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]Aalborg University, Aalborg, Denmark [2]Nokia Siemens Networks, Milano, Italy [3]Politecnico di Milano, Milano, Italy

1 Motivation

- 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 works in the AMR compressed domain performing the discrimination by exploiting the statistical behavior of the set of parameters that characterize a segment of coded signal in presence or absence of speech.

2 AMR Codec

Analysis made on 20ms frame (160 samples @8kHz) divided further on in 5ms subframes, for each subframe are extracted:

- 10 short-term coefficients $\{a_i\}$ then converted in 10 Line Spectral Frequencies $\{L_i\}$;
- pitch delay T_p and gain g_p (long-term predictor);
- algebraic codeword and gain g_{fc}.

For the n^{th} subframe, the AR model in which the excitation is passed through:

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

3 Discriminative Measures performed on the AMR parameters

- **Entropy of the Line Spectral Frequencies**

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

where $L'(n) = l_{n+1} - l_n$. For highly structured spectra (voiced speech) the LSF tend to position themselves close to where the formants are located (low entropy). For flat spectra, the LSF will tend to spread equally along the unit circle (high entropy).

- **Variance of the Pitch Period**

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

For voiced speech segments the pitch period will have low variance. For noise segments the pitch period estimation does not present a clear pattern (high variance).

- **Fixed Codebook Gain**

$$GFC = g_{fc}$$

This is the parameter that is most directly related to the energy level in a subframe.

4 Structure of the VAD

- **VAD Hangover**. The purpose is 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 $y(n-1) > x(n)$:

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

where $a_R = 1 - e^{-5/N}$, where $N = 100$ (0.5s).

- **Initial Training**. In the first 100 ms (20 subframes) a Gaussian pdf of the background noise with μ_{bn}^f and the standard deviation σ_{bn}^f for each feature f is used to define five thresholds.
- **Fuzzy Decision**. Comparison of the value of the feature f with the thresholds found, A fuzzy VAD on 6 levels is found.

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 by summing using different weights ρ for each feature, determined empirically by analyzing their discriminative performances under different conditions of noise and SNR: $\rho_{ENT} = 0.41$, $\rho_{GFC} = 0.33$ and $\rho_{TV} = 0.26$.

- **Smoothing Rule**. Used to prevent the algorithm from clipping unvoiced sound by making a decision based using the fuzzy values from the previous 15 subframe:

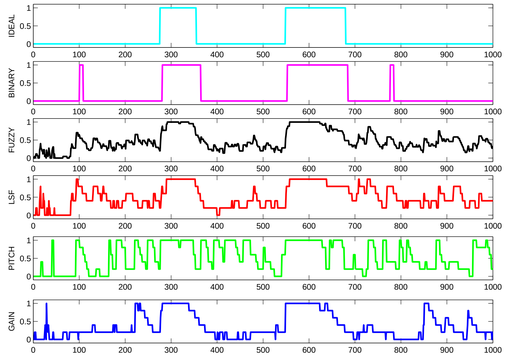
$$VAD_{bin}(n) = \begin{cases} 1 & \text{if } \sum_{k=n-15}^{n+1} VAD_{fuzzy}(k) > 0.003 \\ 0 & \text{otherwise} \end{cases}$$

- **Thresholds Updating**. The background noise in mobile networks can also change drastically during the course of a normal conversation. To compensate, the thresholds found in the initial training stage are changed when $VAD_{bin} = 0$ by updating the mean value μ_{bn}^f and the standard deviation σ_{bn}^f of the background noise for each feature f:

$$\mu_{bn}^f(k) = a_\mu \mu_{bn}^f(k-1) + \frac{1 - a_\mu}{N} \sum_{n=k-N+1}^{k} x(n)$$

$$\sigma_{bn}^f(k) = a_\sigma \sigma_{bn}^f(k-1) + \frac{1 - a_\sigma}{\sqrt{N}} \sqrt{\sum_{n=k-N}^{k} |x(n) - \frac{1}{N}\sum_{l=k-N+1}^{k} x(l)|^2}$$

In both cases $a_\sigma = a_\mu = 1 - e^{-5/N}$, where $N = 100$ (0.5s).



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.

5 Experimental Results

VAD tested under different SNR conditions and noise types (wgn, rain, car, street and babble). The results of the best and worst conditions for our VAD (wgn and babble) and the average over the whole five noise types are shown in the table below. The proposed algorithm is compared with the ETSI AMR-2 voice activity detector.

VAD Performances					
SNR	NOISE	P_D%		P_{FA}%	
		COD	LIN	COD	LIN
5 dB	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
12 dB	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
20 dB	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

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

6 Conclusion

- Reducing the complexity of the VAD process by transposing the operations on the AMR codec parameters is not only possible but preferable as the performances are comparable to the VADs commercially available.
- These techniques are suitable for implementation in mobile networks and other kind of networks working with AMR-coded speech.
- A good alternative to the existing VAD procedures.

References

[1] H. Taddei *et al.*, "Noise Reduction on Speech Codec Parameters", *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2004.*

[2] D. Giacobello, "Study and Evaluation of Innovative Algorithms for Voice Quality Enhancement in Speech Signals Encoded Using ACELP (Algebraic Code Excited Linear Prediction)", M.Sc. Thesis, Politecnico Di Milano, Milano, Italy, July 2006.