C-Means Clustering Applied to Speech Discrimination

J.M. Górriz[1], J. Ramírez[1], I. Turias[2],
C.G. Puntonet[3], J. González[3], and E.W. Lang[4]

[1] Dpt. Signal Theory, Networking and communications, University of Granada, Spain
gorriz@ugr.es
http://www.ugr.es/~gorriz
[2] Dpt. Computer Science, University of Cádiz, Spain
[3] Dpt. Computer Architecture and Technology, University of Granada, Spain
[4] AG Neuro- und Bioinformatik, Universität Regensburg, Deutschland

Abstract. An effective voice activity detection (VAD) algorithm is proposed for improving speech recognition performance in noisy environments. The proposed speech/pause discrimination method is based on a hard-decision clustering approach built over a set of subband log-energies. Detecting the presence of speech frames (a new cluster) is achieved using a basic sequential algorithm scheme (BSAS) according to a given "distance" (in this case, geometrical distance) and a suitable threshold. The accuracy of the Cl-VAD algorithm lies in the use of a decision function defined over a multiple-observation (MO) window of averaged subband log-energies and the modeling of noise subspace into cluster prototypes. In addition, time efficiency is also reached due to the clustering approach which is fundamental in VAD real time applications, i.e. speech recognition. An exhaustive analysis on the Spanish SpeechDat-Car databases is conducted in order to assess the performance of the proposed method and to compare it to existing standard VAD methods. The results show improvements in detection accuracy over standard VADs and a representative set of recently reported VAD algorithms.

1 Introduction

The emerging wireless communication systems are demanding increasing levels of performance of speech processing systems working in noise adverse environments. These systems often benefits from using voice activity detectors (VADs) which are frequently used in such application scenarios for different purposes. Speech/non-speech detection is an unsolved problem in speech processing and affects numerous applications including robust speech recognition [1, 2], discontinuous transmission [3, 4], real-time speech transmission on the Internet [5] or combined noise reduction and echo cancellation schemes in the context of telephony [6]. The speech/non-speech classification task is not as trivial as it appears, and most of the VAD algorithms fail when the level of background noise increases. During the last decade, numerous researchers have developed different

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part I, LNCS 3991, pp. 649–656, 2006.

strategies for detecting speech on a noisy signal [7] and have evaluated the influence of the VAD effectiveness on the performance of speech processing systems [8]. Most of them have focussed on the development of robust algorithms with special attention on the derivation and study of noise robust features and decision rules [9, 10, 11, 7]. The different approaches include those based on energy thresholds, pitch detection, spectrum analysis, zero-crossing rate, periodicity measure or combinations of different features.

The speech/pause discrimination can be described as an unsupervised learning problem. Clustering is one solution to this case where data is divided into groups which are related "in some sense". Despite the simplicity of clustering algorithms, there is an increasing interest in the use of clustering methods in pattern recognition [15], image processing [16] and information retrieval [17, 18]. Clustering has a rich history in other disciplines [12] such as machine learning, biology, psychiatry, psychology, archaeology, geology, geography, and marketing. Cluster analysis, also called data segmentation, has a variety of goals. All related to grouping or segmenting a collection of objects into subsets or "clusters" such that those within each cluster are more closely related to one another than objects assigned to different clusters. Cluster analysis is also used to form descriptive statistics to ascertain whether or not the data consist of a set of distinct subgroups, each group representing objects with substantially different properties.

The essay is organized as follows: in section 2 we describe a suitable signal model to detect the presence of speech frames in noisy environments. In the following section 3, we apply cluster analysis to form "descriptive statistics" transforming the noise sample set into a soft-noise model with low dimensional feature. A complete experimental framework is shown in section 4. Finally we state some conclusions and acknowledgements in the last part of the paper.

