ROBUST UNSUPERVISED DETECTION OF HUMAN SCREAMS IN NOISY ACOUSTIC ENVIRONMENTS

*Mahesh Kumar Nandwana, Ali Ziaei, John H. L. Hansen**

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson, Texas, USA
{mahesh.nandwana,ali.ziaei,john.hansen}@utdallas.edu

ABSTRACT

This study is focused on an unsupervised approach for detection of human scream vocalizations from continuous recordings in noisy acoustic environments. The proposed detection solution is based on compound segmentation, which employs weighted mean distance, T^2-statistics and Bayesian Information Criteria for detection of screams. This solution also employs an unsupervised threshold optimized Combo-SAD for removal of non-vocal noisy segments in the preliminary stage. A total of five noisy environments were simulated for noise levels ranging from -20dB to +20dB for five different noisy environments. Performance of proposed system was compared using two alternative acoustic front-end features (i) Mel-frequency cepstral coefficients (MFCC) and (ii) perceptual minimum variance distortionless response (PMVDR). Evaluation results show that the new scream detection solution works well for clean, +20, +10 dB SNR levels, with performance declining as SNR decreases to -20dB across a number of the noise sources considered.

Index Terms— scream detection, T^2 distance, PMVDR, CompSeg, T^2-BIC SAD

1. INTRODUCTION

The presence of non-speech sounds in continuous audio streams has adverse effects on the performance of speech coding systems, spoken document retrieval, speech and speaker recognition systems. Thus, it becomes necessary to detect and suppress these events in early stages of overall real-time speech systems. Apart from a preprocessing step in speech systems, applications of scream detection can be employed in the area of acoustic surveillance or situation awareness where detecting these events, deviations from a calm monitored space can be observed and appropriate actions taken to address the situation.

Human sounds produced via the oral cavity can be classified into two broad categories: i) speech and ii) non-speech. Non-speech sounds includes vocalizations such as: scream, whistle, cough, laugh, snore, sneeze, hiccups, etc. This study focuses on detection of human screams which reflects a portion of the class of non-speech sounds.

During the past two decades, extensive research has been accomplished in audio signal classification such as speech/music/ environmental sounds. Environmental sniffing is a framework where the environmental acoustics are analyzed to direct speech system reconfiguration [1, 2]. Recently, the research community has gained interest in further detection and classification of non-speech human sounds such as screams, coughs [3], snores [4], laughs [5] etc. because of their increasing applications in various areas including elder care, home and health care, security and safety [6].

In this paper, our goal is to detect human screams from continuous recordings in realistic noisy acoustic environments. We have collected our own corpus and simulated different environments with five different noise levels. Noisy environments include: babble, car, factory, volvo and white Gaussian noise (WGN). For effective detection, we use a two step approach, the first step employs a threshold optimized unsupervised combo-SAD for silence removal [7] followed by the next step which uses the Hotelling's T^2-statistics and Bayesian Information Criteria [8, 9] for speech/scream detection. Performance of the proposed system is compared for two front-ends features namely MFCC and PMVDR [10].

The remainder of this paper is organized as follows: first the corpus collection and simulation is discussed in detail. Sec. 3 discusses the threshold optimized combo-SAD. Next, formulation of the CompSeg algorithm which employs the weighted mean distance, T^2-statistics and Bayesian Information Criteria is discussed. In Sec. 5 experimental evaluation is presented. Finally, concluding remarks and directions for future research are given in Sec. 6.

2. RELATION TO PRIOR WORK

In recent years, there has been great interest in analysis and detection of non-speech sounds, particularly screams because of their increasing applications. Most earlier work focuses either on front-end features or on acoustic modeling.

In [6], analytical and statical features along with an SVM classifier were used for scream detection. In [11], MFCC, MPEG-7 features and HMMs were used for scream and gunshot event classification. Also in [12], a parallel GMM classifier network was utilized for ambient noise, scream and gunshot detection.

Also various approaches proposed for shout detection [13, 14] are not considered here, since shouted speech contains phonetic structure in audio, where as in pure scream there is no phonetic structure.

