Crossroads
The ACM Student Magazine

Voice Activity Detection

by ***Deepti Singh*** and ***Frank Boland***

Voice activity detection (VAD) is used to detect the presence of speech in an audio signal. VAD plays an important role as a preprocessing stage in numerous audio processing applications. For example, in voice over IP (VoIP) and mobile telephony applications, VAD can reduce bandwidth usage and network traffic by transmitting audio packets only if speech is detected. Furthermore, the performance of speech recognition, speaker recognition, and source localization can be improved by applying these algorithms only to parts of the audio that are identified as speech. Video conferencing is another prominent source localization application where VAD is beneficial. In such an application, source localization is performed, and the video camera is steered in the direction of the audio source when speech is detected using VAD.

VAD is also used to identify noise so that it can be suppressed in systems such as hearing aids and audio conferencing devices. Considering all the potential applications of VAD, it would be useful to understand the challenges in designing a VAD system. This article discusses the key steps and the associated challenges in designing a VAD system.

Features and Classes

Figure **1** presents a block diagram of the structure of a typical VAD system. VAD is a classification problem in which the features, or characteristics, of an audio signal are used to separate it into different classes. In the classification process, the audio signal is divided into small fixed-length frames. The values of the features are then calculated for each frame and passed as input to the classification algorithm. Frames are typically about twenty milliseconds long.

The number and types of classes into which the audio signal is classified depends on the application's requirements. For example, if an application must distinguish between speech and other types of signals, as required in VoIP applications, then the audio signal can be classified into two classes: speech and non-speech.

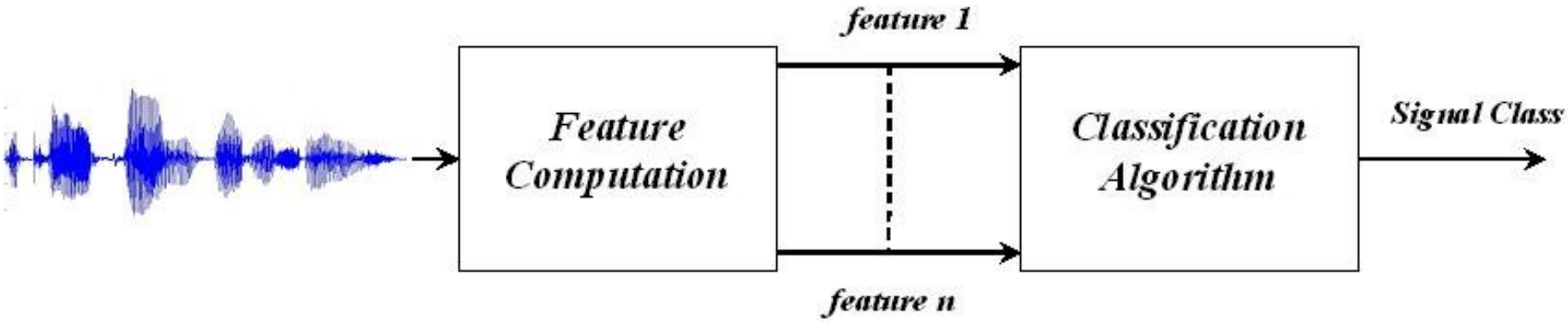

Figure 1. The structure of a VAD system.

Speech can be further classified into voiced and unvoiced speech. Voiced speech is usually a deterministic waveform consisting of vowels and some consonants. Unvoiced speech is generally a stochastic waveform consisting of consonants. This classification is useful in source localization applications in which the source localization algorithm is applied only to voiced sections of speech.

The features needed for classification depend on the class requirements of the application. The short-term energy of the signal, computed as the sum of the squares of the amplitude of the signal samples in a frame, is one of the signal features useful in VAD [**2**,**8**]. Another feature that can be used for VAD is a measure of zero crossings. This feature represents the number of times a signal has changed its sign within the frame. Different types of audio signals have different zero crossing measurements. Generally, voiced speech has lower zero crossing measurements than unvoiced signals. Short-term energy, along with zero crossing, distinguishes between voiced and unvoiced parts of speech.

