All for One: Feature Combination for Highly Channel-Degraded Speech Activity Detection

Martin Graciarena[1], *Abeer Alwan*[4], *Dan Ellis*[5,2], *Horacio Franco*[1], *Luciana Ferrer*[1], *John H.L. Hansen*[3], *Adam Janin*[2], *Byung-Suk Lee*[5], *Yun Lei*[1], *Vikramjit Mitra*[1], *Nelson Morgan*[2], *Seyed Omid Sadjadi*[3],*TJ Tsai*[2], *Nicolas Scheffer*[1], *Lee Ngee Tan*[4], *Benjamin Williams*[1]

[1]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA
[2]International Computer Science Institute (ICSI), Berkeley, CA, USA
[3]Center for Robust Speech Systems (CRSS), U.T. Dallas, Richardson, TX, USA
[4]Speech Processing and Auditory Perception Lab., Univ. of California, Los Angeles, CA, USA
[5]LabROSA, Columbia University, NY, USA

martin@speech.sri.com

Abstract

Speech activity detection (SAD) on channel transmissions is a critical preprocessing task for speech, speaker and language recognition or for further human analysis. This paper presents a feature combination approach to improve SAD on highly channel degraded speech as part of the Defense Advanced Research Projects Agency's (DARPA) Robust Automatic Transcription of Speech (RATS) program. The key contribution is the feature combination exploration of different novel SAD features based on pitch and spectro-temporal processing and the standard Mel Frequency Cepstral Coefficients (MFCC) acoustic feature. The SAD features are: (1) a GABOR feature representation, followed by a multilayer perceptron (MLP); (2) a feature that combines multiple voicing features and spectral flux measures (Combo); (3) a feature based on subband autocorrelation (SAcC) and MLP postprocessing and (4) a multiband comb-filter F0 (MBCombF0) voicing measure. We present single, pairwise and all feature combinations, show high error reductions from pairwise feature level combination over the MFCC baseline and show that the best performance is achieved by the combination of all features.

Index Terms: speech detection, channel-degraded speech, robust voicing features

1. Introduction

Speech activity detection (SAD) on noisy channel transmissions is a critical preprocessing task for speech, speaker and language recognition or for further human analysis. SAD tackles the problem of separation between speech and background noises and channel distortions such as spurious tones, etc.

Numerous methods have been proposed for speech detection. Some simple methods are based on comparing the frame energy, zero crossing rate, periodicity measure, or spectral entropy with a detection threshold to make the speech/nonspeech decision. More advanced methods include long-term spectral divergence measure [1, 2], amplitude probability distribution [3], and low-variance spectrum estimation [4].

This paper presents a feature combination approach to improve speech detection performance. The main motivation is to improve the baseline acoustic feature performance with different novel pitch and spectro-temporal processing features by exploring the complementary information from the presence of a pitch structure or from a different spectro-temporal representation. We combine an MFCC acoustic feature with four speech activity detection features: (1) a GABOR feature representation followed by a multilayer perceptron that produces a speech confidence measure; (2) a Combo feature that combines multiple voicing features and a spectral flow measure; (3) a feature based on subband autocorrelation (SAcC) and MLP postprocessing and (4) a multiband comb-filter F0 (MBCombF0) voicing measure estimated from a multiple filterbank representation. We present speech detection results for highly channel-degraded speech data collected as part of the DARPA RATS program. We show gains from feature level combination, resulting in significant error reductions over the MFCC baseline.

The RATS program aims at the development of robust speech processing techniques for highly degraded transmission channel data, specifically for SAD, speaker and language identification, and keyword spotting. The data was collected by the Linguistic Data Consortium (LDC) by retransmitting conversational telephone speech (CTS) through eight different communication channels [12] using multiple signal transmitters/transceivers, listening station receivers and signal collection and digitization apparatus. The RATS rebroadcasted data is unique in that it contains a wide array of real transmission distortions such as: band limitation, strong channel noises, nonlinear speech distortions (e.g., clipping), frequency shifts, high energy non transmission bursts, etc.

The proposed SAD system is based on a smoothed log likelihood ratio between a speech Gaussian mixture model (GMM) and a background GMM with a multiple feature combination input and Discrete Cosine Transform (DCT) long range modeling. The SAD model is similar to the one presented by Ng *et. al.* [11], however we used different model and likelihood smoothing parameters. The long span feature and dimensionality reduction technique differ from the one from Ng; instead of Heterosedastic linear discrimination (HLDA) we used the DCT technique. In Ng's paper the DCT component is used but on the MLP SAD subcomponent. Finally, the types of features differ as well. In our work we employ the standard acoustic features (i.e., MFCC) as well as four different types of features ranging from spectro-temporal to voicing derived features, whereas Ng's paper a combination of standard acoustic features and cortical based features.

