ROBUST SPEECH ACTIVITY DETECTION IN THE PRESENCE OF NOISE

Ruhi Sarikaya and John H. L. Hansen

Robust Speech Processing Laboratory
Duke University, Box 90291, Durham, NC 27708-0291
http://www.ee.duke.edu/Research/Speech ruhi@ee.duke.edu jhlh@ee.duke

ABSTRACT

This study presents a new approach for robust speech activity detection (SAD). Our framework is based on HMM recognition of speech versus silence. We model speech as one of fourteen large phone classes whereas silence is represented as a separate model. Individual test utterances are concatenated to simulate read continuous speech for testing. The HMM-based algorithm is compared to both an energy based, as well as speech enhancement based, SAD algorithms for clean, 5 dB and 0 dB SNR levels under white Gaussian noise (WGN), aircraft cockpit noise (AIR) and automobile highway noise (HWY). We found that our algorithm provides lower frame error rates than the other two methods especially for HWY noise. Unlike other studies, we evaluate our algorithm on the core test set of the standard TIMIT database. Hence, results can be used as benchmarks to evaluate future systems.

1. INTRODUCTION

Speech activity detection (SAD) is one of the fundamental issues in many speech processing tasks such as continuous speech recognition and speech enhancement. Reliable discrimination between speech and silence becomes very difficult in the presence of noise. Robust SAD is required for pre-recognition noise reduction and recognizer model adaptation.

There are a number of approaches previously used for SAD. One approach is based on energy and it's derivatives [2]. Although energy based algorithms work reasonably well for clean speech, their performance degrades rapidly as SNR levels decrease. In [8], a word boundary detection algorithm is developed for isolated word recognition which does not address the problems encountered in spontaneous and continuous speech recognition, as there is typically no single beginning and end point, in continuous speech. Other endpoint detectors have been proposed based on energy for isolated word recognition. In [1], an optimized strategy for finding endpoints using a three-pass approach is proposed in which energy pulses were located, edited, and endpoint pairs scored in order of most likely candidates. However while it performs well for isolated utterances at SNRs of 30 dB or greater, it fails considerably at lower SNRs. In [9], Lamel's approach was modified to include delta energy besides energy as features, with speech and noise modeled using two HMMs. This method is also used for word boundary detection where training and testing data were single digits embedded in background noise. However the performance was not validated on noisy or continuous speech.

In [4], SAD is formulated in the framework of model-based speech enhancement. However, this approach is both complex and computationally expensive. The testing set was composed of only 2400 frames of speech which can be obtained from 8-10 TIMIT sentences. The performance of the algorithm has not been validated on a larger data set. We observed that detection rates vary considerably among sentence sets. For example while one set of sentences (amounting to 2374 frames) achieved 1.3% total error rate (false alarm + miss), another set with a similar frame count achieved as high as 8.6% frame error rate.

One common theme of these previously proposed SAD algorithms is the lack of a standard evaluation test database. Generally speaking, it may not be difficult to find single or small sentence sets which gives artificially low error rates. The literature lacks a benchmark study for which new systems can be evaluated against. This study establishes the performance of SAD on a well defined core test set.

In this paper we propose a solution to SAD based on broad class phone recognition where further performance improvements can be gained by extending to context independent and context dependent phone recognition based SAD at the expense of increases in computational complexity of the algorithm. Our approach is described in the next section. Furthermore, in order to investigate the viability of our system, we use the standard TIMIT database where test sentences are concatenated into blocks of four sentences to simulate continuous read speech. We compare our system to modified energy based SAD and as well as speech enhancement based SAD algorithms in the presence of WGN, AIR and HWY noises.

The rest of the paper is organized as follows. In the next section we describe our system for SAD. In Sec. 3 we describe the experimental evaluations and results obtained from various SAD systems. Next, we discuss some of the issues and compare our system with other algorithms. Finally, in Sec. 5 we summarize results and point to possible future work for further improvement.

