Voice Activity Detection Based on the Adaptive Multi-Rate Speech Codec Parameters

Giacobello, Daniele; Semmoloni, Matteo; Neri, Danilo; Prati, Luca; Brofferio, Sergio

Published in:
Proceesings of the 11th International Workshop on Acoustic Echo and Noise Control

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

VOICE ACTIVITY DETECTION BASED ON THE ADAPTIVE MULTI-RATE SPEECH CODEC PARAMETERS

Daniele Giacobello[1], *Matteo Semmoloni*[2], *Danilo Neri*[2], *Luca Prati*[2], *Sergio Brofferio*[3]

[1]Department of Electronic Systems, Aalborg University, Aalborg, Denmark
[2]Nokia Siemens Networks, Cinisello Balsamo, Milano, Italy
[3]Dipartimento di Elettronica e Informazione, Politecnico Di Milano, Milano, Italy
dg@es.aau.dk, {matteo.semmoloni,danilo.neri,luca.prati}@nsn.com, sergio.brofferio@polimi.it

ABSTRACT

In this paper we present a new algorithm for Voice Activity Detection that operates on the Adaptive Multi-Rate codec parameters. Traditionally, discriminating between speech and noise is done using time or frequency domain techniques. In speech communication systems that operate with coded speech, the discrimination cannot be done using traditional techniques unless the signal is decoded and processed, using an obviously inherently suboptimal scheme. The proposed algorithm performs the discrimination exploiting the statistical behavior of the set of parameters that characterize a segment of coded signal in case of presence or absence of voice. The algorithm presented provides significantly low misclassification probabilities making it competitive in speech communication systems that require low computational costs, such as mobile terminals and networks.

Index Terms— Voice Activity Detection, Adaptive Multi-Rate Codec

1. INTRODUCTION

Voice Activity Detection (VAD) is an integral part of all modern speech communication devices. In the context of mobile communication, the accurate functioning of the discrimination between voice and noise can improve the total efficiency of the system, allowing to send only the packets corresponding to speech signal and few bits of information about the background noise if the speech signal is not present. A robust VAD can also be used in the Voice Quality Enhancement (VQE) techniques such as Noise Reduction (NR) allowing the algorithm to use the noise information to improve the speech signal quality, for example with spectral subtraction. In this paper we will present a VAD that works directly on the AMR domain, being this the standard speech codec adopted in GSM and UMTS networks. After giving a brief overview on the AMR codec we will present how each parameter is used for the discrimination and how to combine the information in order to have a final binary decision for each coded speech segment. We will conclude our work showing and discussing the performances of the algorithm.

2. OVERVIEW OF THE ADAPTIVE MULTI-RATE CODEC

The AMR [1] was chosen by the 3GPP consortium as the mandatory codec for the UMTS mobile networks working with speech sampled at 8 kHz. Its main advantage is to be a multimodal coder, working on different rates from 12.2 kbit/s to 4.75 kbit/s, with the possibility of changing rate during the voice transmission by interacting with the channel coder. In our studies, mainly centered on the analysis of parameters, we worked on the 12.2 kbit/s mode (AMR 122) considering straightforward the extension to lower bit rates. Below, we will give a brief overview on the main aspects of the encoder.

The AMR codec is based on the Algebraic Code Excited Linear Prediction (ACELP) paradigm that refers to a particular approach for finding the most appropriate residual excitation after the linear prediction (LP) analysis. The speech waveform, after being sampled at 8 kHz and quantize with 16 bits, is divided into frames of 20 ms (160 samples) where each frame contains 4 subframes of equal length. The codec then uses a 10^{th} order linear predictive analysis on a subframe basis and then transform the coefficients obtained into Line Spectral Frequencies (LSF) [2] for more robust quantization.

