Speech Pause Detection for Noise Spectrum Estimation by Tracking Power Envelope Dynamics

Mark Marzinzik and Birger Kollmeier

***Abstract*—A speech pause detection algorithm is an important and sensitive part of most single-microphone noise reduction schemes for enhancement of speech signals corrupted by additive noise as an estimate of the background noise is usually determined when speech is absent. An algorithm is proposed which detects speech pauses by adaptively tracking minima in a noisy signal's power envelope both for the broadband signal and for the high-pass and low-pass filtered signal. In poor signal-to-noise ratios (SNRs), the proposed algorithm maintains a low false-alarm rate in the detection of speech pauses while the standardized algorithm of ITU G.729 shows an increasing false-alarm rate in unfavorable situations. These characteristics are found with different types of noise and indicate that the proposed algorithm is better suited to be used for noise estimation in noise reduction algorithms, as speech deteriorations may thus be kept at a low level. It is shown that in connection with the Ephraim–Malah noise reduction scheme [1], the speech pause detection performance can even be further increased by using the noise-reduced signal instead of the noisy signal as input for the speech pause decision unit.**

***Index Terms*—Envelope dynamics, envelope minima, noise estimation, noise reduction, speech pause detection.**

I. Introduction

NEW technologies in mobile telecommunication, robust speech recognition and digital hearing aids are a strongly driving force in the development of real-time noise reduction algorithms. The number of publications on single-microphone noise reduction algorithms indicates an unbroken interest in this research field over the past two or three decades. A crucial point for these kind of algorithms is the concurrent estimate of the target speech spectrum and the interfering noise spectrum in particular. Since most realistic noisy environments are characterized by nonstationarity, it is necessary to update the noise spectrum estimate as often as possible to maintain an effective noise reduction. This can be done, for example whenever target speech is absent, which means that the input signal consists of noise only. Another constraint is the limited complexity of the algorithm when it is supposed to become implemented in digital circuits. Hence, computational and memory requirements should be as low as possible.

Different algorithms have been proposed which *continuously* update the noise estimate and hence avoid the need for explicit speech pause detection. Martin [2], [3] uses the minimum of the sub-band signal power within a time window of about 1 s as an estimate of the noise power in the respective sub-band. This idea was already formulated by Paul [4]. Doblinger [5] proposed a continuous noise estimation scheme similar to Martin's which is computationally more efficient. This scheme was, however, not systematically tested. Hirsch [6] and Hirsch and Ehrlicher [7] proposed an algorithm which is based on the observation that the most commonly occurring spectral magnitude value in clean speech is zero. Hence, having noisy speech their algorithm measures the distribution density function of the spectral magnitude and determines the maxima which are then used as an estimate of the respective noise magnitude. These kind of algorithms which avoid speech pause detection for noise estimation are supposed to cope better with nonstationary (i.e., fluctuating) noise, since they are generally faster in their adaptation to changing noise levels even during speech activity. On the other hand, the continuous update of the noise estimate (independently in the sub-bands) is susceptible to erroneously capture speech energy. This, however, leads inevitably to speech deterioration in a subsequent noise reduction process. Fischer and Stahl [8] investigated a spectral subtraction noise reduction algorithm with a continuous noise spectrum updating scheme. They found that the corruption of the noise estimate by speech is too large to be further considered and conclude that voice activity detection plays an important role and cannot be fully omitted. Recently, Nemer *et al.* [9] proposed to use the kurtosis (fourth-order statistics) of the noisy signal to continuously estimate speech and noise energies. The examples presented used noisy speech signals with positive signal-to-noise ratios (SNRs) and yield promising results, but further research is required to extend these results to negative SNRs and different classes of noise, respectively.

Most authors reporting on noise reduction refer to speech pause detection when dealing with the problem of noise estimation. As Hirsch [6] pointed out, "this is a very difficult and ultimately unsolved problem for realistic situations with a varying noise level." A lot of studies thus evade the problem by using an ideal speech pause detection using the clean speech signal or by using only short test signals with an initial noise-only period for noise estimation without the need for updating the noise spectrum estimate. In some applications like audio restoration (e.g., restoration of old gramophone recordings) the noise estimation indeed can often be done "manually" off-line. However, other applications like noise reduction for mobile communication and for digital hearing aids require automatic updating of the noise spectrum estimate. Most authors agree that voice activity or speech pause detectors, respectively, are a very sensitive and often limiting part of systems for the reduction of additive noise in speech [10], [11].

Manuscript received May 15, 2001; revised September 24, 2001. This work was supported in part by a research grant from GN ReSound. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hynek Hermansky.

The authors are with the Medical Physics Department, Carl von Ossietzky University Oldenburg, D-26111 Oldenburg, Germany (e-mail: mark@medi.physik.uni-oldenburg.de; Birger.Kollmeier@uni-oldenburg.de).

