A COMPARATIVE STUDY OF SPEECH DETECTION METHODS *

Stefaan Van Gerven † *and Fei Xie* ‡

K. U. Leuven, Department of Electrical Engineering - ESAT
Kardinaal Mercierlaan 94, B-3001 Heverlee, Belgium
E-mail: Stefaan.VanGerven@esat.kuleuven.ac.be

ABSTRACT

This paper adresses the important problem of speech detection. It describes the implementation of 3 speech detection methods and compares their performance under different signal-to-noise ratio (SNR) and stationarity conditions. The method that dynamically adjusts its thresholds is found to be the most reliable, even under very adverse recording conditions. Yet it is of low complexity and has a very moderate processing delay.

1. INTRODUCTION

Speech detection is an ubiquitous problem in speech processing. It consists of the classification of the two clearly distinct signal conditions during a speech recording: periods where the speech signal is present and pauses with background noise only.

- In *speech recognition*, word boundaries must be approximately known to trigger the recognizer correctly. The problem is usually termed *begin-endpoint detection* and a.o. described in [1, pp. 433-435], [2, pp. 246-251]. Word boundaries can often be detected off-line i.e. after the complete utterance has been recorded. An incomplete (too late) detection may cause recognition errors and intra-word gaps should not be classified as noise so that words are detected contiguously. On the other hand, a detection of a sound that is not a recognizer word is not necessarily harmful if a good rejection strategy is used.
- *Adaptive speech enhancement* algorithms typically behave completely different during speech periods than during noise periods. During speech periods the algorithms should learn as much as possible about the speech source and during noise periods as much as possible about the noise source(s). Correct *voice activity detection* (VAD) is therefore crucial to their success. This is true for both single channel spectral subtraction [3] as well as for multi-channel adaptive beamforming-type algorithms [4].
- In *speech coding*, the bit-rate can be lowered drastically during silence periods without effect on the perceived speech quality. Within the recent speech coding standards of the ITU as well as within the half-rate and enhanced full-rate GSM standards full silence compression schemes have been described.

The remainder of the paper is organized as follows. Section 2 describes the different algorithms in detail. An illustrative example is developed throughout that section. Section 3 then presents comparative tests under different signal-to-noise ratio (SNR) and stationarity conditions. Section 4 concludes the paper.

2. ALGORITHMS

All 3 algorithms presented are based on the short-time log-energy of the signal. This parameter is calculated on a frame-by-frame basis with a typical framelength of 10 ms. We will shortly use *log-energy* to denote this quantity.

Figure 1 shows the waveform of the utterance *"speech detection"*, used for the illustration of the different methods. The SNR of this utterance is 29 dB. This utterance is particular since it contains a soft onset *"s"* as well as an intra-word pause around 1 s. because of the plosive *"p"*.

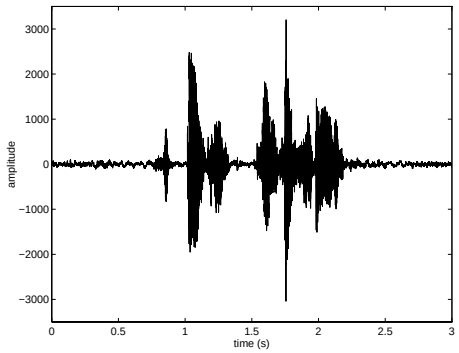

Figure 1: Waveform for *"speech detection"*.

* Work supported by IWT Research Contract 93021 (AKROS).
† Currently with Lernout & Hauspie Speech Products.
‡ Currently with INRS Telecommunications, Quebec, Canada.

2.1. Algorithm 1

The baseline algorithm uses an off-line strategy. It first calculates the global statistics (histogram) of the log-energy for a complete recording and then determines a fixed decision threshold. This method was first presented in [5] as an extension of a principle developed in [6]. It is based on the observation that the histogram of the log-energy of speech-in-noise typically has a bimodal distribution. This distribution can be approximated with two gaussian densities which allows to derive a statistically optimal decision threshold T. For the *"speech detection"* example, the speech-noise histogram together with its bimodal gaussian fit are shown in figure 2.

