Noise Subspace Fuzzy C-Means Clustering for Robust Speech Recognition

J.M. Górriz[1], J. Ramírez[1], J.C. Segura[1],
C.G. Puntonet[2], and J.J. González[2]

[1] Dpt. Signal Theory, Networking and communications,
University of Granada, Spain
gorriz@ugr.es
http://www.ugr.es/~gorriz
[2] Dpt. Computer Architecture and Technology,
University of Granada, Spain

Abstract. In this paper a fuzzy **C**-means (FCM) based approach for speech/non-speech discrimination is developed to build an effective voice activity detection (VAD) algorithm. The proposed VAD method is based on a soft-decision clustering approach built over a ratio of subband energies that improves recognition performance in noisy environments. The accuracy of the FCM-VAD algorithm lies in the use of a decision function defined over a multiple-observation (MO) window of averaged subband energy ratio and the modeling of noise subspace into fuzzy 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, discontinuous transmission, real-time speech transmission on the Internet or combined noise reduction and echo cancelation schemes in the context of telephony [1, 2]. 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 strategies for detecting speech on a noisy signal [3] and have evaluated the influence of the VAD effectiveness on the performance of speech processing systems [4]. Most of them have focussed on the development of robust algorithms with special attention on

M. Gavrilova et al. (Eds.): ICCSA 2006, LNCS 3984, pp. 772–779, 2006.

the derivation and study of noise robust features and decision rules [5, 6, 7, 3]. 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, image processing and information retrieval [9, 10]. Clustering has a rich history in other disciplines [11] 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.

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 $\omega_s = \frac{2\pi \cdot s}{N_{FFT}}$, $0 \le s \le N_{FFT} - 1$ 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 energies for the l-th frame, $E(k, l)$, in K subbands $(k = 0, 1, ..., K - 1)$, are computed by means of:

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

where an equally spaced subband assignment is used and $\lfloor \cdot \rfloor$ denotes the "floor" function. Hence, the signal energy is averaged over K subbands obtaining a suitable representation of the input signal for VAD [12], the observation vector at frame l, $\mathbf{E}(l) = (E(0, l), \dots, E(K - 1, l))^T$. The VAD decision rule is formulated over a sliding multiple observation (MO) window consisting of $2m+1$ observation vectors 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 [8] etc.

3 FCM Clustering over the Observation Vectors

CM clustering is a method for finding clusters and cluster centers in a set of unlabeled data. The number of cluster centers (prototypes) C is a priori known and the CM iteratively moves the centers to minimize the total within cluster variance. Given an initial set of centers the CM algorithm alternates two steps: a) for each cluster we identify the subset of training points (its cluster) that is closer to it than any other center; b) 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.

This previous clustering technique is referred to as hard or crisp clustering, which means that each individual is assigned to only one cluster. For FCM clustering, this restriction is relaxed, and the object can belong to all of the clusters with a certain degree of membership. This is particularly useful when the boundaries among the clusters are not well separated and ambiguous.

3.1 Noise Modeling

FCM is one of the most popular fuzzy clustering algorithms. FCM can be regarded as a generalization of ISODATA [13] and was realized by Bezdek [14]. In our algorithm, the fuzzy approach is applied to a set of N initial pause frames (energies) in order to characterize the noise space. From this energy noise space we obtain a set of clusters, namely noise prototypes [1]. The process is as the following: each observation vector ($\mathbf{E}$ from equation 3) is uniquely labeled, by the integer $i \in \{1, \ldots, N\}$, and 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:

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

FCM attempts to find a partition (fuzzy prototypes) for a set of data points $\mathbf{E}_i \in \mathcal{R}^K$, $i = 1, \ldots, N$ while minimizing the cost function

$$J(\mathbf{U}, \mathbf{M}) = \sum_{i=1}^{C} \sum_{j=1}^{N} (u_{ij})^m D_{ij} \tag{5}$$