2. ALGORITHM DESCRIPTION

The SAD problem can be formulated as a signal detection problem. $\mathbf{S}_0$ denotes the observation vectors for silence and $\mathbf{S}_i$ denotes the observation vectors for the i^{th}

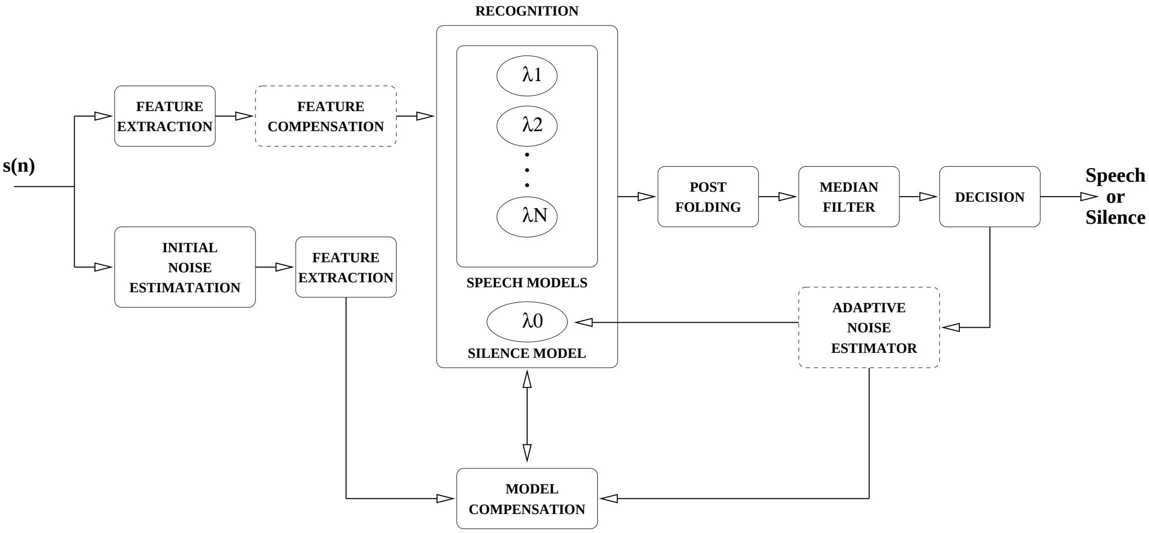

Figure 1: Block diagram for SAD.

speech unit in the feature space. Under the noise-free condition $\mathcal{H}_1$ denotes the hypothesis that the observation vector sequence $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, which is assumed to be Gaussian random vector, belongs to one of the N speech units. $(i: 1 \ldots N)$, whereas $\mathcal{H}_0$ denotes that it belongs to silence. The underlying density under each hypothesis is assumed to have a mixture of multi-D Gaussian densities.

$$\begin{aligned} \mathcal{H}_0: \; X &= \mathbf{S}_0 \\ \mathcal{H}_1: \; X &= \mathbf{S}_i \quad i = 1 \ldots N \end{aligned} \tag{1}$$

Equivalently the same decision criteria can be established in terms of speech versus silence models. An observation vector sequence $\mathbf{X}$ can be classified as one of the two classes of HMMs λ_0 and λ_i. The conditional probabilities $P(\mathbf{X}|\lambda_0)$ and $P(\mathbf{X}|\lambda_i)$ are calculated and the model resulting in the highest likelihood is selected (assuming equal *a priori* probabilities).

$$\begin{aligned} &Choose \quad \lambda_0, \quad P(\mathbf{X}|\lambda_0) > \underset{i:\,1 \ldots N}{\arg\max}\; P(\mathbf{X}|\lambda_i) \\ &Choose \quad \lambda_i, \quad otherwise \end{aligned} \tag{2}$$

In the presence of noise, the observation vector sequence $\mathbf{X}$ is transformed to $\tilde{\mathbf{X}}$ which has a different pdf than $\mathbf{X}$. In order to use the above decision criteria, the models λ_0 and λ_i should be transformed in such a way so that they are able to model the underlying distribution of $\tilde{\mathbf{X}}$.

$$\begin{aligned} \mathbf{X} &\Rightarrow \tilde{\mathbf{X}} \\ \lambda_0 &\Rightarrow \tilde{\lambda}_0 \\ \lambda_i &\Rightarrow \tilde{\lambda}_i \end{aligned} \tag{3}$$

Parallel model combination (PMC) can be used to transform noise free models to noisy models while retaining the likelihood ratio framework. The new decision rule is given below:

$$\begin{aligned} &Choose \quad \tilde{\lambda}_0, \quad P(\tilde{\mathbf{X}}|\tilde{\lambda}_0) > \underset{i:\,1 \ldots N}{\arg\max}\; P(\tilde{\mathbf{X}}|\tilde{\lambda}_i) \\ &Choose \quad \tilde{\lambda}_i, \quad otherwise \end{aligned} \tag{4}$$

Here, speech is modeled as a set of units and silence is a separate single unit. In this respect the problem is similar to a keyword spotting problem where speech units are keywords and silence is a garbage model. Although speech units are defined as one of the fourteen broad phone classes, they can be extended to individual context dependent/independent phone units for further improvement in performance.