This work is different from previously proposed approaches in the following sense: 1) we employ a completely unsupervised approach, to the best of our knowledge almost all the previous approaches are supervised; 2) Data was collected from human subjects rather than using sound effects for scream from movies or internet repositories; 3) a much wider set and range of noise levels are con-

*This project was funded by AFRL under contract FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

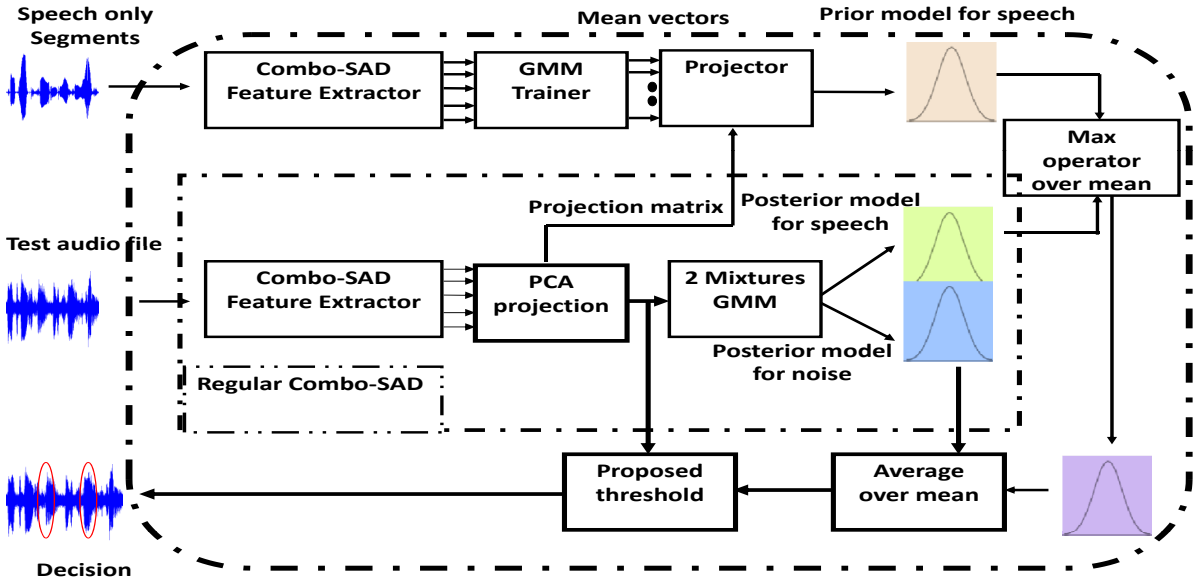

Fig. 1. Block diagram of threshold optimized vocal activity detection system.

sidered. Also, in our previous studies, we considered analysis of human screams along with its impact on the performance of speaker recognition systems [15]. Continuing along that theme, here we develop this pre-processing stage for detection and removal of screams to improve robustness of SID systems.

3. CORPUS DEVELOPMENT

The corpus for this study was collected, simulated and annotated at the University of Texas at Dallas in two phases. In phase I, speech data and scream data was collected from speakers, and in phase II, data was simulated for noisy speech and scream environments.

Phase I was a combination of three parts: Part 1 consisted of recording text dependent neutral read speech (25 TIMIT sentences). Part 2 consists of 12 questions to be answered for recording of spontaneous speech. In Part 3, subjects were told to scream with pause between scream events. More Corpus details are included in [15]. During recording, the gain of the microphone was adjusted to ensure that signal strength was sufficient for analysis as well as to avoid clipping at the same time. Sample audio clips of scream events are available at http://crss.utdallas.edu/Projects/SID_Scream/

In Phase II, a total of 24 audio files were generated consisting of speech and screams with a purely random weighting balance of speech versus scream. The duration of audio files were also random, ranging from 30 seconds to about 4 minutes. After generation of audio files, noise from NOISEX-92 database [16] were added using the Filtering and Noise adding Toolbox (FaNT) [17]. Five different types of noise, namely white Gaussian, babble, car, factory and Volvo were added at SNR levels of 20, 10, 0, -10, and -20 dB. Clean files were transcribed as silence/speech/scream by an expert human annotator which was used as ground truth.

4. VOCAL ACTIVITY DETECTION

Vocal activity detection is the primary step in any non-speech acoustic event detection problem from continuous recordings. Therefore, as a first step, we remove noisy non-vocal segments from these audio streams. Most of the previously proposed detection algorithms assume a homogeneous style of audio data, whereas in surveillance applications the data usually consists of a wide variety of audio depending on indoor versus outdoor situations. Our vocal activity detection system is inspired by the recently proposed UT-Dallas Combo-SAD system for the DARPA RATS program [7]. Combo-SAD was preferred because of its robustness towards non-homogeneous audio, noise and channel distortions.