2. Features Description

This section describes specific aspects of each of the four SAD-specific features: GABOR, Combo, SAcC and MBCombF0.

2.1. GABOR Feature

The GABOR with MLP feature is computed by processing a Mel spectrogram by 59 real-valued spectro-temporal filters covering a range of temporal and spectral frequencies. Each of these filters can be viewed as correlating the time-frequency plane with a particular ripple in time and frequency. Because some of these filters yield very similar outputs for neighboring spectral channels, only a subset of 449 GABOR features is used for each time frame. As the final preprocessing step, mean and variance normalization of the features over the training set is performed. GABOR features are described in [5].

Next, a MLP is trained to predict speech/nonspeech labels given 9 frames of the 449 GABOR features, or 4,041 inputs. The MLP uses 300 hidden units and 2 output units. The size of the hidden layer is chosen so that each MLP parameter has approximately 20 training data points. Although the MLP is trained with a softmax nonlinearity at the output, during feature generation the values used are the linear outputs before the nonlinearity. The resulting 2 outputs are then mean and variance normalized per file, and used as the input features to the classification backend.

2.2. Combo Feature

This section describes the procedure for extracting a 1-dimensional feature vector that has been shown to possess great potential for speech/non-speech discrimination in harsh acoustic noise environments [6]. This “combo” feature is efficiently obtained from a linear combination of four voicing measures as well as a perceptual spectral flux (SF) measure. The perceptual SF and periodicity are extracted in the frequency domain, whereas the harmonicity, clarity, and prediction gain are time domain features.

The combo feature includes the following: (1) Harmonicity (also known as harmonics-to-noise ratio) is defined as the relative height of the maximum autocorrelation peak in the plausible pitch range. (2) Clarity is the relative depth of the minimum average magnitude difference function (AMDF) valley in the plausible pitch range. Computing the AMDF from its exact definition is costly; however, it has been shown [7] that the AMDF can be derived (analytically) from the autocorrelation. (3) Prediction gain is defined as the ratio of the signal energy to the linear prediction (LP) residual signal energy. (4) Periodicity, in the short-time Fourier transform domain, is the maximum peak of the harmonic product spectrum (HPS) [8] in the plausible pitch range. (5) Perceptual SF measures the degree of variation in the subjective spectrum across time. In short-time frames, speech is a quasistationary and slowly varying signal, that is, its spectrum does not change rapidly from one frame to another.

After extracting the features, a 5-dimensional vector is formed by concatenating the voicing measures along with the perceptual SF. Each feature dimension f_i is normalized by its mean and variance over the entire waveform. The normalized 5-dimensional feature vectors are linearly mapped into a 1-dimensional feature space represented by the most significant eigenvector of the feature covariance matrix. This is realized through principal component analysis (PCA), and by retaining the dimension that corresponds to the largest eigenvalue. The 1-dimensional feature vector is further smoothed via a 3-point median filter and passed to the next stage for speech activity detector.

2.3. MBCombF0 Feature

The voicing feature is the estimated degree of voicing of each frame computed by the MBCombF0 algorithm, which is a modification of the correlogram-based F0 estimation algorithm described in [9]. The processing sequence of the MBCombF0 is the following. A frame length of 100 ms is used. First, the input signal is downsampled to 8 kHz and split into four subbands that cover 0 to 3.4 kHz. Each subband has a 1-kHz bandwidth and overlaps the adjacent filter by 0.2 kHz. Envelope extraction is then performed on each subband stream, followed by multichannel comb-filtering with comb filters of different interpeak frequencies.

