques are used to derive acoustic features that capture attributes of speech useful in differentiating speech segments in noise. The acoustic features include short-term spectral features, long-term modulation features both derived using Frequency Domain Linear Prediction (FDLP), and joint spectro-temporal features extracted using 2D filters on a cortical representation of speech. Posteriors of speech and non-speech from a trained multi-layer perceptron are also used as data-driven features for this task. These feature extraction techniques form part of an elaborate feature extraction front-end where information spanning several hundreds of milliseconds of the signal are used along with heteroscedastic linear discriminant analysis for dimensionality reduction. Processed feature outputs from the proposed front-end are used to train SAD systems based on Gaussian mixture models for processing of speech from multiple languages transmitted over noisy radio communication channels under the ongoing DARPA Robust Automatic Transcription of Speech (RATS) program. The proposed front-end performs significantly better than standard acoustic feature extraction techniques in these noisy conditions.

Index Terms: Speech Activity Detection, Features for SAD

1. Introduction

Speech activity detection (SAD) is the first step in most speech processing applications like speech recognition, speech coding and speaker verification. This module is an important component that helps subsequent processing blocks focus resources on the speech parts of the signal. In each of these applications, several approaches have been used to build reliable SAD modules. These techniques are usually variants of decision rules based on features from the audio signal like signal energy [1], pitch [2], zero crossing rate [3] or higher order statistics in the LPC residual domain [4]. Acoustic features have also been used to train multi-layer perceptrons (MLPs) [5] and hidden Markov models (HMMs) [6] to differentiate between speech and non-speech classes. All these approaches in essence focus on characteristic attributes of speech which differentiate it from other acoustic events that can appear in the signal.

The research presented in this paper was funded by the DARPA RATS program under D10PC20015. The views expressed are those of the author(s) 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.

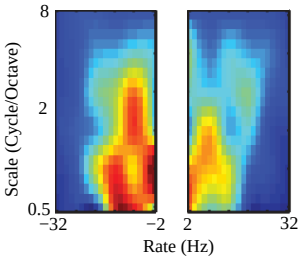

Figure 1: *Spectro-temporal modulations of speech derived from clean TIMIT utterances. Separate positive and negative temporal modulations are used to differentiate sweep directions (upward or downward).*

Speech is a sequence of consonants and vowels, non-harmonic and harmonic sounds with natural silences between them. This makes speech a complex signal with a broad range of spectro-temporal modulations (Fig. 1). Important temporal modulations of speech lie in the 0-20 Hz range, with a peak around 4 Hz. Spectral modulations, on the other hand, span a range between 0-6 cycle/octave. While pitch or voicing introduces modulations in the 2-6 cycle/octave range, modulations less than 2 cycle/octave reflect formant information. As with other pattern recognition tasks, an important step in SAD is to represent speech using features that capture these distinct properties while also being robust to distortions under various noisy conditions.

In this paper we investigate different kinds of acoustic features that capture information based on these modulation properties of speech. These features are generated using different signal processing techniques but can be broadly categorized by the kinds of modulations they capture as -

(a) Short-term spectral features extracted from power spectral estimates in short analysis windows (10-30 ms) of the speech signal,

(b) Long-term modulation frequency components estimated in long analysis windows spanning few hundreds of milliseconds from sub-band envelopes of speech, and

(c) Joint spectro-temporal features derived using 2D selective filters tuned to different rate and scales of the input spectrogram.