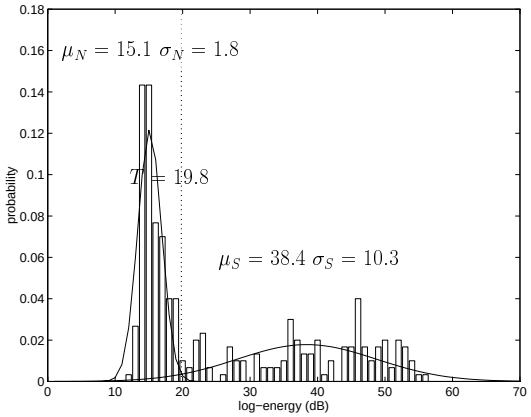

Figure 2: Log-energy histogram for *"speech detection"*.

Accurate histograms require several seconds of speech in noise, but the method has been applied satisfactorily for detection of single short words ($<$ 1 s) as well. The algorithm performs well for high to moderate SNR's and under stationary background noise.

In figure 3 the log-energy over time is shown together with the decision threshold T and the mean noise level μ_N.

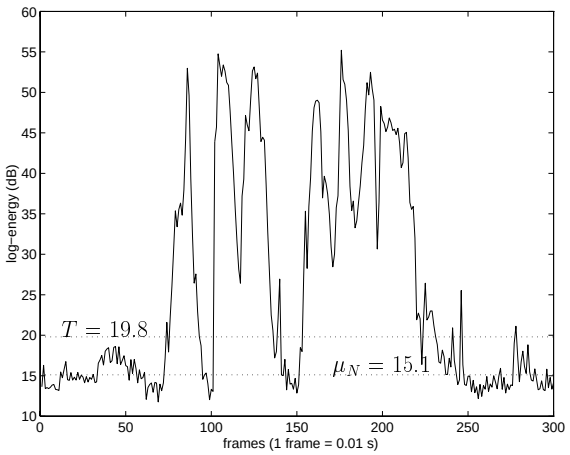

Figure 3: Short time log-energy and decision threshold T for *"speech detection"*, method 1.

The fixed decision threshold leads to the detection of some very short and undesired 'speech'-periods. For adaptive processing they would not harm as they only represent short periods of non-adaptation. An intermediate decision level between the noise mean μ_N and the statistical threshold T (with less adaptation) can even give better results. For a speech-recognizer such addtional triggers would be more difficult to handle. Furthermore the intra-word pause is detected as non-speech which would also be an incorrect input to the recognizer.

Therefore a second stage is usually applied that consists of adding constraints to the speech detection : e.g. a minimum word length of 150 ms. and a minimum gap lenght of 100 ms.. This allows to isolate words correctly before presenting them to the recognizer.

2.2. Algorithm 2

The second algorithm tries to mimic the behavior of the baseline algorithm based on local information only. It is thus an on-line algorithhm and may be implemented in real-time.

The recent mean of the log-energy E_1 over the last t_1 ms. is calculated and compared to a speech threshold T_s. The latter is found by adding a fixed dB value (e.g. E_s = 5 dB) to a past mean log-energy value E_2 which is calculated over the previous t_2 ms.. When E_1 surpasses T_s speech onset is detected and a speech offset (noise) threshold T_n is calculated by adding another fixed dB value (e.g. E_n = 2 dB) to e_2. When E_1 then drops under T_n speech offset is detected. E_2 is not updated during speech.

In figure 4 all quantities involved in this method are shown for the *"speech detection"* example.

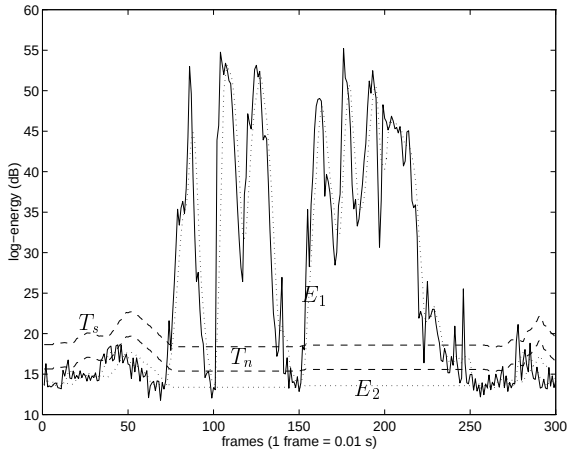

Figure 4: Short time log-energy, mean values E_1 and E_2 (dotted lines) and thresholds T_s and T_n (dashed lines) for *"speech detection"*, method 2.

The choice of E_s is crucial. E_s can be interpreted as the minimal SNR needed for speech detection. E_n on the other hand can be interpreted as the noise variation. The algorithm performs well for high to moderate SNR's and

under moderately varying background noise levels. In particular the noise level should not increase too much during a word as otherwise the speech offset will not work.