The general framework of the algorithm is given in Fig. 1. The dotted blocks shown in Fig. 1 are optional processing steps which can be included. We assumed that speech is proceeded by a very short segment of silence from which the initial noise estimate is computed. Mel-frequency cepstral parameters are used in the feature extraction block. From the noise features, an estimate of the noise mean and covariance vectors are computed. These estimates are submitted to the Parallel Model Combination (PMC)[6] block which is used to transform the noise-free model pdfs to the noise-corrupted pdfs (i.e., $\lambda_i \Rightarrow \tilde{\lambda}_i$. Viterbi based recognition is used in the recognition block. All recognized speech units are folded into an overall speech class in the post-folding block. Confusions among speech models do not effect final speech/silence decision due to post-folding. A median filter of length 11 has been used to smooth the output of post-folding. This prevents frame-to-frame toggling among speech and silence states. In our simulations, since the noise is not varying over time we disable the noise update block, and used the initial noise estimate during the entire test scenario.

2.1 Broad Class Phone Recognition Based SAD

We consider speech as one of the following 14 broad phone classes: *nasals, unvoiced fricatives, voiced fricatives, affricates, unvoiced stops, voiced stops, u/v whispers, front-vowels, mid-vowels, back-vowels, schwa-vowels, diphthongs, liquid and glides.* Silence is considered as a separate class which is composed of {*epi, pau, q, qcl*} . The 61 TIMIT phones are folded into one of the above classes. The Viterbi algorithm is used for recognition, with the broad-class phone recognition output folded into either speech or silence. The feature vector is composed of 12 static, 12 delta, energy, delta energy. The zeroth cepstral parameter is appended to the feature vector to facilitate PMC compensation. Each of the broad classes as well as silence is modeled with 3 state left-to-right, 32 mixture HMMs. Since only 15 HMM models are used to perform recognition, the computational complexity is small compared to any typical speech enhancement schemes. A more complex recognition based SAD system would be based on context independent phone recognition. The best system is based on gender and context dependent phone recognition which is more complex than the first two systems. However the performance is expected to increase as more prior knowledge of the speech is taken into account.

2.1 Energy and Speech Enhancement based SAD

There are two other main approaches used in the past for SAD. The first is based on energy detection whereas the second is based on speech enhancement. There are a number of energy based SAD algorithms in the literature [2, 1, 9, 8, 4]. We used [2] for noise-free conditions and modified the same algorithm in a noise adaptive manner for noisy simulations. The algorithm proposed in [2] has two empirical thresholds which are functions of silence energy. In our simulations we optimized these thresholds for minimum error rate. For noisy cases the first ten frames are assumed to be noise alone. The mean estimate of noise energy is computed from the first ten frames. The utterance energy contour is normalized by subtracting the noise energy estimate in pointwise fashion. The same threshold setting is used for noisy conditions after normalization.

We also implemented a speech enhancement based SAD. Noisy speech is first enhanced by using the constrained iterative Wiener filter (Auto-LSP) approach [3]. This algorithm is based upon a two-step maximum *a posteriori* (MAP) estimation of the all-pole speech parameters and noise-free speech. In the first step, a MAP estimation of the clean speech is obtained from the noisy input speech (via Wiener filtering). In the second step, MAP estimation is used to produce the all-pole model parameters given the previous speech estimate. In between MAP estimation steps, spectral constraints are applied in order to (i) ensure stability of the all-pole model, (ii) to ensure that it possesses speech-like characteristics, and (iii) to provide frame-to-frame continuity in vocal tract characteristics. *Inter-frame* constraints are applied to the Line Spectrum Pair (LSP) parameters while *intra-frame* constraints are applied across iterations to the autocorrelation lag sequence. After enhancement, energy based detection is applied on the enhanced speech where an optimal threshold is selected for minimum frame error rate.