4.1. Combo-SAD

Combo-SAD is an unsupervised approach for speech activity detection. It uses five front-end features which are computed at frame level. Five features include four voicing measures (harmonicity, clarity, prediction gain, and periodicity) along with perceptual spectral flux. These features are combined and normalized to form a five dimensional combo feature vector. Both mean and variance normalization is performed using,

$$x_i' = \frac{x_i - \mu}{\sigma}. \quad (1)$$

where x_i is the i-th feature frame, and the corresponding mean μ and variance σ were computed across all frames of an utterance. Finally, this five dimensional combo feature vector is projected into single dimension using principle component analysis (PCA), preserving the dimension corresponding to the largest eigenvalue. It is observed that this one dimensional combo feature vector acts as a great discriminator between vocal/non-vocal segments. The Combo feature vector has a bimodal distribution, with values higher for vocal and

lower for non-vocal segments. Therefore, a 2-mixture GMM is used for classification between vocal and non-vocal segments. The mixture with the higher mean is assigned to the vocal segment. The threshold (θ) is computed using following equation,

$$\theta = k\mu_v + (1-k)\mu_{nv}., \quad (2)$$

where μ_v and μ_{nv} are the means of mixtures of vocal and non-vocal respectively, and k is a weighting factor ranging from 0 to 1.

4.2. Threshold Optimization

The threshold estimation method in combo-SAD assumes that the audio stream always contains some speech and pause in balanced proportions. However, in the case of surveillance systems there are generally long periods of silence during night monitoring resulting in a greater number of false alarms. For these cases, combo-SAD results in relatively poor estimates of the vocal and non-vocal model distributions and threshold, resulting in high error. In order to overcome this issue, a new solution for threshold optimization was proposed in [18].

For better estimation, we first train a GMM with a larger number of mixtures (256 in this work) using an annotated speech corpus (usually Switchboard, Fisher etc.). Next, we project the means of this GMM onto a single dimension decision space of combo-SAD. We denote the mean of these projected values as μ_{tv}.

Here, μ_{tv} (developed using a secondary speech corpus) and μ_v (developed using combo-SAD) is denoted as prior and posterior models of speech or vocalizations. For decision making on the vocalized GMM, we use the following criteria: (i) if $\mu_v \geq \mu_{tv}$, we use the posterior model; (ii) if $\mu_v < \mu_{tv}$ we use the prior model. Therefore, the new method for threshold estimation is,

$$\theta = k\max(\mu_v, \mu_{tv}) + (1-k)\mu_{nv}. \quad (3)$$

A block diagram of threshold optimized vocal activity detection is shown in Fig. 1. Hence, the proposed approach achieves better decision making in regions where there is more silence compared to speech using a prior model for decision making.

5. COMPOUND SEGMENTATION ALGORITHM

After removing silence and noisy portions from the audio streams, next we detect the boundaries of scream and speech. Most distance metrics used for segmentation such as BIC or Kullback-Leibler distance (KL2), result in more estimation error if the data is insufficient because they require second order statistics (i.e., the covariance). The T^2-statistic combined with BIC has previously been used to formulate audio stream segmentation/SAD, resulting in the T^2-BIC algorithm[8], be employed in a new scenario here for scream detection.

CompSeg is an unsupervised algorithm, first proposed by Huang and Hansen [9], which detects the input acoustic change points based on features and uses three different distance metrics based on the length of the analysis window size. It also improves segmentations for short segments which are generally used in speaker ID trials. This algorithm uses an equal covariance assumption. Thus, more data is used for covariance estimation thereby reducing the effect of insufficient data in the estimation process. One more assumption is that if the analysis window is less than 2 sec., we assume the global covariance to be an identity matrix, which is termed as weighted mean distance. The distance metric rules for the CompSeg algorithm is summarized:

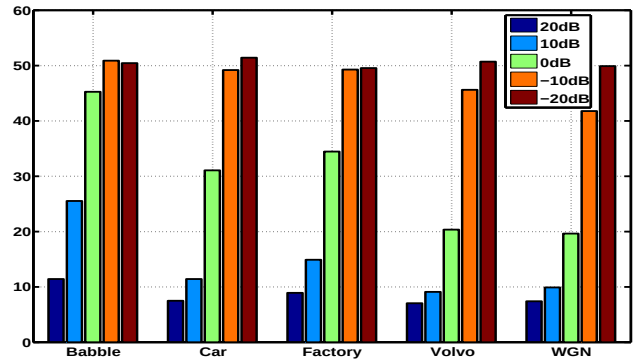

Fig. 2. TO-Combo-SAD Vocal activity detection EER(%) for different environments.