In addition to these acoustic features, we also evaluate data-driven features derived using multi-layer perceptrons trained on large amounts of data. Posteriors of speech/non-speech classes

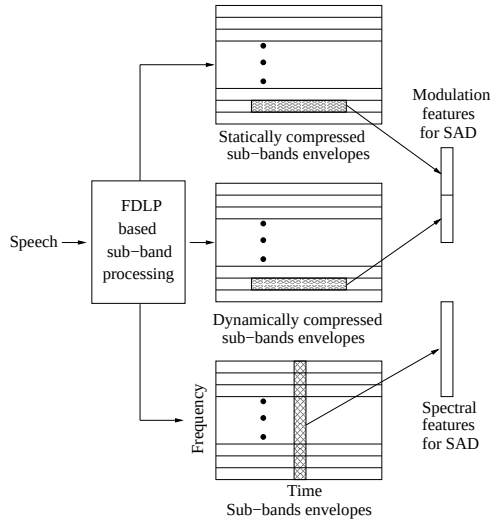

Figure 2: *Short-term spectral and long-term modulation features for SAD derived using FDLP*

estimated by these networks are used as features similar to acoustic features.

Each of these features is evaluated in terms of SAD on recordings from multiple languages, using a Gaussian mixture model (GMM) based back-end SAD system. The audio data for the DARPA RATS program is collected under both controlled and uncontrolled field conditions over highly degraded, weak and/or noisy communication channels making the SAD task very challenging [7]. The rest of the paper is organized as follows. Section 2 describes the different feature extraction techniques we employ. Section 3 talks about the GMM based speech/non-speech (S/NS) acoustic models we train on these features. These models are trained after the features have been pre-processed for better classification. S/NS scores are then smoothed before a threshold based decision module generates the final speech and non-speech timing intervals. Section 4 discusses our experiments and results. The paper concludes with a discussion in Section 5.

2. Features for SAD

We use two different acoustic processing techniques to derive features for SAD. The first technique uses an autoregressive (AR) model to representing the long-term amplitude modulations (AM) of speech. Short-term spectral features and long-term modulation frequency features are derived from this representation. The second approach uses a bank of modulation selective filters at the output of a computational auditory model. Joint spectro-temporal features are extracted from this representation.

2.1. Frequency Domain Linear Prediction (FDLP)

FDLP is an efficient technique for auto regressive (AR) modeling of temporal envelopes of a signal [8]. The magnitude response of the all pole filter approximates the Hilbert envelope of the signal in a manner similar to the approximation of the power spectrum of the signal using time domain linear prediction (TDLP). In this approach, we first apply the discrete cosine transform (DCT) on long segments of speech to obtain a real valued spectral representation of the signal. The DCT transform of the signal is decomposed using critical-band-sized windows. Linear prediction is performed on each sub-band DCT signal to obtain a parametric model of its temporal envelope. We compute a spectrogram of speech by stacking the individual sub-band temporal trajectories derived using FDLP.

Short-term spectral features are derived from sub-band temporal envelopes, by integrating the envelopes in short term frames (of the order of 25 ms with a shift of 10 ms) [9]. These short term sub-band energies are then converted into 15 cepstral features. To extract long-term modulation frequency features, we first compress the sub-band temporal envelopes statically using the logarithmic function and dynamically with an adaptation circuit consisting of five consecutive nonlinear adaptation loops. The compressed temporal envelopes are then transformed using the Discrete Cosine Transform (DCT) in long term windows (200 ms long, with a shift of 10 ms). We use 10 modulation frequency components from each cosine transform, yielding modulation spectrum in the 0-35 Hz range with a resolution of 5 Hz [9]. Fig. 2 is a schematic representation of how we derive short and long-term features using FDLP.

2.2. Cortical Representations of Speech

Cortical representations of speech are derived using a two stage computational auditory model based on neurophysiological investigations of various stages of the human auditory system [10]. The first stage which models the cochlear filter bank, hair cell and lateral inhibitory networks, transforms the acoustic signal into an auditory spectrogram representation. The second stage analyzes this spectrogram to estimate the content of its spectral (scale) and temporal (rate) modulations using a bank of 2D modulation selective filters. These filters mimic the behavior of neurons in the primary auditory cortex. Mathematically these filtering operations are equivalent to two-dimensional wavelet transforms of the auditory spectrogram, with wavelets resembling 2D Gabor functions. For our experiments we use a bank of directional selective filters tuned to different rates and scales, with both symmetric and asymmetric shapes.