After passing the signal through the LP filters, a residual signal is obtained. The codec then looks for a codeword that best fits the residual. There are two codebooks in the ACELP paradigm: an adaptive codebook and an algebraic codebook (also called fixed codebook). The parameters of the adaptive codebook are the pitch gain and pitch period; these are found through a closed-loop long-term analysis. The parameters of the fixed codebook are found analyzing the residual signal subtracted of its pitch excitation. The calculations make possible to find a codeword with only 10 non-zero coefficients. It has been shown [3] that a good approximation for the transfer

function of the n^{th} subframe is given by:

$$H_n(z) = \frac{g_{fc}(n)}{\left(1 - g_p(n)z^{-T_p(n)}\right)\left(1 - \sum_{i=1}^{10} a_i(n)z^{-i}\right)}, \quad (1)$$

where $g_{fc}(n)$ is the fixed codebook gain, $g_p(n)$ and $T_p(n)$ are the parameters of the pitch excitation and $\{a_i(n)\}$ are the linear prediction coefficients or equivalently the line spectral frequencies $\{L_i(n)\}$.

The decoder performs the synthesis of the speech using the transmitted parameters. The excitation that is passed through the LP filter is created by combining the fixed codeword, multiplied by its gain, and the adaptive codeword.

3. DISCRIMINATIVE MEASURES PERFORMED ON THE AMR PARAMETERS

3.1. Line Spectral Frequencies

The LSF from the way they are constructed, are directly related to the frequency response of the LPC filter [2]. For this reasons they have been studied also regarding their speech recognition performances [4]. It is then clear that they can also be used for VAD purposes. In particular, it is easy to notice that for highly organized spectra (voiced speech) the LSF tend to position themselves close to where the formants are located; as opposed to the case of white noise where, having this a flat spectrum, the LSF will tend to spread equally along the unit circle. In order to exploit this behavior, a measure similar to the spectral entropy has been chosen by calculating the entropy of the LSF differential vector $\mathbf{L}' = (l_2 - l_1, \ldots, l_{10} - l_9)$:

$$ENT = -\sum_{n=1}^{9}\left[\frac{L'(n)}{\sum_{n=1}^{9} L'(n)} \log_2\left(\frac{L'(n)}{\sum_{n=1}^{9} L'(n)}\right)\right]. \quad (2)$$

The calculation of (2) is similar to the spectral entropy in the sense that, given the LSF vector $L = (l_1, \ldots, l_{10})$, the frequency response of the LPC filter $H(\omega)$ can be approximated with rectangular impulses [5]:

$$\hat{H}_i(\omega) = \frac{A}{l_i - l_{i-1}}, \quad l_i < \omega < l_{i-1}, \quad (3)$$

where A is a scaling factor and the domain of ω is the one of the normalized frequencies $[0, \pi]$. Summing all the rectangular impulses we obtain an approximation of the spectrum:

$$\hat{H}(\omega) = \sum_{i=2}^{10} \hat{H}_i(\omega), \quad (4)$$

The entropy of the LSF differential vector (2) is then an approximation of the spectral entropy of $\hat{H}(\omega)$.

This highly reliable feature will be used as a main discriminative factor in our algorithm, being weakly influenced by the SNR and the energy level in a conversation.

3.2. Pitch Period

The pitch period can be particularly useful to perform VAD due to its properties. In particular, for voiced speech the pitch period will tend to maintain itself around a certain value that can differ depending on the speaker, usually between 18 and 143 samples at 8 kHz (56 Hz and 450 Hz in the frequency domain). In particular, we will analyze its variance in a AMR frame making it also speaker-independent (by removing its mean value):

$$TV = \sum_{n=1}^{4}\left[T_p(n) - \frac{1}{4}\sum_{n=1}^{4} T_p(n)\right]^2. \quad (5)$$

The statistical behavior of the pitch period during unvoiced speech and voiced speech does not show any difference: in both cases it will have a quasi-uniform density probability over the possible values. Nevertheless, its variance feature TV has shown to be very robust in detecting voiced speech: high during unvoiced speech and noise, low during voiced speech.