- $L_w < 2s$: weighted mean distance,
- $2s \leq L_w < 5s$: T^2-distance,
- $L_w \geq 5s$: traditionl BIC,

here, L_w is the length of the analysis window. The weighted mean and T^2 distance collectively are called T^2-mean, which is used for the processing window of size $<$ 5s. For windows of size $>$ 5s, the traditional BIC [19] is applied to detect break points.

6. EXPERIMENTAL EVALUATION

In this section, we evaluate the performance of the proposed scream detection approach. For evaluations, data was down sampled to 8kHz. First, we have processed the continuous audio streams through threshold optimized combo-SAD for vocal activity detection, and then used CompSeg for speech/scream detection.

6.1. TO-Combo-SAD Results

For TO-Combo-SAD, combo features were extracted using a frame size of 40ms and a skip rate of 10ms. The threshold was optimized using the approach described in Sec. 4.2, for this task k=0.4 was used in Eq. (3). Performance of the vocal activity detection system was computed in terms of equal error rate (EER). The results for TO-Combo-SAD are shown in Fig. 2. We observe that for low levels of noise, the EER is below or near 10% but as noise levels increase, the EER degrades rapidly (i.e. EER ∼50%).

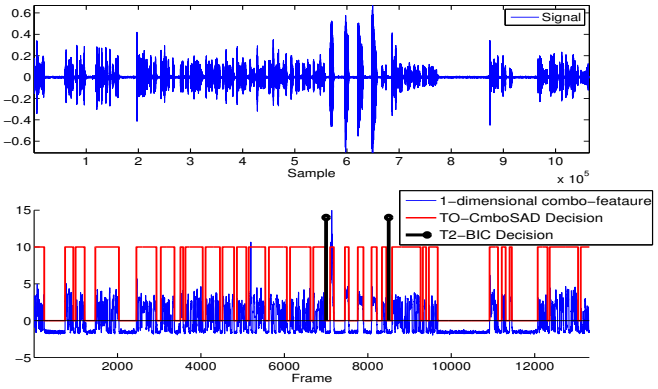

Fig. 3. TO Combo-SAD decisions and scream detection using CompSeg in an audio stream.

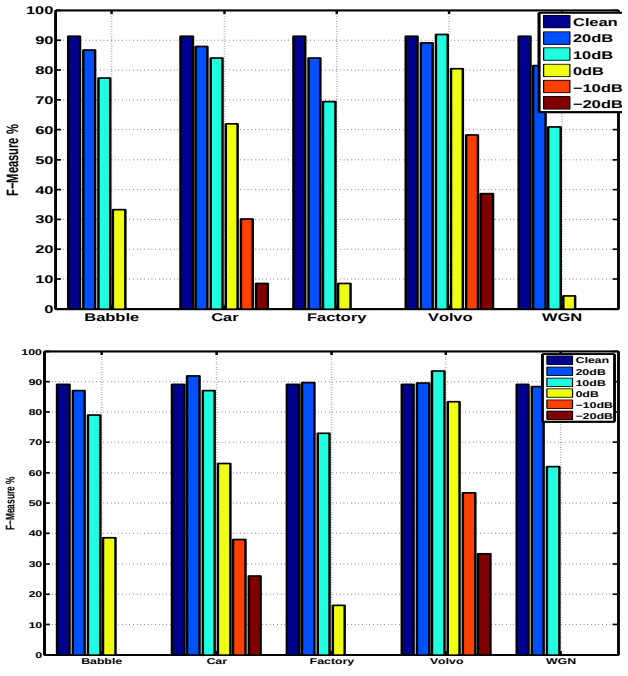

Fig. 4. Scream/Speech detection results for MFCC (top) and PMVDR (bottom) front-end.

6.2. CompSeg Scream/Speech Detection Results

After vocal activity detection, silence and noisy frames were discarded from the continuous streams. The remaining stream was further processed and the compound segment detection algorithm applied. CompSeg can be used with a number of front-ends [9]. For CompSeg evaluation, we have used two front-ends. Performance was evaluated in terms of percentage F-measure.

6.2.1. Mel-frequency cepstral coefficients(MFCC)

MFCCs are the most common features used for analysis of speech. They are computed by applying a Mel-scaled filter bank either to the short-term FFT magnitude spectrum or short term LPC-based spectrum to obtain a perceptually meaningful smoothed overall spectrum. DCT is then applied.

6.2.2. Perceptual minimum variance distortionless response (PMVDR)

The PMVDR feature is obtained by incorporating perceptual warping of FFT power spectrum, and replacing the Mel-scaled filter bank with the minimum variance distortionless response (MVDR) spectral estimator. The MVDR based spectrum has a better spectral modeling ability for high pitch signals [10]. In previous studies, PMVDR has been found to perform better than MFCC for scream modeling [15].