The minimal processing delay is a single frame i.e. 10 ms. but the algorithm can be made more accurate if a delay of e.g. 30 ms. is acceptable. Using such a longer delay, the speech onset detection has been used successfully for speech recognition demonstrations with continuous recording (open microphone) and even for a real-time speech recognition system. With proper parameter settings short clicks, coughs and background noises were not detected as words.

2.3. Algorithm 3

The third algorithm calculates the speech onset and offset thresholds as a function of the local noise statistics only. The idea stems from the fact that the log-energy distribution of (stationary) noise can be modeled easier than the log-energy distribution of speech and therefore no assumptions should be made about the speech signal. Rather than speech detection it can thus be interpreted as detection of log-energy values that do not belong to the estimated noise distribution.

Noise mean μ_N and noise variance σ_N estimates are continuously updated during non-speech periods. The thresholds derived from these parameters are thus adaptive and can cope with (slowly) varying noise levels. The time constants involved in the recursive estimation of the noise parameters can be varied as a function of the expected degree of non-stationarity.

The algorithm works as follows. From the estimated noise parameters μ_N and σ_N calculate the speech threshold as :

$$T_s = \mu_N + \alpha\sigma_N$$

When the log-energy exceeds T_s, speech onset is detected and a noise threshold T_n is fixed as :

$$T_s = \mu_N + \beta\sigma_N$$

When the log-energy drops under T_n speech offset is detected and the estimation of μ_N and σ_N restarts. Typical values are : $\alpha = 4$ and $\beta = 1.2$.

The illustration for the *"speech detection"* example is given in figure 5. Note that as in method 2 and in the unconstrained version of method 1 the intra-word pause in *"speech"* is detected as a noise period.

The delay of this algorithm is higher than the previous algorithm, not conceptually, but because better results were reported when the framelengths were increased to 20 ms. and above (smoothed log-energy). The main advantage of this algorithm is that it performs well even under low SNR and higly non-stationary conditions.

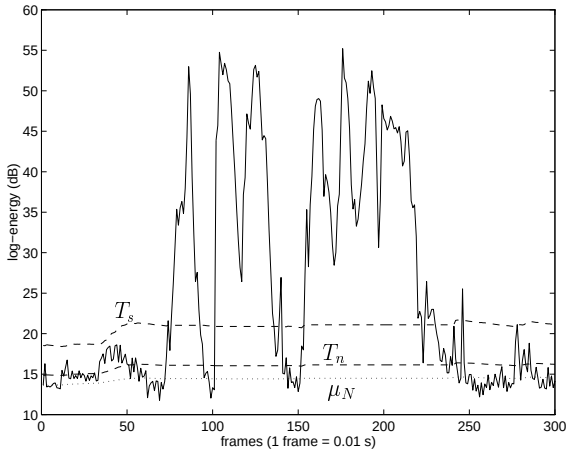

Figure 5: Short time log-energy, mean noise estimate μ_N (dotted line) and thresholds T_s and T_n (dashed lines) for *"speech detection"*, method 3.

3. COMPARATIVE TESTS

To compare the different methods, a criterion for 'optimal' speech detection must be put forward. This is difficult as the 'optimality' is very much application dependent and the parameters of the speech detection algorithms have to be tuned towards the application. We have therefore not focused on a particular application but were interested in the speech detection as such. We have tested the algorithms in different conditions and provide the test results graphically, which allows a visual evaluation.

3.1. Artificially added noise

For the tests with artificially added noise we have used a nearly 9 s. utterance containing 5 triplets of connected English digits. The utterance was selected from the NOISEX database [7]. Noise of 6 different types, also selected from the NOISEX database, was added at 4 different SNR's (18, 12, 6 and 0 dB). Figures 6 and 7 show the result for a speech-like noise at 12 dB and at 0 dB. Method 1 was applied without the additional constraints and the framelength used for all methods was 16.66 ms..

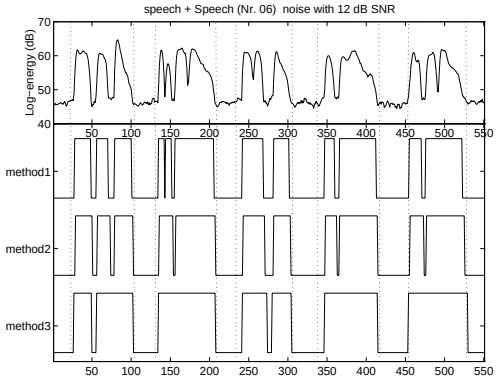

Figure 6: Log-energy and speech detection for all 3 methods at 12 dB SNR.