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

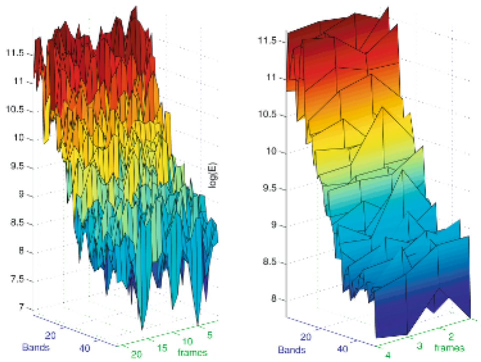

Fig. 1. a) 20 log Energies of noise frames, computed using $N_{FFT} = 256$, averaged over 50 subbands. b) Clustering approach applied to the a set of log-energies using hard decision CM (C=4 prototypes).

where $\mathbf{U} = [u_{ij}]_{C \times N}$ is the fuzzy partition matrix, $u_{ij} \in (0, 1)$ is the membership coefficient of the j-th individual in the i-th prototype; $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_C]$ denotes the cluster prototype (center) matrix, $m \in [1, \infty)$ is the fuzzification parameter (set to 2) and $D_{ij} = d(\mathbf{E}_j, \mathbf{m}_i)$ is the distance measure between $\mathbf{E}_j$ and $\mathbf{m}_i$.

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

3.2 Soft Decision Function for VAD

In order to classify the second labeled data (energies of speech frames) we use a sequential algorithm scheme 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:

$$\hat{\mathbf{E}}(l) = \max\{\mathbf{E}(i)\}, \quad i = l - m, \ldots, l + m \tag{6}$$

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 [7, 6] in speech-pause transition).

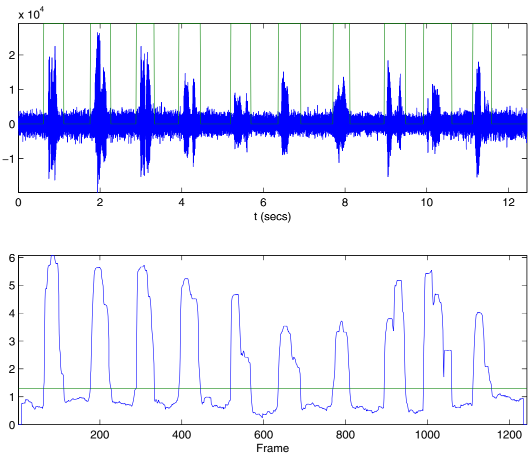

Fig. 2. VAD operation: Top- Decision function and threshold versus frames. Bottom- Input signal and VAD decision versus time.

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

$$F(l) = \log\left(1/K \sum_{k=0}^{K-1} \frac{\hat{E}(k,l)}{<\bar{\mathbf{m}}_c>}\right) > \gamma \tag{7}$$

where $< \bar{\mathbf{m}}_c >$ is the averaged noise prototype center and γ is the decision threshold.

The set of noise prototypes are updated in pause frames (not satisfying equation 7)) using the adaptation of the standard FCM, replacing the oldest energy in the noise model, consisting of N samples, by the actual feature vector $\hat{\mathbf{E}}(l)$. The initial prototype matrix $\mathbf{M}(l)$ at decision frame l is the previous one $\mathbf{M}(l-1)$, and the following update is applied to the fuzzy partition and prototype center matrices:

$$\begin{aligned}
u_{ij}^{(t+1)} &= 1/\left(\sum_{l=1}^{C}(D_{lj}/D_{ij})^{1/(1-m)}\right)\\
\mathbf{m}_i^{(t+1)} &= \left(\sum_{j=1}^{N}\left(u_{ij}^{(t+1)}\right)^m \mathbf{E}_j\right)/\left(\sum_{j=1}^{N}\left(u_{ij}^{(t+1)}\right)^m\right)\\
&\mathit{until} \quad ||\mathbf{M}^{(t+1)}(l) - \mathbf{M}^{(t)}(l)|| < \epsilon\\
&\mathit{for} \quad i = 1,\ldots,C, \quad j = 1,\ldots,N
\end{aligned} \tag{8}$$

This sequential adaptation doesn't involve high computational effort although other kind of static adaptation rules could be applied. The algorithm described so far is presented as pseudo-code in the following:

1. Initialize Noise Model:
   - Select N feature vectors $\{\mathbf{E}(i)\}, i = 1,\ldots,N$.
   - Compute threshold γ.

2. Apply FCM clustering to feature vectors extracting C noise prototype centers $\{\mathbf{m}(c)\}, c = 1, \ldots, C$
3. for l=init to end
   (a) Compute $\hat{\mathbf{E}}(l)$ over the MO window
   (b) if equation 7 holds then VAD=1
   else Update noise prototype centers $\mathbf{m}(c)$ with equations 8.

