Evaluation of Voice Activity and Voicing Detection

Bojan Kotnik¹, Pierre Sendorek², Sergey Astrov³, Turgay Koc⁴, Tolga Ciloglu⁴, Laura Docio Fernandez⁵, Eduardo Rodriguez Banga⁵, Harald Hoge³, Zdravko Kacic¹

¹ Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia
² Department TSI Signal-Images, Telecom Paristech, France
³ Corporate Technology, Siemens AG, Munich, Germany
⁴ Electrical and Electronics Eng. Dept, Middle East Technical University, Ankara, Turkey
⁵ Signal Processing and Communications Dept., University of Vigo, Spain
bojan.kotnik@uni-mb.si, sergey.astrov@siemens.com

Abstract
This paper describes the ECESS evaluation campaign of voice activity and voicing detection. Standard VAD classifies signal into speech and non-speech, we extend it to VAD+ so that it classifies a signal as a sequence of non-speech, voiced and unvoiced segments. The evaluation is performed on a portion of the Spanish SPEECON database with manually labeled segmentation. To avoid errors caused by the limited precision of manual labeling we introduce “dead zones” – tolerance intervals ±5 ms around label changes in the data set. In these tolerance intervals we don’t evaluate the signal.

Index Terms: VAD, voicing detection, evaluation

1. Introduction
Voice activity detection (VAD) distinguishes between speech and non-speech segments in a given audio signal. VAD is widely used in modern speech processing and communication systems for different purposes: estimation and detection of speech signals, speech recognition, real time speech transmission, noise reduction and echo cancellation schemes.

Speech communication systems with VAD send only wanted signal, thus allowing more information to be transmitted through the same medium. VAD speeds up speech recognition and coding because no computations are performed on non-speech time intervals.

The goal of the VAD is to classify speech and non-speech segments as precisely as possible, but various applications have different preferences in accuracy: VAD for speech recognition must keep maximum of information by detecting speech segments with high accuracy, the misclassification of non-speech segments as speech is not critical: it doesn't reduce the speech intelligibility. On the other hand, VADs for speaker characterization (identification and classification) in security systems have to remove non-speech segments from the signal with high accuracy in order to avoid speaker characterization on background noise. The misclassification of a speech signal as a non-speech has higher tolerance.

Modern VAD and voicing detection algorithms apply different strategies to enable high classification performance: 1) extraction of parametric features, 2) classification procedure using a difference or threshold comparison measures, or using the pretrained statistical classifiers, and 3) post-processing to smooth the output decision. The typical set of parameters includes MFCCs, LSFs, frequency band energies, and zero crossing rates [5]. The classification is straightforward: it compares the extracted parameters with predetermined thresholds [5]. The postprocessing is performed to reflect the long-term stationary nature of the speech signal by applying a hangover and filtering of erroneous fluctuations of decisions. A more computationally demanding VAD that achieves higher classification performances in noise adverse conditions is presented in [7], where the system is based on the concatenation of so-called frequency filtering and Linear Discriminant Analysis-extracted features and energy dynamics features. The classification is performed by Support Vector Machines (SVM) [11]. The state-of-the-art voicing detectors and VADs use similar classification and postprocessing techniques. The main difference is in the determination of parametric features, which must discriminate the voiced and unvoiced properties of the speech signal [13]. The most typical features are based on a high-resolution spectrum and sinusoidal modeling [13], derivatives of the speech signal correlation [12], wavelet packet decomposition [10], [12], and harmonic modeling [6]. Different VAD algorithms (ETSI GSM AMR, ETSI AFE, ITU-T G.729, CIVAD) were evaluated in [2]. The evaluation criteria are 1) receiver operating curves for changing parameters and 2) word accuracy of speech recognition performed on the speech signal after VAD. In our paper we perform an evaluation without a priori assumptions of the target application. Therefore, a new application-independent evaluation criterion is defined.

As mentioned before, the VAD classifies a signal only into speech and non-speech segments. Several speech applications could profit from voice activity and voicing detection (VAD+) which separates speech segments into voiced and unvoiced parts. A typical case where the accurate VAD+ is beneficial is noise reduction: the VAD+ decision could be used to adaptively set the appropriate noise reduction strengths regarding the contents of the input speech signal. During non-speech regions, the reduction of noise could be stronger than in unvoiced regions. Moreover, the segmental SNR in unvoiced regions is lower than in voiced regions, therefore special care must be dedicated to handle unvoiced segments. The VAD+ can be applied in PDA (pitch determination) or PMA (pitch marking) algorithms, where the F0 and epochs need to be estimated only in voiced regions.

In the ECESS [1] evaluation campaign three VAD+ algorithms are compared according to a predefined criterion: the accuracy of voiced/unvoiced/non-speech detection. The
computer requires ideal reference labels, but even the most accurate manual labeling has a precision limitation: a human person is able to build approximate labels reflecting an intuitive model; the VAD+ tries to formalize this intuition. Nowadays error rates are negligible compared to the variance generated by those uncertainties, but the question of the choice of a reasonable evaluation criterion still remains.

