SVM-based Voice Activity Detection for Distributed Speech Recognition System

Azzedine Touazi¹ and Mohamed Debyèche¹²

¹Université des Sciences et de la Technologie Houari Boumediene, Faculté d’Électronique et d’Informatique, Laboratoire de Communication Parlée et de Traitement du Signal (LCPTS), El Alia, Bab Ezzouar, 16111, Algiers, ALGERIA
e-mail: {touazi.azzedine, mdebyèche}@gmail.com, mohamed.debyèche@enp.edu.dz

Abstract—Voice Activity Detection (VAD) algorithms based on machine learning techniques have shown competitive results in the area of automatic speech recognition. This paper describes a new approach of VAD based on Support Vector Machines (SVM) for Distributed Speech Recognition (DSR) system. In the proposed scheme, the speech and the non-speech frames are detected from the compressed Mel Frequency Cepstral Coefficients (MFCCs), at the back-end (e.g. server) side, with the aim of improving the VAD performance and reducing the compression bit-rate from the front-end side (e.g. client). By using the trained SVM with polynomial kernel, the SVM-based VAD show an encouraging detection results. The classification task conducted from the Aurora-2 speech database, using different noise conditions, illustrates comparable VAD performance, with respect to ETSI Advanced Front-End (ETSI-AFE) VAD algorithm.

Keywords- voice activity detection; support vector machines; DSR system; mel frequency cepstral coefficients

I. INTRODUCTION

Nowadays, the development of speech technologies over mobile devices, and the requirement of improved speech recognition performance, has led to Distributed Speech Recognition (DSR) systems [1], where the processing is distributed between the terminal and the server. In DSR architecture, the basic idea consists of using a local front-end terminal, from which the Mel Frequency Cepstral Coefficients (MFCCs) feature vectors are extracted and then transmitted through an error protected data channel to a remote back-end recognition server (see Figure 1). Voice activity detection (VAD), which plays an important role in DSR systems, is a technique used to detect automatically the speech and non-speech frames from a speech signal. This can considerably reduce the number of insertion errors from the recognition task at the back-end side.

The traditional VAD approaches include those based fundamentally on energy levels [2], Linear Prediction Coding (LPC) parameters [3], zero crossing rate [4], spectrum analysis [5], and periodicity measure [6], etc. Most of these traditional approaches fail when the level of background noise increases. In case of the European Telecommunications Standards Institute Advanced Front-End (ETSI-AFE) DSR standard [7], the process of detecting voice activity, based on signal energy, is implemented into two parts, (i) one VAD is used for noise estimation in the high-frequency bands and (ii) a second VAD is designed for non-speech frame dropping.

With the progress of speech recognition systems, the machine learning-based VAD techniques are getting more attention during the last decade, e.g. [8-14]. The most of these techniques are based on Support Vector Machines (SVM) Classification. The machine learning approaches are very competitive to traditional ones, by the fact that they can consider the interaction between features and their discriminative power in order to better detect the speech frames, especially in case of noisy environments.

Figure 1. ETSI-AFE DSR Architecture.
This paper pays attention to the detection of speech and non-speech frame dropping from MFCC coefficients using SVM classification. Where, the VAD is performed from the back-end side for a DSR system. This approach is motivated by two main advantages: (i) It can reduce the compression bit-rate for a DSR system (i.e. 4.3 kbps rather than 4.4 kbps) by the fact that the VAD information will not be transmitted and (ii) It can improve the VAD performance by designing a noise robust machine learning-based VAD from MFCCs.

We have conducted several experiments from Aurora-2 [15] database, and the Library for Support Vector Machines (LIBSVM) [16]. A comparison of the proposed approach with ETSI-AFE is performed throughout this series of experiments.

The remainder of this paper is structured as follows: A general overview of ETSI DSR standards is presented in Section II. A description of the SVM-based VAD is the object of section III. The experimental set up, results, and discussions are detailed in section IV. Finally, conclusion and further work are laid out in Section V.

II. DSR STANDARDS OVERVIEW

In the conventional DSR ETSI Front-End (ETSI-FE) standard [17] the speech features (i.e., MFCCs), used in the front-end part, are derived from the extracted speech frames at frame length of 25 ms with frame shift of 10 ms, using Hamming windowing. Then, a Fourier transform is performed and followed by Mel filter bank with 23 frequency bands in the range from 64 Hz up to 4 kHz. The extracted coefficients are the first 12 Mel frequency cepstral coefficients (c1, c2,..., c12), the zeroth cepstral coefficient (c0), and the log energy (log E) in each frame.

In the compression task, the 14-dimensional feature vector (c1-c12, c0, and log E), is split equally into seven sub-vectors, and each of them is quantized with its own 2-dimensional vector quantizer using Split Vector Quantization (SVQ) technique. The resulting compression bit-rate is 4.4 kbps, and 4.8 kbps with including channel bit-rate.

ETSI published new DSR version known as the advanced version (ETSI-AFE) [7], which provides considerably improvements in recognition performance, in presence of background noise. In the feature extraction part of the ETSI-AFE standard, noise reduction is performed first, which is based on Wiener filtering theory. Then, waveform processing is applied to the de-noised signal and cepstral features are calculated. VAD for the non-speech frame dropping is also implemented, where the VAD flag is transmitted as single bit to the back-end side. Table I shows the bit allocation used in ETSI-AFE standard.