Voiced speech signals are usually low frequency signals, and unvoiced speech signals generally are high frequency signals. It is expected that voiced speech will have a higher proportion of its energy in a low frequency band (less than 2 kHz), whereas unvoiced signals will have a more uniform distribution of energy. The ratio of low frequency band to full band energy distinguishes between voiced and unvoiced speech signals.

Features related to linear predictive coding (LPC) are also employed for VAD. LPC predictor coefficients are obtained from the LPC analysis of an input signal. These coefficients represent the system model and can be used to reconstruct the original signal. LPC coefficients are useful in classification because for speech signals, the first few LPC predictor coefficients contain most of the information about the signal. A feature related to LPC is the LPC residual. It can be defined as the difference between the reconstructed signal using the LPC predictor coefficients and the original input signal.

The accuracy with which the LPC is able to predict a signal is a measure of its randomness. As a result the LPC residual for voiced speech is lower compared to that for unvoiced speech [**5**]. Thus, this feature helps to distinguish between voiced speech and signals with a more random waveform. Higher order statistics like skewness and kurtosis of the LPC residual are also adopted for classification [**7**]. Gaussian signals are characterized by their mean and variance. They do not have higher order statistics. Speech signals are non-Gaussian and hence higher order statistics are utilized to distinguish between speech and Gaussian noise.

Feature Selection

A major challenge in designing a VAD system is the process of identifying the features that can effectively classify the audio into different classes. A good feature set for VAD will ensure that there is little, if any, overlap between the classes. Features that are strongly related are said to be highly correlated. The features for classification should not be highly correlated as they may adversely affect each others' contribution to the classification process. Feature selection algorithms can be used to select an optimal feature subset from a given feature set. The following algorithms identify the features that do not make a significant contribution towards classification.

Fisher's F ratio [**8**] is one of the most commonly used feature selection algorithms. F ratio can be applied on only one feature at a time to determine its significance for classification. If the features are uncorrelated (which is not always the case), F ratio is applied on each feature to assess its significance towards classification and the most useful features are selected for VAD. A drawback of this algorithm is that it cannot detect overlap between classes. If the features are correlated, then a variation of the F ratio algorithm is applied to all the features together to estimate their significance.

The optimal feature subset for classification can be determined by applying the feature selection algorithm on all the feature subsets and then selecting the subset that makes the most significant contribution towards classification [**1**,**4**]. This seems simple, but when the feature set is large this approach can be computationally demanding, and impractical to apply. The forward selection or backward selection approaches are compromising solutions for this problem.

The forward selection method starts with the assumption that none of the features are important. In the first iteration, it considers the contribution of each feature towards classification and selects the feature that helps the classifier to perform the best [1,4]. In the successive iterations it takes the selected feature with the rest of the features one at a time, and selects the combination that contributes the most towards the performance of the classifier. This procedure is repeated until either the addition of another feature has insignificant or no improvement in the classifier's performance or a certain criteria is satisfied.

Backward selection is the reverse of forward selection. It starts with the assumption that all the features in the feature set are important. The first iteration eliminates the feature with the least contribution to the classification. Successive iterations continue to discard those features whose exclusion produces no (or insignificant) improvement in the classifier's performance.

Usually both forward selection and backward selection return the same feature subset. Both approaches are greedy search algorithms, as they do not consider all feature subsets. The significance of the features is not questioned in the successive iterations once they are found to be significant in forward selection. Similarly in backward selection, once a feature is found to be insignificant it is eliminated, and is not considered for significance testing in further iterations. Thus, the feature set obtained is not necessarily optimal.