2. VAD+ Evaluation Procedure

The desire to get accurate results led to the revision of the evaluation method.

We perform the evaluation of every sample (62.5 microseconds) in order to have enough evaluation points and fine time quantization of the reference and the VAD+ output. We count the number of correct and wrong decisions ("dead zones" are discarded in evaluation):

\[
N_{ACC} = \frac{M_{NN}}{M_{NN} + M_{VN} + M_{UN}}
\]

Similarly, the accuracies of voiced and unvoiced detection are presented below:

\[
V_{ACC} = \frac{M_{VV}}{M_{VV} + M_{UV} + M_{UW}}
\]

\[
U_{ACC} = \frac{M_{UU}}{M_{UU} + M_{UV} + M_{UW}}
\]

For further explanation we introduce error rates: \(E_{XY}\) is the error rate of misclassification of \(Y\) samples as \(X\). For example, \(E_{UY}\) is an error rate of voiced segments recognized as unvoiced:

\[
E_{UW} = \frac{M_{UW}}{M_{VV} + M_{UV} + M_{UW}}
\]

3. VAD+: Voice Activity and Voicing Detection

In the following we describe three VAD+ algorithms realized by University of Maribor/Siemens AG (UMB), Middle East Technical University (METU) and University of Vigo (UVIGO).

3.1. UMB VAD+

The UMB VAD+ is realized in the form of two cascaded artificial neural networks (ANN). For every frame \(m\), the first ANN1 detects speech and non-speech. The speech frames are classified further as voiced or unvoiced using the ANN2. The two outputs \(SN[m]\) (speech/non-speech binary classification result), and \(VU[m]\) (voiced/unvoiced binary output) are merged into one final VAD+ hard decision belonging to voiced, unvoiced or non-speech.

During the VAD+ operation, the speech signal is divided into overlapping frames of length \(L_{VAD} = 160\) ms after high-pass filtering (DC and 50Hz noise removal). Such relatively long analysis frame size is selected to ensure stable output VAD+ decisions. The selection of frame shift is not critical: in our case it is set to \(S = 10\) ms for ANN training and \(S = 1\) ms for tests. Then we divide each frame into 5 non-overlapped subframes of equal duration: \(s_0, s_1, s_2, s_3\), and \(s_4\). For all of the subframes we extract 12 MFCCs and one energy parameter. Figure 2 depicts the details about the framing of the high-pass filtered input speech signal.
the input speech signal. The 65-dimensional input vector for the ANN1 (speech/non-speech) is obtained from all 5 subframes: (12 MFCCs + 1 energy) × 5 subframes results in 65 parameters. The feature vector for ANN2 (voiced/unvoiced) is constructed from the middle subframes (s1, s2, s3) and one voicing degree parameter VD[m] resulting in a 40-dimensional feature vector: (12 MFCCs + 1 energy) × 3 subframes + 1 voicing degree. A summary about the two staged VAD+ feature extraction procedure is presented in Figure 3.

The voicing degree VD[m] provides information about the periodic content level inside the particular frame m of the analyzed speech signal: voiced speech has higher values of \( VD[m] \) than an unvoiced signal. To compute the voicing degree we apply a Hamming window to two consecutive half-frames each with a length of 48 ms or \( H_D = 256 \) samples. Due to the high sampling rate (16 kHz), the cross-correlation would require extensive computations which are accelerated to the high sampling rate (16 kHz), the cross-correlation result:

\[
X_{corr}[m,n] = \frac{X_{corr}[m,n]}{10^6} \left( 0.54 - 0.46 \cos \frac{2 \pi n}{2 H_D - 1} \right)
\]

In the next step, the 256-order FFT of \( X_{corr}[m,n] \) is computed:

\[
\text{FFT} \left[ X_{corr}[m,n] \right]
\]

where \( n = 0, \ldots, 2H_D-1 \). Finally, the voicing degree \( VD[m] \) is obtained as a spectral energy of \( X_{corr}[m,n] \) up to the frequency 520.8 Hz:

\[
VD[m] = \sum_{k=0}^{25} \left| X_{corr}[m,n] \right|
\]

### 3.2. METU VAD+

METU VAD+ is an ANN based classification algorithm. The classification of speech/non-speech and voiced/unvoiced segments is performed with two different ANNs. The algorithm consists of feature extraction and training phases.

The feature extraction phase has two main steps. In the first step, the speech signal is segmented into overlapped frames with a frame length of 16 ms and 5 ms frame shift. Then, using a minimum statistics noise estimation algorithm [8], the smoothed noise power spectral density (PSD) is estimated and subtracted from the smoothed signal PSD, in such a way the clean signal PSD is estimated. The resulting spectrum is divided into 16 equally spaced frequency bands, whose bandwidths are 250 Hz. For each band \( k \) we obtain energy \( e_k \) and generate a 16-dimensional energy vector \( E[n] \), where \( n \) is the frame number. In addition, the zero crossing rate \( ZR[n] \) of each frame is found.