3.3. Fixed Codebook Gain

The Fixed Codebook Gain $g_{fc}(n)$, as can be seen from (1), is the parameter that is most directly related to the energy of an n^{th} AMR subframe; it is therefore used as an indicator of the energy level in a subframe and a feature in the VAD process without any processing:

$$GFC = g_{fc}. \quad (6)$$

The feature GFC is not very robust in terms of SNR, nevertheless using adaptive thresholds we will see that can guarantee a good discriminative behavior.

4. STRUCTURE OF THE VOICE ACTIVITY DETECTOR

In this section we show how the features have been combined and how the voice activity detection takes place and brings to the final decision.

4.1. VAD Hangover

One of the main problems in the creation of any voice activity detector is the similarity of the statistical behavior of the discriminative features in presence of noise and unvoiced speech. In order to mitigate this effect, we use a recursive filter on the values with the purpose to conserve the effect of the voiced speech for the duration of the unvoiced speech. Considering $x(n)$ the feature value for the n^{th} subframe, the output $y(n)$ will be, if $y(n-1) > x(n)$:

$$y(n) = a_R x(n) + (1 - a_R)y(n-1), \quad (7)$$

where $a_R = 1 - e^{-5/N_R}$ and N_R is the length of the step response of the filter, in our experimental analysis we used $N_R = 100$, equivalent to $0.5s$. The choice of this value is related to the characteristics of the speech signal and therefore is the same for each feature. In the case $y(n-1) \leq x(n)$ the filtering will not take place. Thus, if the value is decreasing after being high, most likely due to the presence of voiced speech, the signal $y(n)$ will decrease less rapidly preventing the signal to go below the voice-noise threshold in presence of unvoiced speech. It should be noted that operating this filtering, we highly reduce the temporal clipping that can be introduced in the middle and at the end of the speech signal that can highly lower the quality of the signal [7]. On the other hand, the probability of false alarm (misdetecting noise for speech) will necessarily be higher; nevertheless, it is clear that perceptually speaking, it is preferable to misdetect noise for speech than the other way around.

4.2. Initial Training

Our algorithm supposes an initial period of 100 ms for training (20 subframes). In this period of time, supposedly of only background noise, the features (ENT, TV, GFC) are calculated and processed to determine the initial discriminative thresholds. Under the hypothesis of gaussianity that holds well in this case, we first find the mean value μ_{bn}^f and the standard deviation σ_{bn}^f for each parameter f and these values will characterize the probability density function of features during noise conditions.

In our algorithm we will use five thresholds; This is done to create a fuzzy VAD and postpone the final binary decision to a latter stage in order to take into account other factors. The determination of the thresholds is done dividing the noise probability density functions obtained in confidence zones; for ENT and GFC the thresholds are $TH_1 = \mu_{bn}^f$, $TH_2 = \mu_{bn}^f + \sigma_{bn}^f$, $TH_3 = \mu_{bn}^f + 2\sigma_{bn}^f$, $TH_4 = \mu_{bn}^f + 3\sigma_{bn}^f$, $TH_5 = \mu_{bn}^f + 5\sigma_{bn}^f$ and for TV the thresholds are (considering that $\mu_{bn}^{TV} = 0$) $TH_1 = 1/72\sigma_{bn}^{TV}$, $TH_2 = 1/36\sigma_{bn}^{TV}$, $TH_3 = 1/27\sigma_{bn}^{TV}$, $TH_4 = 1/18\sigma_{bn}^{TV}$, $TH_5 = 1/9\sigma_{bn}^{TV}$. After this initial stage, each feature value, after being filtered by (7) will be compared to its respective thresholds in order to define a likelihood value; for example for the entropy feature ENT the cycle at the n^{th} subframe will be:

if $ENT(n) < TH_1$ **then**
  $VAD_{ENT}(n) = 0$
else if $ENT(n) \geq TH_1$ **and** $ENT(n) < TH_2$ **then**
  $VAD_{ENT}(n) = 0.2$
  ...
else if $ENT(n) \geq TH_4$ **and** $ENT(n) < TH_5$ **then**
  $VAD_{ENT}(n) = 0.8$
else
  $VAD_{ENT}(n) = 1$
end if