Next, reliable comb-channels are selected individually for each subband using a 3-stage selection process. The first selection stage is based on the comb-channel's harmonic-to-subharmonic-energy ratio in the respective subband, those with a peak magnitude greater than one. In the second stage, comb-channels and their corresponding subharmonic channels (with an interpeak frequency that is half of the former) are retained if both are present in this initial selected set. In the final selection stage, channels whose maximum autocorrelation peak location (computed from their comb-filtered outputs) is close to their corresponding comb-filters' fundamental period are selected. A subband summary correlogram is then derived from the weighted average of selected energy-normalized autocorrelation functions. Finally, the four subband summary correlograms are combined using a subband reliability weighting scheme to form the multiband summary correlogram. The weighting of each subband depends on its maximum harmonic-to-subharmonic-energy ratio and the number of the subband summary correlogram whose maximum peak location is similar to its own. Time-smoothing is then applied to the multiband summary correlogram as described in [9], and the maximum peak magnitude of the resulting summary correlogram is the MBCombF0 voicing feature extracted.

2.4. SAcC Feature

The SAcC feature (for Subband Autocorrelation Classification) [10] is derived from our noise-robust pitch tracker. SAcC involves an MLP classifier trained on subband autocorrelation features to estimate, for each time frame, the posterior probability over a range of quantized pitch values, and one "no-pitch" output. We trained a RATS-specific MLP by using the consensus of conventional pitch trackers applied to the clean (source) signal to create a ground truth for each of the noisy (received) channels; we trained a single MLP for all channels. For this system, we used only the "no-pitch" posterior as a feature to indicate the absence of voiced speech in the signal frame.

2.5. Feature Figures

Figure 1 shows a plot of the channel-degraded waveform for channel A, spectrogram, labels, and the GABOR, Combo, SAcC and MBCombF0 feature outputs per frame. Rectangles superimposed on the waveform highlight the speech regions. Notice the highly channel-degraded waveform and low signal to noise ratio (SNR). The labels are 1 for speech, 0 for non-speech and -1 for no transmission (NT) regions. The NT regions are high energy white noise type of sounds interleaved between valid signal transmissions. GABOR features are much smoother with a long time span. Other SAD features are frame-based so they have a more dynamic behavior. However, they all achieve good detection of the speech regions.

Interestingly, the three voicing based features provide somewhat different outputs.

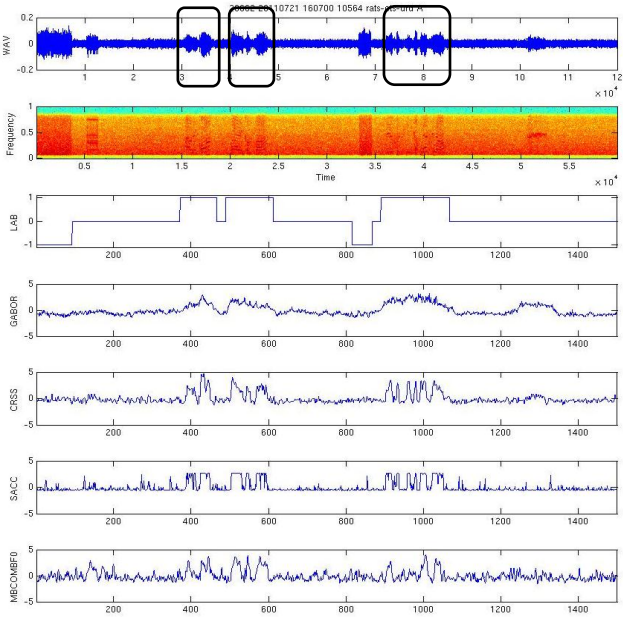

Figure 1: Waveform, spectrogram, ground truth speech and nonspeech labels, GABOR, Combo, SAcC and MBCombF0 features. Speech regions marked in black rectangles.

3. SAD Description

The SAD system is based on a frame-based smoothed log-likelihood ratio (LLR) setup. The LLR is computed between speech and nonspeech Gaussian mixture models (GMM). Then the LLR is smoothed with a multiple window median filter of length 51 frames. Finally, the speech regions are obtained from the smoothed LLR frames which are higher than a given threshold. No padding was used to artificially extend the speech regions.

Additionally, we used long range modeling using a 1-dimensional Discrete Cosine Transform (DCT). For each feature dimension we first created a window of multiple frames. Next, we computed the DCT transform and only preserved a subset of the initial DCT coefficients to obtain the desired number of features. This results in a low-dimensional representation of the feature modulation within the multi-frame window. We found that a 30 frame window was optimal for most features. We then concatenated the DCT dimensionality-reduced features for all dimensions and applied waveform level mean and variance normalization.