In the second step, we segment the speech signal into overlapped frames with a frame length of 40 ms and with a frame shift of 5 ms. The chirp Z-transform [9] of each frame is found over 50-1500 Hz. The autocorrelation function of the magnitude of the chirp Z-transform is calculated and the maximum \( F[n] \) of the autocorrelation signal after the first minimum greater than 40 ms is found for each frame. The \( F[n] \) values are normalized so that the maximum \( F[n] \) value is equal to one for each speech file.

Two different feature sets are formed for speech/non-speech and voiced/unvoiced classifications. The first set, used by the first ANN, consists of vectors \( E[n] \), \( E[n+3] \), \( E[n+6] \), \( E[n+9] \) and \( E[n+12] \) while the second \( E[n] \), \( E[n+3] \), \( E[n+6] \), \( E[n+9] \) and \( E[n+12] \). The first ANN is used to classify speech/non-speech segments. Finally, the speech segments are discriminated as voiced/unvoiced by the second ANN.

### 3.3. UVIGO VAD+

The VAD methods based only on the energy parameters often fail in noisy environments. The UVIGO VAD+ system uses 39 acoustic characteristics: 12 Mel frequency cepstral coefficients plus the C0 coefficient, along with the first and second order derivatives. These feature vectors are normalized on a sentence-by-sentence basis using cepstral mean removal and variance normalization.

As mentioned above, the VAD+ must detect and segment a continuous audio stream in three different acoustic classes: non-speech, voiced speech and unvoiced speech. This task can be considered as an acoustic segmentation problem. Thus, the UVIGO VAD+ system uses an approach that was mainly introduced to assist Automatic Speech Recognition (ASR) systems within the context of broadcast news transcription, where one of the objectives of acoustic segmentation was to provide ASR systems with an acoustic event classification to
discard non-speech signal (silence, music, commercials, etc). Basically, in the UVIGO system the segmentation is done via the Viterbi algorithm which is used to find the best possible model sequence corresponding to non-speech, voiced and unvoiced speech classes that could have produced the input feature sequence.

The models used to represent each class are Gaussian Mixture Models (GMM) with 32 Gaussian components trained via the Expectation-Maximization (EM) algorithm in a supervised way.

4. Evaluation Setup

The manually VAD+ segmented SPEECON Spanish database [4] is used for training/development and evaluation purposes. For the training/development set we selected 2 audio files from each of 50 speakers, four SPEECON channels, this results in 400 utterances. The evaluation part of the database consists of 3680 evaluation audio files which are not part of the training/development set. Moreover, 10 speakers are completely unseen.

60 sentences (one sentence per speaker) of the close talking channel C0 and the corresponding sentences from the other three channels (C1, C2, and C3) were selected for training UMB VAD+ ANN classifiers [3]. The total amount of ANN training files is therefore 60 × 4 = 240.

All participants of the evaluation campaign have submitted the VAD+ segmentation results to the University of Maribor. The result of the evaluation is a set of accuracies (\(N_{\text{ACC}}\), \(V_{\text{ACC}}\) and \(U_{\text{ACC}}\)) and a confusion matrix (misclassification errors).

5. Results

The accuracy results of the VAD+ evaluation campaign for 3 participants are presented in Table 1. As can be seen, University of Maribor (UMB) has the best voiced detection accuracy, and University of Vigo (UVIGO) has the best unvoiced detection accuracy. As expected, the accuracy of VAD+ algorithms decreases with increasing noise levels: channel 0 has the highest accuracy, and the VAD+ accuracy on utterances from channel 3 is the lowest. Channel 0 is recorded with close talk microphone; for channel 1 a lavaliere microphone is used; channels 2 and 3 are recorded with distantly placed microphones [4].

The detection of unvoiced segments becomes harder with increasing noise levels. This has two explanations. First, it is hard to distinguish between speech and non-speech segments in a noisy environment. Second, the motor noise can be confused with “voice” in a signal due to its tonal contents.

Analyzing confusion matrices (not provided here) we observe that for the low noise channels 0 and 1, the misclassification of voiced segments as non-speech and non-speech as voiced are lower (\(E_{\text{VSN}}\) and \(E_{\text{NSV}}\)) than other confusions. This may be explained that non-speech and voiced segments have dissimilar acoustic models. Furthermore, the performance analysis for 10 unseen speakers does not show a statistically important degradation.

6. Conclusion

In this paper we have presented three voice activity and voicing detection algorithms. They were compared using our evaluation procedure. Training and evaluation was performed using a manually segmented subset of the Spanish SPEECON speech database. In order to exclude errors caused by manual labeling we introduced “dead zones” and discarded from considerations intervals around manual label time stamps.

7. References