2.3 Parallel Model Combination (PMC)

We have used parallel model combination (PMC) [6] to update our HMM models in noisy conditions. The idea behind PMC is to adapt continuous density HMMs trained on clean cepstral speech data to make it more robust to noise. Given a segment of the noise itself, PMC combines the parameters of the corresponding pairs of speech and noise states to yield compensated sets of parameters. A thorough description of this technique can be found in the literature [7]. Although both static and delta parameter compensation are possible, we use only static parameter compensation which resulted in satisfactory performance.

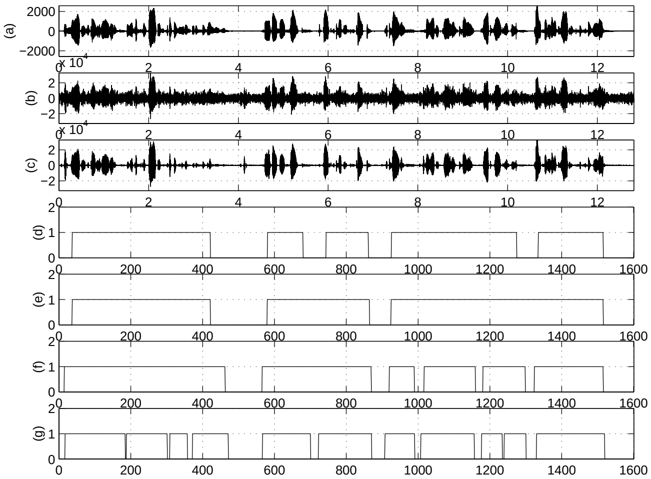

Figure 2: (a) clean speech, (b) noisy speech (0 dB highway noise), (c) enhanced speech, (d) energy-based SAD, (e) Speech enhancement based SAD, (f) Broad-Class recognition based SAD, (g) reference transciption.

3. EVALUATIONS

The SAD algorithms are evaluated on the clean (8kHz sampled) TIMIT core test set which inherently has 35 dB SNR. The models are estimated from data in the training set of TIMIT. The core test set of TIMIT is used for evaluation. Core test set is composed of 192 sentences contributed by 16 male and 8 female speakers. The testing data results in 57,700 frames to classify with a frame length of 20 msec and skip rate of 10 msec. The performance is established based on frame level error rate. Three types of noise sources are used for noisy simulations: white Gausssian (WGN), aircraft cockpit (AIR) and automobile highway (HWY) at 5 dB and 0 dB SNRs. A complete description of noise types can be found in [5]. In Fig. 2, a typical example of clean speech, noisy speech and enhanced speech is shown respectively in the first three plots. The next portion of the figure shows the decision using energy based SAD, speech enhancement based SAD and broad phone-class based SAD, respectively. The last graph shows the reference speech/silence regions of the speech. Here 1 denotes speech and 0 denotes silence. The speech file is obtained by concatenating four sentences which simulates read continuous speech. As seen in the plot, broad phone-class SAD closely traces the speech/silence portions of the speech whereas the other two algorithms make errors especially in the transition regions where many silence deletions and insertions occur.

3.1 Noise Free Simulations

In the noise-free case, broad phone-class recognition based SAD and energy based SAD are used. The results are shown in Table 1. The first algorithm achieved an error rate of 5.8% while the energy based SAD achieved 7.6%.

Speech Activity Detection (SAD) for (8kHz) TIMIT core test set Total silence frames: 9083, Total speech frames: 48617, Total frames: 57700													
		ALG. 1: Broad Class + PMC				ALG. 2: Energy Based				ALG. 3: SE + Energy Based			
Noise	SNR (dB)	FA	Miss	Corr	P(e) (%)	FA	Miss	Corr	P(e)(%)	FA	Miss	Corr	P(e)(%)
WGN	5	5058	2496	50146	**13.1**	2815	7483	47402	17.9	4661	4198	48841	15.4
	0	4923	4050	48709	**15.6**	6373	6001	45327	21.4	4400	5828	47472	17.7
AIR	5	3002	3562	52411	**11.1**	1439	8157	48104	16.6	5688	2799	49213	14.7
	0	3399	5413	48888	**15.3**	1066	11374	45260	21.6	5507	3442	48751	15.5
HWY	5	2167	1789	53744	**6.9**	2267	6375	49058	14.0	1848	6641	49211	14.7
	0	1576	3624	52500	**9.0**	1609	8448	47643	17.4	1545	7700	48455	16.0
Clean		2924	405	54371	**5.8**	2049	2315	53336	7.6	-	-	-	-

Table 1: Detection Errors, FA: false alarm, Corr: correct, P(e): probability of error, P(e)=P(false alarm) + P(miss)

The empirical thresholds for energy based detector are optimized over the testing set allowing an artificially low error rate for clean speech. For clean speech, many of the errors are made in transition frames, which contain speech as well as silence.