III. SVM-BASED VAD FOR DSR SYSTEM

As outlined in section I, in this proposed approach, the main idea is the implementation of the SVM classification from the back-end side of a DSR system, in order to detect voice activity. As depicted by Figure 2, the VAD is performed through the transmitted MFCC vectors. We should highlight that we take into account only the effect of source coding of MFCCs (i.e. quantized MFCCs). Therefore, for each frame of 10 ms, the decompressed feature vector (c1-c12, c0, and log E) is recognized as speech or non-speech frame from the trained SVM model. On the other hand, The SVM-based VAD can be viewed as a binary-class classification problem.

![Figure 2. Proposed SVM-based approach.](image)

| Table I. Bits allocation in ETSI-AFE at 4.4 kbps |
|-----------------|-----------------|
| MFCC sub-vector | Bits allocation |
| (c1, c2)        | 6               |
| (c3, c4)        | 6               |
| (c5, c6)        | 6               |
| (c7, c8)        | 6               |
| (c9, c10)       | 6               |
| (c11, c12)      | 5               |
| (c0, log E)     | 8               |
| VAD             | 1               |

IV. EXPERIMENTAL FRAMEWORK

Several experiments from Aurora-2 are conducted. Aurora-2 database provides speech samples and scripts to perform speaker independent speech recognition experiments in clean and multi-condition training modes. This database has been prepared by down-sampling the speech samples to 8 kHz, filtering with the G.712 and MIRS characteristics; noise is artificially added to the filtered speech utterances at a desired Signal to Noise Ratio (SNR) (20 to -5dB) with including clean condition, and eight different noise conditions, such as subway, babble, car, exhibition hall, restaurant, street, airport, and train station. There are three training test sets (test set A, B, and C), and two training modes, such as clean and multi-condition. For further details, a full description of Aurora-2 is given in [15].

In this paper, 300 noisy speech utterances, in the range from 0 to 20 dB, selected from Aurora-2 database are applied to SVM-based VAD training and testing. In which 150 sentences, taken from (test set A, B, and C), are used for training and the remaining ones are applied for testing. For frame labeling, we consider a VAD flag of each 10 ms ETSI-AFE clean speech frame as a reference decision. To simulate various noisy environments, four different types of noise (subway, babble, restaurant, and street) are considered.

The experimental part is conducted from the Library for SVM (LIBSVM) [16] using different types of kernels, such as, linear, polynomial (with different degrees $d=1,...,7$), and Radial Basis Function (RBF) kernels. In order to identify the optimal values of the cost and the kernel parameters ($C$, $\gamma$), respectively, a grid-search [18] on $C$ and $\gamma$ is applied to the training set. Various pairs of ($C$, $\gamma$) values in which $C=[2^5, 2^4, ..., 2^1]$ and $\gamma=[2^5, 2^4, ..., 2^{10}]$ are tested using fivefold cross-validation. Then, the best parameters set ($C$, $\gamma$) are used for accurately predicting the testing set.

**Figure 3.** Overall statistical measures, A: (Uncompressed MFCCs), B: (Compressed MFCCs).

A. Evaluation Criteria

We compared the performance of SVM-based VAD to ETSI-AFE standard in terms of statistical measures such as, sensitivity (true positive rate), specificity (true negative rate), and accuracy, which are described as follows:

\[
\text{Sensitivity}(\%) = \frac{TP}{(TP + FN)} \times 100, \quad (1)
\]

\[
\text{Specificity}(\%) = \frac{TN}{(TN + FP)} \times 100, \quad (2)
\]

\[
\text{Accuracy}(\%) = \frac{(TN + TP)}{TN + TP + FN + FP} \times 100, \quad (3)
\]

where $TP$, $TN$, $FN$, $FP$ are defined as the true positive, true negative, false negative, and false positive, respectively.

B. Results and Discussions

Figure 3 shows the obtained statistical measures for both cases, (i) detecting voice activity from uncompressed MFCCs (i.e. front-end side), and (ii) detecting voice activity from compressed MFCCs (i.e. back-end side). According to the results obtained, it can be seen that the polynomial kernels provide more improvements and a good compromise in terms of statistical measures, comparing to RBF and linear ($d=1$) kernels. We can also show that there is no significant performance degradation in case of detecting voice activity from the compressed MFCCs.
It is worth noting that the proposed approach provides competitive VAD performance at different noise levels. However, one may need to enhance the SVM-based classifier, or to use the existing machine learning techniques that can provide more information about the discriminative power of the MFCC features.

As it can be shown from the bit allocation in Table I, in case of detecting voice activity from the back-end side, we can achieve 2.3% of bandwidth reduction (i.e. 4.3 kbps rather than 4.4 kbps).

From the complexity point of view, we should highlight that the proposed approach can only reduce the computational cost in the compression task (i.e. coding the VAD flag). As in the case of ETSI-AFE, the VAD information is required for noise reduction (or even for feature extraction).

V. CONCLUSION

We have presented a case study of detecting voice activity for DSR system. In which the main idea is to estimate the VAD, at the back-end side, using machine learning techniques. Experiments have been conducted from Aurora-2 database by considering different noise conditions and SNR levels. The proposed machine learning approach when compared with the traditional energy-based ETSI-AFE, gives comparable statistical performance measures. From the experimental part, it shown that the SVM-based VAD approach applied to DSR systems holds promise and should be further explored.

REFERENCES