A total of 36-dimension features were computed for both MFCC and PMVDR which include 12 static, delta and delta-delta. We use a frame rate of 100 frames/sec, where each frame is 20 ms in duration with an overlap of 50% between adjacent frames.

Fig. 3 illustrates the TO-Combo-SAD decision and detected break points for scream in a continuous audio stream. The overall detection results for both the front-ends are summarized in Fig. 4. We have observed that for upto 10dB SNR, the proposed scream detection algorithm provides satisfactory performance. Even for noise types volvo and car performance is above 60% for 0dB. However for noise type babble, factory and WGN, the system performance is near zero at low SNRs of -10dB and -20dB. Therefore, the ability to detect screams in continuous audio streams over a range of noise type have been shown up to 10dB SNR.

7. CONCLUSION AND FUTURE WORK

In this study, we presented an unsupervised approach for robust detection of human screams and applied it across continuous noisy audio recordings. The proposed approach is a hybrid solution which combines threshold optimized combo-SAD and a CompSeg system. It has been shown that the proposed algorithm is noise robust and gives better performance even at SNR levels as low as 0dB for noise types car and volvo. In future work, we plan to extend this algorithm for the UT-Non-Speech II corpus which includes more than two hours of screams along with speech from 57 speakers. Perceptual tests will also be conducted to compare human vs. machine detection accuracies.

8. REFERENCES

[1] M. Akbacak and J.H.L. Hansen, "Environmental sniffing: noise knowledge estimation for robust speech systems," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. IEEE, 2003, vol. 2, pp. II–113.

[2] M. Akbacak and J.H.L. Hansen, "Environmental sniffing: noise knowledge estimation for robust speech systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 465–477, 2007.

[3] S. Matos, S.S. Birring, I.D. Pavord, and D.H. Evans, "Detection of cough signals in continuous audio recordings using hidden markov models," *Biomedical Engineering, IEEE Transactions on*, vol. 53, no. 6, pp. 1078–1083, 2006.

[4] W.H. Liao and Y.K. Lin, "Classification of non-speech human sounds: Feature selection and snoring sound analysis," in *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, Oct 2009, pp. 2695–2700.

[5] L.S. Kennedy and D.P.W. Ellis, "Laughter detection in meetings," in *NIST ICASSP 2004 Meeting Recognition Workshop, Montreal*. National Institute of Standards and Technology, 2004, pp. 118–121.

[6] W. Huang, T.K. Chiew, H. Li, T.S. Kok, and J. Biswas, "Scream detection for home applications," in *Industrial Electronics and Applications (ICIEA), 2010 the 5th IEEE Conference on*, June 2010, pp. 2115–2120.

[7] S.O. Sadjadi and J.H.L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *Signal Processing Letters, IEEE*, vol. 20, no. 3, pp. 197–200, March 2013.

[8] B. Zhou and J.H.L. Hansen, "Efficient audio stream segmentation via the combined t 2 statistic and bayesian information criterion," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 467–474, 2005.

[9] R. Huang and J.H.L. Hansen, "Advances in unsupervised audio classification and segmentation for the broadcast news and ngsw corpora," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 907–919, May 2006.

[10] U.H. Yapanel and J.H.L. Hansen, "A new perceptually motivated mvdr-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Communication*, vol. 50, no. 2, pp. 142–152, 2008.

[11] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 165–168.

[12] L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection in noisy environments," in *15th European Signal Processing Conference (EUSIPCO-07), Sep. 3-7, Poznan, Poland*, 2007.

[13] J. Pohjalainen, P. Alku, and T. Kinnunen, "Shout detection in noise," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4968–4971.

[14] V.K. Mittal and B. Yegnanarayana, "Production features for detection of shouted speech," in *Consumer Communications and Networking Conference (CCNC), 2013 IEEE*. IEEE, 2013, pp. 106–111.

[15] M.K. Nandwana and J.H.L. Hansen, "Analysis and identification of human scream: Implications for speaker recognition," *INTERSPEECH 2014*, pp. 2253–2257, 2014.

[16] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[17] H.G. Hirsch, "FaNT-filtering and noise adding tool," 2005.

[18] A. Ziaei, L. Kaushik, A. Sangwan, J.H.L. Hansen, and D.W. Oard, "Speech activity detection for nasa apollo space missions: Challenges and solutions," in *INTERSPEECH 2014*, 2014.

[19] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*. Virginia, USA, 1998, p. 8.