3.1 Noisy Simulations

In all noisy simulations, broad phone-class recognition outperformed speech enhancement based SAD, which in turn outperformed energy based SAD. We observed that broad phone-class based SAD is especially well suited to HWY noise, even at 5dB SNR the error rate is only 1% higher than the clean condition. On average broad phone-class based SAD outperformed speech enhencement based SAD and energy based SAD by 2.2% and 5.3% respectively for white Gaussian noise. The average difference is again in favor of our algorithm by 1.9% and 5.9% for AIR noise. Finally, our algorithm outperformed the other methods by 7.4% and 8.3% for HWY noise.

4. DISCUSSION

Although it is a standard database, one of the issues that comes up when TIMIT is used for SAD is the ill-proportioned amount of speech versus silence data. In the test set there are 9083 frames are silence compared to 48617 frames of speech. This imbalance might lead to a bias in performance. If the entire test set is labeled as speech the error rate will be 15.8%. However higher error rates can be obtained as neither very small false alarm nor miss are allowed. Therefore neither total number of misses nor the total number of false alarms should be small. Rather, their value should be comparable for accurate performance assessment.

Although PMC can compensate static and dynamic parameters, only static parameters were compensated. In [7] it was shown that compensating delta and delta-delta in addition to static parameters halves the relative error rate for continous speech recognition. We expect a similar performance improvement would also to translate to our SAD.

We can further improve the detection rates by using context-dependent phone recognition system. Context independent phone recognition based SAD is feasible since only fourty six models must be compensated. However context-dependent phone recognition based SAD would also be computationally expensive.

5. CONCLUSIONS AND FUTURE WORK

The problem of speech activity detection is addressed by formulating a broad phone-class recognition system. The detector is shown to perform well even at low SNRs. It's performance is compared to a modified-energy and speech enhancement based detectors. While providing lower error rates than the other two methods for all noise types, it is especially well suited for automobile highway (HWY) noise resulting in half the error rate of the other two methods. This study can be used as a benchmark for future systems as it used the core TIMIT set for simulations. Currently we are working on two new algorithms which will be integrated into our system. The first is normalized likelihood ratio scoring which is intended to reduce the number of false alarms. The second is to build boundary-HMMs for robust decision on speech-noise boundaries.

ACKNOWLEDGEMENT

The authors thank to B. L. Pellom for providing the PMC routine.

1. REFERENCES

[1] L.F. Lamel, L.R. Rabiner, A.E. Rosenberg and J.G. Wilpon,"An improved Endpoint Detector for Isolated Word Recognition," *IEEE Trans. ASSP,* **29**:777-85, 1981.

[2] L.R. Rabiner and M.R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances," *AT&T Bell Lab. Tech. J. ,* **54**(2):297-315, 1975.

[3] B.L. Pellom and J.H.L. Hansen, "An Improved (Auto:I;LSP:T) Constrained Iterative Speech Enhancement Algorithm for Colored Noise Environments," *IEEE Trans. on SAP,* (to appear late 1998/early 1999).

[4] B.L. McKinley and G.H. Whipple, "Model Based Speech Pause Detection," *ICASSP-97,* vol. 2, pp. 1179-1182, 1997.

[5] J.H.L. Hansen and L.M. Arslan, "Robust Feature-Estimation and Objective Quality Assessment for Noisy Speech Recognition Using the Credit Card Corpus," *IEEE Trans. on SAP,* **3**(3):169-184, 1995.

[6] M.J.F. Gales and S.J. Young, "Cepstral Parameter Compensation for HMM Recognition in Noise," *Speech Communication,* **12**:231-239, 1993.

[7] M.J.F. Gales and S.J. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination," *IEEE Trans. on SAP,* **4**(5):352-360, 1996.

[8] J.C. Junqua, B. Mak and B. Reaves, "A robust Algorithm for Word Boundary Detection in the Presence of Noise," *IEEE Transactions on SAP,* **2**(3):406-412, 1994.

[9] A. Acero, C. Crespo, C. de la Torre and J.C. Torrecilla, "Robust HMM-Based Endpoint Detector," *EUROSPEECH-93,* vol. 3, pp. 1551-1554, 1993.