2 A Suitable Model for VAD

Let $x(n)$ be a discrete time signal. Denote by y_j a frame of signal containing the elements:

$$\{x_i^j\} = \{x(i + j \cdot D)\}; \quad i = 1 \dots L \tag{1}$$

where D is the window shift and L is the number of samples in each frame. Consider the set of $2 \cdot m + 1$ frames $\{y_{l-m}, \dots y_l \dots, y_{l+m}\}$ centered on frame y_l, and denote by $Y(s, j)$, $j = l - m, \dots l \dots, l + m$ its Discrete Fourier Transform (DFT) resp.:

$$Y_j(\omega_s) \equiv Y(s, j) = \sum_{n=0}^{N_{FFT}-1} x(n + j \cdot D) \cdot exp\left(-j \cdot n \cdot \omega_s\right). \tag{2}$$

where $\omega_s = \frac{2\pi \cdot s}{N_{FFT}}$, $0 \leq s \leq N_{FFT} - 1$ and N_{FFT} is the number of points or resolution used in the DFT (if $N_{FFT} > L$ then the DFT is padded with zeros). The log-energies for the l-th frame, $E(k, l)$, in K subbands ($k = 0, 1, ..., K - 1$), are computed by means of:

$$E(k,l) = \log\left(\frac{K}{N_{FFT}}\sum_{s=s_k}^{s_{k+1}-1}|Y(s,l)|^2\right)$$
$$s_k = \left\lfloor\frac{N_{FFT}}{2K}k\right\rfloor \quad k = 0,1,...,K-1, \tag{3}$$

where an equally spaced subband assignment is used and $\lfloor\cdot\rfloor$ denotes the "floor" function. Hence, the signal log-energy is averaged over K subbands obtaining a suitable representation of the input signal for VAD [19], the observation vector at frame l, $\mathbf{E}(l) = (E(0,l),\ldots,E(K-1,l))^T$. The VAD decision rule is formulated over a sliding window consisting of $2m+1$ observation (feature) vectors (log-energies) around the frame for which the decision is being made (l), as we will show in the following sections. This strategy consisting on "long term information" provides very good results using several approaches for VAD such as [13, 14] etc.

3 C-Means Clustering over the Feature Vectors

C-means clustering is a method for finding clusters and cluster centers in a set of unlabeled data [20]. The number of cluster centers (prototypes) C is a priori known and the **C**-means iteratively moves the centers to minimize the total within cluster variance. Given an initial set of centers the **C**-means algorithm alternates two steps [21]:

- for each cluster we identify the subset of training points (its cluster) that is closer to it than any other center;
- the means of each feature for the data points in each cluster are computed, and this mean vector becomes the new center for that cluster.

3.1 Noise Modeling

In our algorithm, this procedure is applied to a set of initial pause frames (log-energies) in order to characterize the noise space. Then we call this set of clusters noise prototypes [1]. Each observation vector (**E** from equation 3) is uniquely labeled, by the integer $i \in \{1,\ldots,N\}$, and uniquely assigned to a prespecified number of prototypes $C < N$, labeled by an integer $c \in \{1,\ldots,C\}$. The dissimilarity measure between observation vectors is the squared Euclidean distance:

$$d(\mathbf{E}_i,\mathbf{E}_j) = \sum_{k=0}^{K-1}(E(k,i)-E(k,j))^2 = ||\mathbf{E}_i-\mathbf{E}_j||^2 \tag{4}$$

and the loss function to be minimized is defined as:

$$W(C) = \frac{1}{2}\sum_{c=1}^{C}\sum_{C(i)=c}\sum_{C(j)=c}||\mathbf{E}_i-\mathbf{E}_j||^2 = \sum_{k=1}^{C}\sum_{C(i)=c}||\mathbf{E}_i-\bar{\mathbf{E}}_k||^2, \tag{5}$$

[1] The word cluster is assigned to different classes of labeled data, that is **K** is fixed to 2 (noise and speech frames).

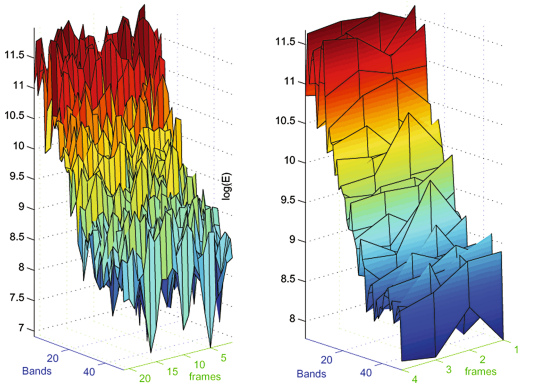

Fig. 1. a) 20 log Energies of noise frames, computed using $N_{FFT} = 256$, averaged over 50 subbands. b) Clustering approach to the latter set of log Energies using hard decision **C**-means (C=4 prototypes).