Figure 2 shows the operation of the proposed FCM-VAD on an utterance of the Spanish SpeechDat-Car (SDC) database [15]. The phonetic transcription is: "tres", "nueve","zero", "siete", "μinko", "dos", "uno", "otSo", "seis","cuatro". We also show the soft decision function and the selected threshold in the FCM-VAD operation for the same phrase.

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 [7], and the influence of the VAD decision on speech processing systems [4]. 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 [15] 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 can be shown that increasing the number of observation vectors m improves the performance of the proposed FCM-VAD. The best results are obtained for $m = 8$ while increasing the number of observations over this value reports no additional improvements. The proposed VAD outperforms the Sohn's VAD [3], 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 [2, 1]. It also improve recently reported methods [3, 6, 5, 7]. Thus, the proposed VAD works with improved speech/non-speech hit rates when compared to the most relevant algorithms to date. Table 1 shows the recognition performance for the Spanish SDC database for the different training/test mismatch conditions (HM, high mismatch, MM: medium mismatch and WM: well matched) when WF and FD are performed on the base system [8]. The VAD outperforms all the algorithms used for reference, yielding relevant improvements in speech recognition.

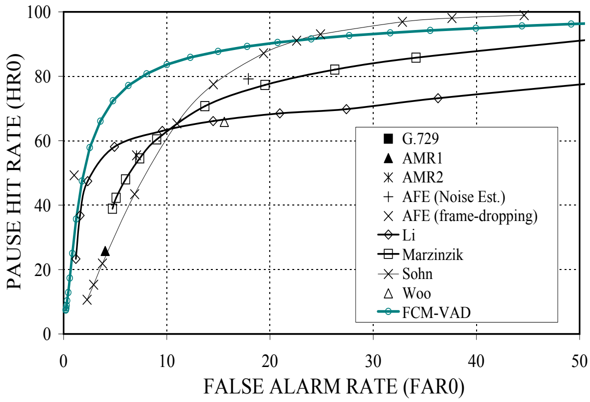

Fig. 3. ROC curves of proposed FCM-VAD in high noisy conditions for $m = 8$, $K = 32$ and $C = 2$ and comparison to standard and recently reported VADs

Table 1. Average word accuracy (%) for the Spanish SDC database

	Base	G.729	AMR1	AMR2	AFE
WM	92.94	88.62	94.65	95.67	95.28
MM	83.31	72.84	80.59	90.91	90.23
HM	51.55	65.50	62.41	85.77	77.53
Average	75.93	75.65	74.33	90.78	87.68
	Woo	Li	Marzinzik	Sohn	FCM-VAD
WM	95.35	91.82	94.29	96.07	**96.68**
MM	89.30	77.45	89.81	91.64	**91.82**
HM	83.64	78.52	79.43	84.03	**86.05**
Average	89.43	82.60	87.84	90.58	**91.51**

5 Conclusions

A new VAD for improving speech detection robustness in noisy environments is proposed. The proposed FCM-VAD is based on noise modeling using FCM clustering and benefits from long term information for the formulation of a soft decision rule. The proposed FCM-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 discrimination analysis or the ROC curves are effective to evaluate a given algorithm, the influence of the VAD in a speech recognition system depends on its discrimination accuracy [12]. Thus the proposed VAD improves the recognition rate when it is used as a part of a Automated Speech Recognition (ASR) system.

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. ETSI, Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels, 1999, ETSI EN 301 708 Recommendation.
2. ITU, A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70, 1996, ITU-T Recommendation G.729-Annex B.
3. 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,.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. Anderberg, M. R. 1973. Cluster Analysis for Applications. Academic Press, Inc., New York, NY.
10. 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
11. Jain, A. K. and Dubes, R. C. 1988. Algorithms for Clustering Data. Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ.
12. 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.
13. J. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters, J. Cybern., vol. 3, no. 3, pp. 32-57, 1974.
14. J. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum, 1981.
15. 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.