The fuzzy VAD values for each feature $VAD_{ENT}(n)$, $VAD_{GFC}(n)$ and $VAD_{TV}(n)$ are then combined into one value using a different weights ρ for each feature, determined empirically by analyzing their discriminative performances. In particular each VAD has been tested alone under different conditions of noise (car, wgn, babble, rain, street) and SNR (-15$dB \div$ 25dB). The results where following the initial statistical analysis: $\rho_{ENT} = 0.41$, $\rho_{GFC} = 0.33$ and $\rho_{TV} = 0.26$.

4.3. Smoothing Rule

Once we have found a fuzzy VAD as a linear combination of the three values used in the discriminative process, we have to make a final binary decision. To strengthen the effort made by the filter in (7) to prevent the algorithm from clipping unvoiced sound, we introduce a smoothing rule based on the principle that an unvoiced sound is never an isolated phenomenon but comes always before of after a voiced sound that is much easier to detect. In order to do so, the algorithm makes a decision based not only on the current subframe but uses also the fuzzy values from the previous 15 subframe. In other word:

$$VAD_{bin}(n) = \begin{cases} 1 & \text{if} \quad \sum_{k=n-15}^{n} VAD_{fuzzy}(n-k) > H, \\ 0 & \text{otherwise}, \end{cases} \tag{8}$$

where $H = 0.55$ is a constant value found empirically that gave us the best performances in the trade-off between keeping the rate of correct classification of speech high and the false alarm rate low. An example of the functioning of the algorithm is shown in figure 1.

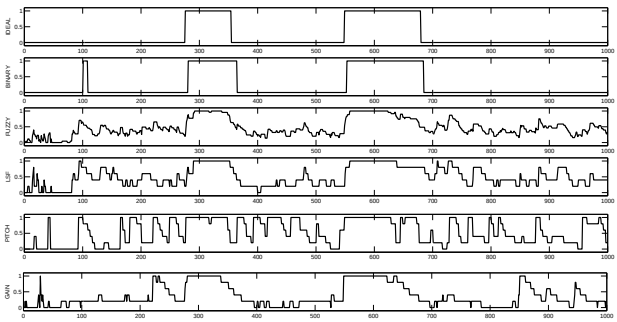

Fig. 1. *Example of the VAD functioning ($SNR = 12dB$, street noise). From below we have VAD_{GFC}, VAD_{TV}, VAD_{ENT}, VAD_{fuzzy}, VAD_{bin} and the ideal reference VAD.*

4.4. Thresholds Updating

The background noise in mobile networks, other than being highly non-stationary, can also change drastically during the course of a normal conversation. In order to compensate

VAD Performances					
SNR	NOISE	P_D%		P_{FA}%	
		COD	LIN	COD	LIN
$5dB$	WGN	88.8	91.7	10.5	7.2
	BABBLE	79.1	82.5	29.2	25.3
	AVERAGE	80.7	81.7	26.2	23.1
$12dB$	WGN	94.1	96.2	9.3	5.4
	BABBLE	91.4	93.2	26.1	18.3
	AVERAGE	91.5	92.9	21.1	17.1
$20dB$	WGN	96.2	98.6	6.2	3.4
	BABBLE	95.6	97.5	17.5	11.3
	AVERAGE	96.0	97.1	15.8	10.7