The output of the auditory model is a multidimensional array with modulation components presented along four dimensions of time, frequency, rate, and scale. For our current experiments, the time axis of absolute values is averaged over a 250 ms sliding time window resulting in a three mode tensor for each time window. We finally use a tensor-PCA dimensionality reduction technique [11] to reduce the feature dimensionality for each time window independently in each subspace to a total of 140 dimensions.

2.3. Data-driven Features for SAD

The acoustic features described above explicitly incorporate information present in different kinds of spectro-temporal modulations to differentiate between speech and other acoustic events. In a different approach, we train MLPs on large amounts of data to differentiate between two classes - speech versus non-speech. Instead of using these models directly to produce S/NS decisions, the trained models are used as a data-driven front-end to derive features for SAD.

The proposed front-end has a multi-stream architecture with several levels of MLPs. The motivation behind this multi-stream front-end is to use parallel streams of data that carry complementary or redundant information while at the same time degrading differently in noisy environments. We use 5 feature streams that include sub-band energies corresponding

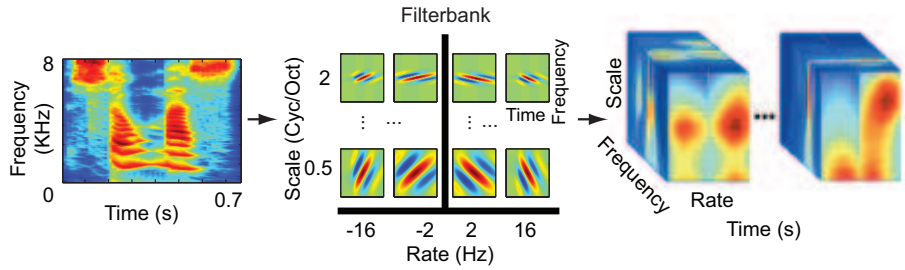

Figure 3: *Cortical multi-scale representation of speech - The auditory spectrogram is analyzed by a bank of spectro-temporal modulation selective filters to produce spectro-temporal representations of speech. The filters have different temporal and spectral selectivity, with different symmetry and up-down sweep.*

to different frequency ranges of the spectrum along with two kinds of temporal modulations.

The MLPs are trained on close to 660 hours of audio from the RATS development corpus [14] using LDC provided S/NS annotations. Sub-band energies of speech are derived using the FDLP technique with 45 bark scale filters. Separate MLPs are trained on 3 sets of sub-band energies corresponding to high, mid and low frequency ranges. Each 15-dimensional feature vector is also appended with contextual information from 50 adjacent left and right frames. Fast and slow temporal modulations correspond to the the first and last 5 long-term modulation frequency features also derived using FDLP respectively. Outputs from these 5 sub-systems are then fused by a merger MLP at the second level to derive the final S/NS posterior features. These features are derived from the pre-softmax outputs of the final layer.

3. Acoustic Models for SAD

Speech activity detection is carried out on the proposed features in three main steps. In the first step, the input frame-level features are projected to a lower-dimensional space. The reduced features are then used to compute per-frame log likelihood scores with respect to speech and non-speech classes, each class being represented separately by a GMM. The frame-level log likelihood scores are mapped to S/NS classification decisions to produce final segmentation outputs in the last step. Each of these steps are described in detail in the next sections.

3.1. Feature projection to lower dimensions

With both the FDLP and cortical feature extraction techniques producing high dimensional feature vectors per frame of speech, it is necessary to apply a dimensionality reduction scheme to ensure robust acoustic modeling in subsequent steps. Linear Discriminant Analysis (LDA), although very popular in many classification tasks, is not very appealing for speech/non-speech classification since it can project to only one dimension. This is due to its use of a between-class scatter matrix of rank N-1, where N is the number of classes (in our case, N=2). Heteroscedastic linear discriminant analysis (HLDA) [12], on the other hand, is better suited for our purposes, as its output dimensionality is not constrained by the number of classes and can accommodate Gaussian Mixture Models. Since HLDA is a maximum likelihood method, it is not guaranteed to find an optimal projection in terms of increasing class separability. However in practice we find consistent gains in SAD accuracy from combining various types of features using this technique.