Feature Extraction

Feature extraction is another technique that is applied to find suitable features for classification. While feature selection selects a feature subset from the feature set, feature extraction methods reduce the dimensionality of the input data by forming new features that are a combination of two or more original features from the feature set. This is extremely useful in cases where nonoverlapping boundaries of the classes cannot be obtained using the original features. The features can be transformed into a new feature space such that separation between the classes can be obtained with the new features.

Principal component analysis is the simplest and most widely used method for feature extraction. It performs linear transforms on the input features. Discriminant analysis, discussed below, is another example of this category.

The advantage of feature extraction is that it reduces the storage space for the features. The drawback of feature extraction is increased computation requirements. The computations are increased because in addition to the transformed features, all the original features might also have to be calculated.

Classification Algorithms

Aside from deciding on the features for classification, the choice of the classification algorithm is also crucial in designing a VAD system. The existing algorithms for classification used in VAD can be broadly divided into two major categories: user-defined threshold based algorithms and machine learning algorithms.

The user-defined threshold based algorithms compute the threshold values of one or more features to come up with the classification decision. These threshold values are obtained by applying some user-imposed rules. The advantage of this approach is that the user is in full control of the decision function obtained for classification because the user decides the rules that are used to compute the threshold values. The major drawback of this approach is that finding the threshold values can be difficult, especially when the feature set is large.

On the other hand, supervised machine learning algorithms use data with known classes (training data) to construct a decision function. This decision function classifies an audio signal into one of the predefined classes. Some of the supervised machine learning algorithms used for VAD include discriminant analysis (DA) [**5**], artificial neural networks (ANN) [**9**,**10**], and support vector machines (SVM) [**1**,**3**]. DA, SVM, and ANN are applied for feature selection as well.

A parametric classifier assumes prior knowledge of the data distribution whereas a nonparametric classifier does not assume prior information about the data distribution. Discriminant analysis is a parametric classifier, and assumes that the distribution of the data is normal (Gaussian). If the data is not normally distributed, DA's performance is still robust, provided the data set does not contain important outliers. The Mahalanobis distance classifier (MDC) is the simplest of the DA classifiers. Mahalanobis distance (MD) is similar to Euclidean distance, and is a measure of the distance of the data from the centroid of the classes. In MDC the input data is assigned to the class that has the minimum MD value. The centroid of the classes is obtained from the training data. Other examples of DA classifiers are linear discriminant analysis and quadratic discriminant analysis.

Unlike DA, SVM and ANN are nonparametric classifiers. ANNs consist of computational units called neurons that can take multiple input signals and generate an output of 1 if the weighed sum of the input signals is above some threshold value. ANN can be represented using a directed graph in which the vertices represent neurons and the edges denote connections between the neurons. There exists algorithms that can use the training data to decide on the threshold values for the neurons in the ANN.

SVM uses training data to find an optimal hyperplane that separates different classes of the training

data. The optimal hyperplane maximizes the distance between itself and the training data of the classes that it separates. This improves the generalization ability of the classifier. The hyperplane is then used to classify the input data of an unknown class into their respective classes. If the training data is not linearly separable, then the training data samples are transformed to a higher dimensional space to achieve linear separation. It should be noted that it is also possible to avoid data transformation in this situation by working on the input data itself by employing special functions.

Statistical classification algorithms such as hidden Markov models [**6**] and Gaussian mixture models [**11**] are also suitable for VAD. In the Markov model, the probability that the current frame belongs to a particular class depends on the class of the previous frame. In hidden Markov models, the class of the current frame cannot be observed. The likelihood of the current frame belonging to a particular class depends on the matrix stating the probability of transition from one class to another and the probability of observing feature values for a particular class.

The Gaussian mixture model assumes that the feature vectors have a Gaussian distribution. Gaussian mixture models are used to obtain the probability density for each class based on features of the training data. This probability density estimation involves the weighted sum of a multidimensional Gaussian distribution. This probability density is used to classify the data.