Table 1. Performances comparison between the proposed algorithm (COD) and the ETSI AMR-2 (LIN)

this phenomenon, an update of the thresholds found in the initial training stage is necessary. In order to do so, when $VAD_{bin} = 0$, the algorithm will update the thresholds by updating the mean value μ_{bn}^f and the standard deviation σ_{bn}^f of the background noise for each feature f. In order to do so, we used a linear estimation of the first and second order moments:

$$\begin{aligned}\mu_{bn}^f(k) &= a_\mu \mu_{bn}^f(k-1) + \frac{1-a_\mu}{N}\sum_{n=k-N}^{k} x(n),\\ \sigma_{bn}^f(k) &= a_\sigma \sigma_{bn}^f(k-1) + \\ &\quad \frac{1-a_\sigma}{N}\sum_{n=k-N}^{k}|x(n) - \frac{1}{N}\sum_{l=k-N}^{k} x(l)|.\end{aligned} \tag{9}$$

In both cases $a_\sigma = a_\mu = 1 - e^{-5/N}$, where $N = 100$ (0.5 s) is the length of the window considered during the calculations and approximately the length of the step response of the filter. The value of N has been found empirically considering the trade-off between the possibility to adapt rapidly and the robustness to noise bursts.

5. EXPERIMENTAL RESULTS

In order to evaluate the algorithm, several hours of conversation from both male and female speakers have been analyzed. The VAD was tested under different SNR conditions and noise types (wgn, rain, car, street and babble). The results, for different kinds of SNR and noise are shown in Table 1, for brevity we show only the best and worst conditions for our VAD (wgn and babble) and the average over the whole five noise types. The proposed algorithm is compared with the ETSI AMR-2 voice activity detector [6]. It is clear from the experimental results that the VAD implemented can compete in complexity and performances with modern commercial VAD. The algorithm has been designed to privilege the probability to detect speech when present P_D over the false-alarm probability P_{FA}. In this way, it smoothens the rapid decay of perceived quality when clipping of speech is present [7]. In fact, the mid-speech and end-speech clipping are almost not present thanks to the solutions implemented in the VAD. On the other hand, the front-end clipping is still present because, in order to keep the delay (one of the major constraints in mobile networks) as low as possible, no look-ahead has been being used.

6. CONCLUSIONS

In this paper we have presented an innovative VAD structure that operates directly on the AMR compressed domain. In particular, we have shown that reducing the complexity of the VAD process by transposing the operations on the AMR codec parameters is not only possible but preferable as the experimental results have shown to be comparable with the VADs commercially available. These techniques are suitable for implementation in mobile networks and other kind of networks working with AMR-coded speech. Given the interesting results of all the algorithms tested on the UMTS network, we can see these as a good alternative to the existing VAD procedures.

7. REFERENCES

[1] 3GPP, *TS 26.071; AMR speech codec: General Description*, Version 7.0.0, 2007.

[2] T. Bäckström, C. Magi, Properties of line spectrum pair polynomials - A review, *Signal Processing*, vol. 86, no. 11, november 2006, pp. 3286-3298.

[3] H. Taddei,C. Beaugeant,M. de Meuleneire, Noise Reduction on Speech Codec Parameters, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2004.

[4] K. K. Paliwal, A Study of Line Spectrum Pair Frequencies for Vowel Recognition, *Speech Communication*, vol. 8, 1989, pp. 2733.

[5] F. Zheng, Z. Song, W. Yu, F. Zheng, W. Wu, The Distance Measure for Line Spectrum Pairs Applied to Speech Recognition, *Journal of Computer Processing of Oriental Languages*, vol. 11, march 2000, pp. 221-225.

[6] 3GPP, *TS 26.094; AMR speech codec: Voice Activity Detector (VAD)*, Version 7.0.0, 2007.

[7] L. Ding, A. Radwan, M. S. El-Hennawey, R. A.. Goubran, Measurement of the Effects of Temporal Clipping on Speech Quality, *IEEE Transaction On Instrumentation and Measurement*, vol. 55, no. 4, august 2006, pp. 1179-1203.