For most of the experiments we used 256-mixture GMMs with full covariance matrices for speech and nonspeech classes. We trained channel dependent models, therefore at test time we ended up with 16 models, 8 for speech and 8 for nonspeech. When testing SAD features we used 32-mixture GMMs with full covariance matrices due to their reduced feature dimension. During testing we obtained the LLR from the numerator obtained as the sum of the log probability of the speech models given the current feature and the denominator obtained from the sum of the log probability of the nonspeech models given the current feature. No channel selection is performed during testing.

4. Experiments

4.1. Data Description

This section discusses speech detection in RATS data. We present the results of each feature in isolation and then the feature level combination results.

The data used belongs to the LDC collections for the DARPA RATS program: LDC2011E86, LDC2011E99, and LDC2011E111. We trained models on the train subsets and tested on the Dev-1 and Dev-2 sets. These two development sets contain similar data but we found Dev-2 to contain speech at lower SNR. The data was annotated with speech and background labels. More details are presented in Walker and Strassel [12].

The audio data was retransmitted using a multilink transmission system designed and hosted at LDC. Eight combinations of analog transmitters and receivers were used covering a range of carrier frequencies, modes and bandwidths, from 1MHz amplitude modulation to 2.4GHz frequency modulation.

The audio material for retransmission was obtained from existing speech corpora such as Fisher English data, Levantine Arabic telephone data and RATS program specific collections, which included speech in several languages such as English, Pashto, Urdu, Levantine Arabic, etc.

4.2. Error Computation

The equal error rate (EER) was computed from two error measures using SAIC's RES engine which is the official SAD scoring software for the RATS program. One error measure is the probability of missing speech (Pmiss), and the second is the likelihood of falsely accepting the speech presence hypothesis (Pfa). These are computed as follows:

Pmiss = total_missed_speech / total_scored_speech
Pfa = total_false_accept_speech / total_scored_nonspeech

where total_missed_speech is the duration of the undetected speech regions, and total_scored_speech is the duration of all the speech regions from transcripts. Total_false_accept_speech is the duration of the falsely detected speech segments, and total_scored_nonspeech is the total duration of the nonspeech regions from transcripts.

4.3. Speech Detection Results

Table 1 shows the % EER for different input features on Dev-1 and Dev-2 sets. We first tested all the features in isolation. Next, in Table 2 we performed a two way combination between MFCC and each of the alternative SAD features. For example, in the first case we appended the 40-dimensional MFCC to a 4-dimensional GABOR feature, resulting in a 44 dimensional feature vector. Finally, in Table 2 we performed full feature combination between MFCC and the four SAD features resulting in a 56-dimensional feature vector. In Table 3 we present the channel specific results for the all feature combination system. Notice that channel D is missing, as it was officially excluded from scoring.

Table 1: Single Feature Speech Detection % EER Results.

Input Features	Feat Dim	Model Gauss	% EER Dev-1	% EER Dev-2
MFCC (baseline)	40	256	2.05	2.70
GABOR	4	32	4.00	5.45
Combo	4	32	4.20	4.75
SAcC	4	32	4.75	4.90
MBCombF0	4	32	4.10	6.15

In Table 1 we used the performance of the MFCC feature with DCT processing as the baseline, which resulted in a very low EER on both Dev-1 and Dev-2 sets. Next, we compared the other four SAD features in isolation. On Dev-1 GABOR achieves the lowest EER, followed by MBCombF0, and finally Combo and SAcC. However, on Dev-2 the best feature is Combo, followed by SAcC, GABOR and MBCombF0. This reveals that some features might be more robust to the specific types of distortions in one set but fail to generalize to the other set. The increased errors on Dev-2 might be due to the fact that SNRs are lower than that on Dev-1.

Table 2: Feature Combination Speech Detection % EER Results.

Input Features	Feat Dim	Model Gauss	% EER Dev-1	% EER Dev-2
MFCC + GABOR	44	256	1.70	2.50
MFCC + Combo	44	256	1.85	2.40
MFCC + SAcC	44	256	1.90	2.45
MFCC + MBCombF0	44	256	1.65	2.45
MFCC + All SAD	56	256	1.55	2.10

Analyzing the results in Table 2 we found big error reductions on both sets when combining one SAD feature with the MFCC feature compared to the baseline performance. On Dev-1 the best pairwise combination is with MBCombF0 followed by the combination with GABOR. Interestingly, this reverses the order of performance from Table 1 for each of these features in isolation. The combination with Combo and SAcC also produces error reductions compared to the baseline. On Dev-2 the best pairwise combination is achieved with Combo feature, followed closely by the combination with SAcC, MBCombF0 and finally GABOR. This trend in Dev-2 additionally shows that these different features in combination and over different test sets produce different gains, therefore it is expected that the combination of all features will result in further performance improvements.