where $C(x)$ denotes the prototype associated to observation x and

$$\bar{\mathbf{E}}_c = (\bar{E}(1,c), \ldots, \bar{E}(K,c))^T \tag{6}$$

is the mean vector associated with the c-th prototype. Thus, the loss function is minimized by assigning the N observations to the C prototypes in such a way that within each prototype the average dissimilarity of the observations is minimized. Once convergence is reached, N K-dimensional pause frames are efficiently modeled by C K-dimensional noise prototype vectors denoted by $\bar{\mathbf{E}}_c^{opt}$, $c = 0, \ldots, C-1$. In figure 1 we observed how the complex nature of noise can be simplified (smoothed) using a clustering approach. The clustering approach speeds the decision function in a significant way since the dimension of feature vectors is reduced substantially ($N \rightarrow C$).

3.2 Soft Decision Function for VAD

In order to classify the second labeled data (log energies of speech frames) we use a BSAS using a MO window centered at frame l, as shown in section 2. For this purpose let consider the same dissimilarity measure, a threshold of dissimilarity γ and the maximum clusters allowed $\mathbf{K} = 2$.

Let $\hat{\mathbf{E}}(l)$ be the decision feature vector that is based on the MO window as follows:

$$\hat{\mathbf{E}}(l) = \max\{\mathbf{E}(i)\}, \quad i = l-m, \ldots, l+m \tag{7}$$

This selection of the feature vector describing the actual frame is useful as it detects the presence of voice beforehand (pause-speech transition) and holds the detection flag, smoothing the VAD decision (as a hangover based algorithm [11, 10] in speech-pause transition).

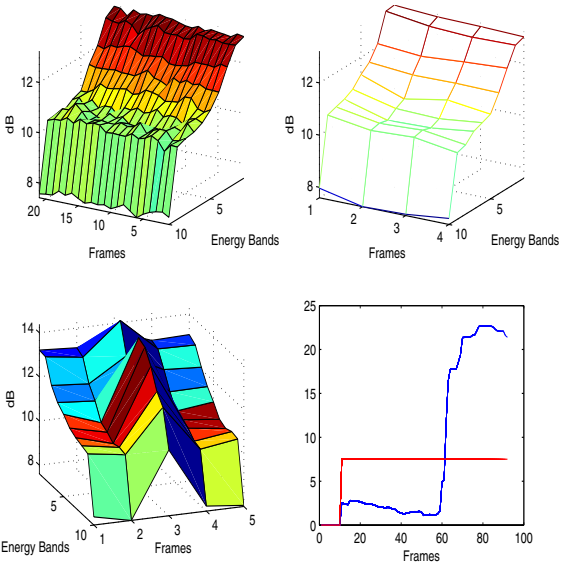

Fig. 2. Step of the algorithm. The frame selected is classified as speech frame (VAD=1) as is shown in the decision function a) Noise log-energy subbands. b) **C**-means centers prototypes. c) comparison between noise prototypes and the log-energy of the current frame. d) decision function and threshold versus frames.

Finally, the presence of a new cluster (speech frame detection) is satisfied if:

$$||\hat{\mathbf{E}}(l) - < \bar{\mathbf{E}}_c > ||^2 > \gamma \tag{8}$$

where $< \bar{\mathbf{E}}_c >$ is the averaged noise prototype and γ is the decision threshold. The set of noise prototypes are updated in pause frames (not satisfying equation 8)) in a competitive manner (only the closer noise prototype is moved towards the current feature vector):

$$\bar{\mathbf{E}}_{c'} = arg_{min} \left(||\bar{\mathbf{E}}_c - \hat{\mathbf{E}}(l)||^2 \right) \quad \Rightarrow \quad \bar{\mathbf{E}}_{c'}^{new} = \alpha \cdot \bar{\mathbf{E}}_{c'}^{old} + (1 - \alpha) \cdot \hat{\mathbf{E}}(l) \tag{9}$$

where α is a normalized constant with value close to one for a soft decision function (i.e. we selected in simulation $\alpha = 0.99$).