3.2. Speech/Non-speech likelihood computation

During training, we pool all the feature vectors generated by HLDA into two classes, speech and non-speech, and estimate a GMM for each class using standard ML re-estimation methods. We do this starting with a single Gaussian component for each class. The means of the single component are then randomly perturbed to increase the number of components to two. This is followed by few iterations of the expectation-maximization (EM) algorithm to re-estimate the new means and variances. The interleaved Gaussian splitting and EM training procedure is continued until we reach the desired number of mixture components.

During testing, for each frame of speech we use its HLDA representation to calculate log likelihood scores with respect to each of the two GMMs models. The two scores are then subtracted to form a per-frame S/NS log likelihood ratio.

3.3. Speech/Non-Speech classification

The final classification is done by computing the average per-frame log likelihood ratio, based on a sliding window of 81 frames. The resulting scores are then compared against a fixed threshold. Frames with scores above the threshold are classified as speech, and the rest as non-speech.

4. Experiments

The features described in Section 2 are evaluated in terms of speech activity detection (SAD) accuracy on noisy radio communications audio provided by the Linguistic Data Consortium (LDC) for the DARPA RATS program [14]. Most of the RATS data released for SAD were obtained by retransmitting existing audio collections - such as the DARPA EARS Levantine/English Fisher conversational telephone speech (CTS) corpus - over eight radio channels, labeled A through H. covering a wide range of radio channel transmission effects [7].

The development corpus used in our experiments consisted of 11 hours of audio from the Arabic Levantine and English Fisher CTS corpus, retransmitted over the eight channels. The training corpus consisted of 73 hours of audio (62 hours from the Fisher collection, and 11 from new RATS collection). Although the entire data was also retransmitted over eight channels, we selected a channel at random for each audio file to reduce the turnaround time for our experiments. In the initial release of the above LDC RATS corpora several audio files from channel F were unusable, so we excluded all data from that channel from both training and development.

We trained SAD models, as described in Section 3, for

Features	Dimensionality			Equal Error Rate (%) on different channels							
	#Dims. /Frame	#Frame Context	Total #Dims.	A	B	C	D	E	G	H	All
PLP	15	31	465	3.55	3.00	5.03	2.51	2.75	3.48	2.34	3.34
FDLPS	15	31	465	3.42	3.10	4.46	2.42	2.78	3.40	2.29	3.20
FDLPM	340	1	340	3.88	3.80	4.12	3.26	3.52	3.60	2.51	4.15
MLP	2	31	62	3.05	2.96	3.76	2.20	2.71	3.35	2.10	3.17
CORT	140	1	140	3.81	3.33	4.02	2.46	3.41	3.46	2.20	3.27
PLP+MLP	17	31	527	3.10	2.84	3.20	2.25	2.63	2.96	2.07	2.84
FDLPS+MLP	17	31	527	3.15	2.94	3.04	2.17	2.67	2.89	1.93	2.82
FDLPM+MLP	402	1	402	3.02	2.90	3.73	2.26	2.84	2.42	1.89	2.88
PLP+CORT	605	1	605	3.35	3.08	3.21	2.29	2.71	2.62	1.97	2.85

Table 1: Equal Error Rate (%) on different channels using different acoustic features and combinations

each type of features, as well as for certain feature combinations. In each case, HLDA was used to reduce dimensionality prior to GMM training. Table 1 shows the dimensionality of the original space, prior to the application of HLDA, for each feature type used. We explicitly use a context of 31 frames for short-term features. In all cases, the output dimensionality of HLDA was set to 45. A single Gaussian was used to represent each of the two classes (speech, non-speech) during HLDA estimation. After the dimensionality reduction, we trained 512-component GMMs for S/NS classification. The number of contextual frames, HLDA dimensionality, and number of GMM components were optimized using separate experiments [13].