A desired feature of a good classification algorithm is a high hit (correct classification) rate and a low false positive rate. The ability of the classifier to classify unknown data successfully into its respective class is known as generalization. Sometimes the classifiers classify the training data well but do not classify new data properly. This is known as overfitting. Typically, it is desired that the classifier is trained such that it has a good generalization ability and low overfitting.

To assess the generalization ability of a classifier the "leave one out" method can be employed. The classifier trains with all but one data sample of the training data and predicts the class of the data sample that was left out. This process is repeated for all data samples, and the total classification rate that is obtained gives an idea of the classifier's performance.

Another approach is to use another data set that contains data independent from the training data. Comparing the performance of these algorithms on a test data set after the classifiers have been trained using a training data set is the simplest way to choose an appropriate classification algorithm. The hit rate obtained in this case is not the true hit rate. Statistical significance tests can be performed to obtain the "chance" hit rate.

Table 1. The key steps and associated challenges in designing a VAD system.

	Step	Challenges
1.	Select feature set and apply feature selection or feature extraction	• No overlap between groups • Uncorrelated • No insignificant features
2.	Select the classification algorithm	• High hit rate • Low false alarm rate
3.	Select the training data	• Cover most signal scenarios • Good generalization • Low overfitting

Selecting the appropriate training data is another important factor. In SVM algorithms the choice of the training data affects whether the decision function will be able to classify unknown data effectively. To achieve this it is recommended to use a large number of inputs in the training data. The training data should cover most of the signal scenarios in which the VAD system will operate [9].

Conclusion

This article presents the steps and associated challenges in designing a VAD system (see Table 1). Classification features, feature selection, feature extraction, and classification algorithms are discussed. The choice of features and classification algorithm for VAD depends on the problems associated with the environment for which the VAD system is being designed, the number of classes the signal is being classified into, and the extent to which the user is willing to compromise in terms of the individual hit/false positive rates and the overall hit rate. This information is directly relevant for anyone interested in VAD and could be helpful in designing a classifier for any purpose.

References

1 Abe, S., *Support Vector Machines for Pattern Classification*, Springer-Verlag London Limited, 2005.

2 Deller, J.R.; Proakis, J.G. & Hansen, J.H.L., *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Company, 1993.

3

Enqing, D.; Guizhong, L.; Yatong, Z. & Xiaodi, Z., *Applying support vector machines to voice activity detection*, Proceedings of the sixth International Conference on Signal Processing, 2002, pg. 1124- 1127.

4

Guyon, I. & Elisseeff, A., *An Introduction to Variable and Feature Selection*, Journal of Machine Learning Research, 3 (2003), 157-1182.

5

Huberty, C.J., *Applied Discriminant Analysis*, John Wiley and Sons Inc., 1994.

6

Lawrence R. Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceedings of the IEEE, 77, 2 (1989), pg. 257-286.

7

Nemer, E.; Goubran, R. & Mahmoud, S., *Robust voice activity detection using higher-order statistics in the LPC residual domain*, IEEE Transactions on Speech and Audio Processing, 9, 3 (2001), pg. 217-231.

8

Parsons, T.W., *Voice and Speech Processing*, McGraw-Hill Inc., 1987.

9

Principe, J.C.; Euliano, N.R. & Lefebvre, W.C., *Neural and Adaptive Systems: Fundamentals through Simulations*, John Wiley and Sons Inc., 2000.

10

Shao, C. Bouchard, M., *Efficient classification of noisy speech using neural networks*, Proceedings of the Seventh International Symposium on Signal Processing and Its Applications, 2002, pg. 357-360.

11

Wrigley, S.N.; Brown, G.J.; Wan, V. & Renals, S., *Speech and crosstalk detection in multichannel audio*, IEEE Transactions on Speech and Audio Processing, 13, 1 (2005), pg. 84-91.

Biography

Deepti Singh is a PhD student in the Department of Electronic and Electrical Engineering at Trinity College, Dublin, Ireland.

Frank Boland is a professor in the Department of Electronic and Electrical Engineering at Trinity College, Dublin, Ireland.