In figure 2 we show an step detail in the algorithm. We display the noise log energy model (top-left), the clustering **C**-means approach, the log-energy of current frame (frame=3) included in the noise prototypes ($C = 4$) and the decision rule versus time.

4 Experimental Framework

Several experiments are commonly conducted to evaluate the performance of VAD algorithms. The analysis is normally focused on the determination of misclassification errors at different SNR levels [11], and the influence of the VAD

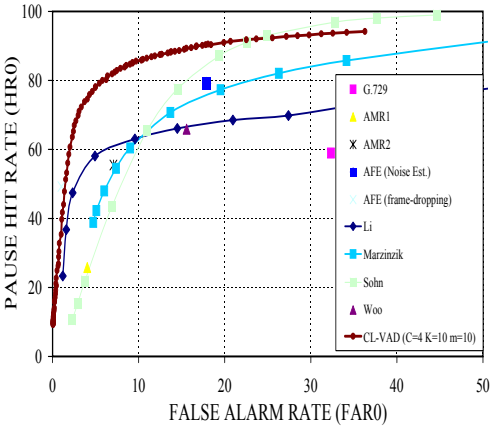

Fig. 3. ROC curves of proposed Cl-VAD in high noisy conditions for $m = 10$, $K = 10$ and $C = 8$ and comparison to standard and recently reported VADs

decision on speech processing systems [8, 1]. The experimental framework and the objective performance tests conducted to evaluate the proposed algorithm are described in this section.

The ROC curves are used in this section for the evaluation of the proposed VAD. These plots describe completely the VAD error rate and show the trade-off between the speech and non-speech error probabilities as the threshold γ varies. The Spanish SpeechDat- Car database [22] was used in the analysis. This database contains recordings in a car environment from close-talking and hands-free microphones. Utterances from the close-talking device with an average SNR of about $25dB$ were labeled as speech or non-speech for reference while the VAD was evaluated on the hands-free microphone. Thus, the speech and non-speech hit rates $(HR1, HR0)$ were determined as a function of the decision threshold γ for each of the VAD tested. Figure 3 shows the ROC curves in the most unfavorable conditions (high-speed, good road) with a 5 dB average SNR. It was shown that increasing the number of observation vectors m improves the performance of the proposed Cl-VAD. The best results are obtained for $m = 10$ while increasing the number of observations over this value reports no additional improvements. The proposed VAD outperforms the Sohn's VAD [7], which assumes a single observation likelihood ratio test (LRT) in the decision rule together with an HMM-based hangover mechanism, as well as standardized VADs such as G.729 and AMR [4, 3]. It also improve recently reported methods [7, 10, 9, 11]. Thus, the proposed VAD works with improved speech/non-speech hit rates when compared to the most relevant algorithms to date.

5 Conclusions

A new VAD for improving speech detection robustness in noisy environments is proposed. The proposed Cl-VAD is based on noise modeling using **C**-means

clustering and benefits from long term information for the formulation of a soft decision rule. The proposed Cl-VAD outperformed Sohn's VAD, that defines the LRT on a single observation, and other methods including the standardized G.729, AMR and AFE VADs, in addition to recently reported VADs. The VAD performs an advanced detection of beginnings and delayed detection of word endings which, in part, avoids having to include additional hangover schemes or noise reduction blocks. Obviously it also will improve the recognition rate when it is considered as part of a complete speech recognition system. The experimental work on this part is on the way. In addition a soft decision based clustering approach for modeling noise prototypes and decision function is currently on progress.

Acknowledgements

This work has received research funding from the EU 6th Framework Programme, under contract number IST-2002-507943 (HIWIRE, Human Input that Works in Real Environments) and SESIBONN and SR3-VoIP projects (TEC2004-06096-C03-00, TEC2004-03829/TCM) from the Spanish government. The views expressed here are those of the authors only. The Community is not liable for any use that may be made of the information contained therein.