The derived SAD models were evaluated on the development set in terms of equal error rate (EER%), which is the operating point at which the falsely rejected speech rate (probability of missed speech) is equal to the falsely accepted non-speech rate (probability of false alarm). The results are shown in Table 1 for conventional features (PLP), short-term features derived using FDLP (FDLPS), long-term modulation features (FDLPM), joint spectro-temporal cortical features (CORT), and data-driven features (MLP). Although each of the feature sets have varying performance in each of the individual noisy channels, they are comparable to each other in terms of overall SAD performance. In a second set of experiments we combine features which capture various kinds of information about speech. We observe close to 15% relative improvement for two kinds of feature combinations - combination of acoustic and data-driven features (for example FDLPS+MLP), and the combination of different acoustic features (PLP+CORT). We draw the following conclusions from these experiments -

1. Contextual information needs to be captured for good S/NS discrimination. While we provide this explicitly in short-term features (31 frames of speech), long-term modulation features implicitly capture this information.
2. It is useful to use dimensionality reducing techniques, such as HLDA, to project high-dimensional features to lower dimensions before modeling.
3. MLP based models, which are traditionally used to directly produce S/NS decisions, can be used as data-driven front-ends to produce complementary data-driven features.
4. Different acoustic features capture complementary attributes leading to further performance improvements when combined.

5. Conclusions

We have evaluated different kinds of acoustic and data-driven features in terms of SAD on a very challenging audio corpus. The proposed features are first pre-processed by appending sufficient context information before projecting them to lower dimensions. These features provide comparable performances when used individually. Significant improvements are obtained when features which capture different spectro-temporal properties and data-driven attributes are combined together.

6. References

[1] K. Woo, T. Yang, K. Park, C. Lee, "Robust Voice Activity Detection Algorithm for Estimating Noise Spectrum", IEEE Electronics Letters, 2000.

[2] R. Chengalvarayan, "Robust Energy Normalization using Speech/non-speech Discriminator for German Connected Digit Recognition", In Proc. of ISCA Eurospeech, 1999.

[3] ITU-T Recommendation G.729-Annex B.,"A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70", 1996.

[4] E. Nemer, R. Goubran, S. Mahmoud, "Robust Voice Activity Detection using Higher-order Statistics in the LPC Residual Domain", IEEE Trans. Speech and Audio Processing, 2001.

[5] J. Dines, J. Vepa, and T. Hain, "The Segmentation of Multichannel Meeting Recording for Automatic Speech Recognition", In Proc. of ISCA ICSLP, 2006.

[6] 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 IEEE ICASSP, 2002.

[7] K. Walker and S. Strassel, "The RATS Radio Traffic Collection System", In Proc. of ISCA Odyssey, 2012.

[8] R. Kumerasan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications", JASA, 1999.

[9] S. Thomas, S. Ganapathy and H. Hermansky, "Phoneme Recognition using Spectral Envelope and Modulation Frequency Features", In Proc. of IEEE ICASSP, 2009.

[10] T. Chi, P. Ru, and S. Shamma, "Multiresolution spectrotemporal analysis of complex sounds", JASA, 2005.

[11] N. Mesgarani, M. Slaney, and S. Shamma, "Discrimination of Speech from Non-speech based on Multiscale Spectro-temporal Modulations", IEEE Trans. on Audio, Speech, and Language Processing, 2006.

[12] N. Kumar and A.G. Andreou, "A Generalization of Linear Discriminant Analysis in Maximum Likelihood Framework", Johns Hopkins University Technical Report, 1996.

[13] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Vesely and P. Matejka, "Developing a Speech Activity Detection System for the DARPA RATS Program", In Proc. of ISCA Interspeech 2012.

[14] X. Ma, D. Graff, K. Walter, "RATS - First Incremental SAD Audio Delivery, LDC2011E86", 2011.