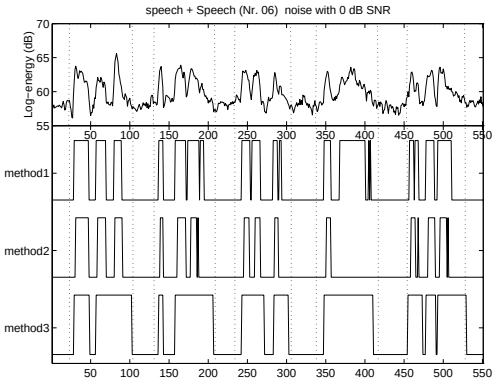

Figure 7: Log-energy and speech detection for all 3 methods at 0 dB SNR.

3.2. Car recordings

We have also tested the speech detectors on real car recordings in different driving conditions. These recordings were multi-microphone recordings made to test adaptive multi-microphone speech enhancement algorithms as preprocessing for speech recognition in cars. Here we have selected a 5 word utterance from a recording made on the highway with open window. This is one of the most severe noise conditions we have in our database. The result of the speech detection is shown in figure 8.

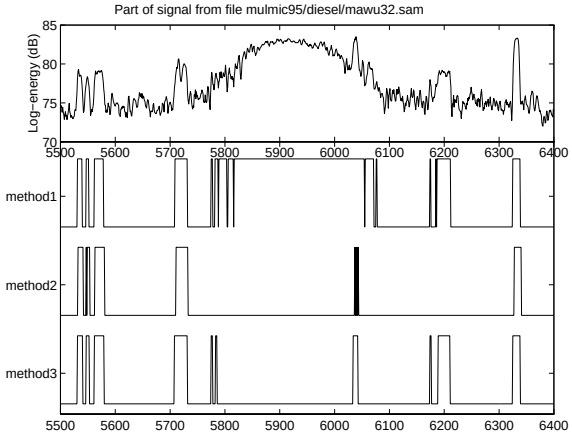

Figure 8: Speech detection for 5 words of a recording made in a car driving on the highway (90-110 km/h) with driver window open. The words are located around frames 5550, 5700, 6050, 6150, 6350. The huge noise burst is caused by another car passing by. Due to the fixed threshold, method 1 classifies the noise burst as being speech. Method 2 does not but only marginally detects word 3 and fails to detect word 4. Method 3 detects word 3 although not completely.

3.3. Summary of test results

From these and other tests it can be concluded that algorithms 1 and 2 behave similar for low to moderate SNR's under stationary conditions. Algorithm 2 outperforms for slowly varying background noises but fails more rapidly as the SNR decreases. Algorithm 3 is the best choice for lower SNR's and remains reliable for strongly non-stationary signals. Algorithm 3 has finally been selected to 'trigger' the adaptive filters in the speech enhancement experiments on the multi-microphone database. Note that a priori knowledge of the expected SNR value and the degree of non-stationarity helps in setting the algorithm parameters for optimal performance.

4. CONCLUSIONS

This paper has shown that it is possible to develop reliable speech detection algorithms with low complexity and moderate processing delay.

Starting from an off-line algorithm, two on-line algorithms have been derived. The algorithms discriminate speech and noise based on the short-time log-energy of the signal only.

The algorithm that dynamically adjusts its speech threshold as a function of the estimated noise statistics was found to be the most reliable even in very adverse recording conditions.

REFERENCES

1. D. O'Shaughnessy. *Speech Communication: Human and Machine*. Addison-Wesley Publishing Company, 1987.

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

3. R. Le Bouquin and G. Faucon. Study of a voice activity detector and its influence on a noise reduction system. *Speech Communication*, 16:245–254, 1995.

4. D. Van Compernolle. Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 833–836, Albuquerque, April 1990.

5. D. Van Compernolle. Noise adaptation in a hidden Markov model speech recognition system. *Computer Speech and Language*, 3(2):151–168, 1989.

6. R. J. McAulay and M. L. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(2):137–145, 1980.

7. A. Varga, H. J. M. Steenneken, M. Tomlinson, and D. Jones. The NOISEX-92 study on the effect of additive noise on automatic speech recognition, 1992. (Documentation included in the NOISEX-92 CD-ROM's).