References

1. L. Karray and A. Martin, Towards improving speech detection robustness for speech recognition in adverse environments, 2003, Speech Communitation, number 3,pages 261-276.
2. J. Ramírez, J. C. Segura, M. C. Benítez, A. de la Torre and A. Rubio, A New Adaptive Long-Term Spectral Estimation Voice Activity Detector,Proc. of EUROSPEECH 2003, 2003, Geneva, Switzerland, September, pages 3041-3044.
3. ETSI, Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels, 1999, ETSI EN 301 708 Recommendation.
4. ITU, A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70, 1996, ITU-T Recommendation G.729-Annex B.
5. A. Sangwan, M. C. Chiranth, H. S. Jamadagni, R. Sah, R. V. Prasad, V. Gaurav, VAD Techniques for Real-Time Speech Transmission on the Internet, IEEE International Conference on High-Speed Networks and Multimedia Communications, 2002, pages 46-50.
6. F. Basbug, K. Swaminathan and S. Nandkumar, Noise Reduction and Echo Cancellation Front-End for Speech Codecs, 2003, IEEE Transactions on Speech and Audio Processing, 11, num 1, pages 1-13.
7. J. Sohn, N. S. Kim and W. Sung, A statistical model-based voice activity detection, 1999, IEEE Signal Processing Letters,vol 16,num 1, pages 1-3,.
8. R. L. Bouquin-Jeannes and G. Faucon, Study of a voice activity detector and its influence on a noise reduction system, 1995, Speech Communication, vol 16, pages 245-254.
9. K. Woo, T. Yang, K. Park and C. Lee, Robust voice activity detection algorithm for estimating noise spectrum, 2000, Electronics Letters, vol 36, num 2, pages 180-181.

10. Q. Li, J. Zheng, A. Tsai and Q. Zhou, Robust endpoint detection and energy normalization for real-time speech and speaker recognition, 2002, IEEE Transactions on Speech and Audio Processing, vol 10, num 3, pages 146-157.
11. M. Marzinzik and B. Kollmeier, Speech pause detection for noise spectrum estimation by tracking power envelope dynamics, 2002, IEEE Transactions on Speech and Audio Processing, vol 10, num 6, pages 341-351.
12. Fisher, D. 1987. Knowledge acquisition via incremental conceptual clustering. Machine Learning 2:139–172.
13. J.M. Górriz, J. Ramírez, J.C. Segura and C.G. Puntonet, Improved MO-LRT VAD based on bispectra Gaussian model, 2005, Electronics Letters, vol 41, num 15, pages 877-879.
14. J. Ramírez, José C. Segura, C. Benítez, L. García and A. Rubio, Statistical Voice Activity Detection using a Multiple Observation Likelihood Ratio Test, 2005, IEEE Signal Processing Letters, vol 12, num 10, pages 689-692.
15. Anderberg, M. R. 1973. Cluster Analysis for Applications. Academic Press, Inc., New York, NY.
16. Jain, A. K. and Flynn, P. J. 1996. Image segmentation using clustering. In Advances in Image Understanding. A Festschrift for Azriel Rosenfeld, N. Ahuja and K. Bowyer, Eds, IEEE Press, Piscataway, NJ, 65-83.
17. Rasmussen, E. 1992. Clustering algorithms. In Information Retrieval: Data Structures and Algorithms, W. B. Frakes and R. Baeza-Yates, Eds. Prentice-Hall, Inc., Upper Saddle River, NJ, 419-442
18. Salton, G. 1991. Developments in automatic text retrieval. Science 253, 974-980.
19. J. Ramírez, José C. Segura, C. Benítez, A. de la Torre, A. Rubio, An Effective Subband OSF-based VAD with Noise Reduction for Robust Speech Recognition, 2005, In press IEEE Trans. on Speech and Audio Processing.
20. J. B. MacQueen, Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297 (1967)
21. T. Hastie, R. Tibshirani and J. Friedman The Elements of Statistical Learning Data Mining, Inference, and Prediction Series: Springer Series in Statistics 1st ed. 2001. ISBN: 0-387-95284-5
22. A. Moreno, L. Borge, D. Christoph, R. Gael, C. Khalid, E. Stephan and A. Jeffrey, SpeechDat-Car: A Large Speech Database for Automotive Environments, Proceedings of the II LREC Conference, 2000.