Finally, the best performance is found from the all feature combination on both development sets. On Dev-1 the relative gain from the all feature combination system over the MFCC baseline is 24.3% and over the best pairwise combination is 6.0%. On Dev-2 the relative gain from the all feature combination system over the MFCC baseline is 22.2% (about the same as on Dev-1) and over the best pairwise combination is 12.5%. This means that each SAD feature provides different complementary information to the baseline MFCC feature. This is a very relevant result as three out of the four SAD features (Combo, SAcC and MBCombF0) aim at capturing voicing information. Since each of these features approaches the problem from a different perspective and use different processing techniques, the complementary information is expected.

Table 3: % EER Results by Channel on Dev-1 from MFCC and MFCC+All SAD Feature Systems.

Input Feature	A	B	C	E	F	G	H
MFCC	2.50	3.05	2.40	3.05	2.00	0.80	1.90
MFCC + All SAD	2.25	2.40	2.35	2.00	2.45	0.70	1.20

In Table 3 we present the channel specific results on Dev-1 from the MFCC only and the all feature combination systems, this last one is the one which performed best in Table 2. Comparing both systems, there is a gain from the feature combination system in all channels except for F. The best performance for the MFCC+All SAD system is achieved on channel G, followed by channel H and the rest of the channels with similar performance overall. On channel G the signal is very clear and SNR is higher compared to other channels. In addition, channel G data does not contain any non-transmit (NT) regions. Channel H also overall contains high SNR recordings. The other channels contain different types of distortions and vary in SNR and speech degradation types. Overall the performance is similar in those highly degraded channels which reveal a consistent behavior of the proposed SAD. However, performance on those degraded channels lag behind channels G and H, which reveals that there is still work to do to minimize that difference.

5. Conclusions

Our feature combination approach results in a highly accurate speech detector despite high degradation by channel noise and transmission distortions. We found significant gains from combining a MFCC acoustic feature with four speech activity detection features: GABOR, Combo, SAcC and MBCombF0l. These SAD features differ in their processing techniques, one is based on spectro-temporal processing and the other three are based on voicing measure estimation. Their different processing techniques and approaches result in different performances over two different test sets. The complementary information from these features results in important gains when combining with the baseline MFCC feature. Finally, we found important gains in performance when combining all the features, which is the major benefit from the feature combination explored in this paper.

6. Acknowledgements

This material is based on work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20024. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the DARPA or its Contracting Agent, the U.S. Department of the Interior, National Business Center, Acquisition & Property Management Division, Southwest Branch. The views expressed are those of the author 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.

7. References

[1] J. Ramírez, J.C Segura, C. Benítez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information", *Speech Commun*., vol. 42, pp. 271-287, 2004.

[2] J. Ramírez, P. Yelamos, J.M. Gorriz, and J.C. Segura "SVM-based speech endpoint detection using contextual speech features", *IEEE Electron. Lett.,* vol. 42, no.7, pp.426-428, 2006.

[3] S.G. Tanyer and H. Özer, "Voice activity detection in nonstationary noise", *IEEE Trans. Speech Audio Process., vol.* 8, no. 4, pp. 478-482, 2000.

[4] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold", *IEEE Trans. Audio, Speech Lang. Process.,* vol. 14, no. 2, pp. 412-424, 2006.

[5] B.T. Meyer, S.V. Ravuri, M.R. Schadler, and N. Morgan, "Comparing different flavors of spectro-temporal features for ASR," in *Proc. Interspeech*, Aug. 2011, pp. 1269-1272.

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

[7] M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg, and H.J. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust. Speech Signal Process., vol.* ASSP-22, pp. 353-362, 1974.

[8] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. IEEE ICASSP*, Apr. 1997 , pp. 1331-1334.

[9] L. N. Tan, and A. Alwan, "Multi-Band summary correlogram-based pitch detection for noisy speech", accepted to *Speech Commun.*.

[10] B.-S. Lee and D. Ellis, "Noise robust pitch tracking by subband autocorrelation classification," in *Proc. Interspeech*, Sep. 2012..

[11] 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. Interspeech*, Sep. 2012.

[12] K. Walker and S. Strassel, "The RATS Radio Traffic Collection System," in *Proc. Odyssey*, Jun. 2012.