Publisher Item Identifier S 1063-6676(02)01523-7.

Various procedures for speech pause detection have been described in the literature so far. Kang and Fransen [12] proposed a very simple scheme. Whenever the low-pass band energy (in the frequency range from 0 to 1 kHz) of a current signal frame is below a specific fraction of the low-pass band dynamic range as scanned in the past frames, the frame is used for updating the noise spectrum estimate. Obviously, this procedure has strong limitations. It will only work with higher SNRs and will fail in noises with prominently low frequencies. A more elaborate algorithm using adaptive energy thresholds was proposed by van Gerven and Xie [13]. Elberling *et al.* [14] used the so-called synchro method for spectral estimation of the background noise. This procedure makes use of the specific characteristic of voiced speech sounds, i.e., that the energy is confined to pitch-harmonic frequencies. Based on successive multiplication of the envelopes from neighboring pairs of band-pass signals, followed by a summation over all resulting signal-products, a global measure of energy synchronization is obtained which is then used to classify the time frames of the input signal into those dominated by speech (high synchronization) and those not dominated by speech (low synchronization). This patent application is reported to work successfully in SNRs ranging from $+9$ to -9 dB with various noises. However, an increase of wrong speech pause decisions with decreasing SNR is reported. Sheikhzadeh *et al.* [15] proposed a pause detection algorithm based on an auto-correlation voicing detection which was performed on the enhanced signal (i.e., after the noise reduction rather than on the noisy signal). Although extensive testing is mentioned, no performance results are presented. However, the authors state that the algorithm is not supposed to work well below SNRs of 0 dB. Dendrinos and Bakamidis [10] presented an algorithm for determining the starting and ending points of speech segments in colored-noise environments through singular value decomposition based on some thresholds which have been determined experimentally. Good performance was proved for SNRs higher than 0 dB. However, the complexity of the algorithm makes a real-time implementation difficult. Recently, El-Maleh and Kabal [16] performed a comparative study of three voice activity detection (VAD) algorithms: a VAD used in the GSM cellular system [17], the VAD used in the enhanced variable rate codec (EVRC) of the North American CDMA-based PCS and cellular systems [18], and a third-order statistics based VAD [19]. Unfortunately, the authors did not investigate false-alarm rates and hit rates systematically but present only some noisy waveforms with the respective VAD decisions. However, the EVRC VAD is reported to show consistent superiority over the other VADs. Davídek *et al.* [20] implemented a speech activity detector using cepstral coefficients for use in a real-time noise cancellation system. However, a comprehensive evaluation of the detector itself is not given. Abdallah *et al.* [21] introduced a local entropic criterion for speech signal detection. Very good performance down to SNRs of -20 dB is reported. However, only white noise was tested so far. McKinley and Whipple [22] suggested a model based speech pause detection algorithm which is claimed to be robust for low SNRs. The speech pause detection problem is formulated into a decision theory framework. However, this algorithm requires extensive training of a Hidden Markov Model with the set of speech prototypes to be encountered. Itoh and Mizushima [23] proposed a speech/nonspeech identification based on four different parameters. The first is the maximum value of the auto-correlation function of the LPC residual signal, which represents the degree of the periodicity of the signal waveform. Second is a spectral slope parameter, third is a reflection coefficient which itself is computed from some PARCOR coefficients, and fourth is the signal energy. For each of the parameters, Itoh and Mizushima used empirically determined thresholds for a speech/stationary noise/nonstationary noise decision. It seems, however, that the decision for nonstationary noise is made only on the basis of the spectral slope parameter. Unfortunately, the proposed algorithm was not tested in low SNR situations.

Irrespective of the actual kind of speech pause detector used, a comprehensive and fair evaluation should include its hit rate as well as its false-alarm rate using different noises with a large variety of SNRs. These measures reveal most of an algorithm's capabilities and deficiencies. For an application in noise reduction, the problem is that a speech pause detection algorithm with a high false-alarm rate results in remarkably deteriorated speech after the noise reduction. On the other hand, a speech pause detection algorithm that finds too few of the actual speech pauses results in worse reduction of the noise. Hence, noise estimation is a very sensitive stage in the noise reduction process.

The algorithm for speech pause detection that will be described in the next section dynamically tracks the dynamics of the signal's temporal power envelope as well as of its low- and high-pass frequency band power envelopes. After a number of threshold comparisons, a frame-by-frame decision is made on the presence of a speech pause. This approach was motivated by the work of Festen *et al.* [24], who used the minima in the signal envelope for estimating the noise level in a speech-plus-noise signal to control an AGC (automatic gain control) algorithm for hearing aids. The proposed algorithm can be regarded as an extension of the simple scheme proposed by Kang and Fransen [12]. In order to assess its applicability to real-time noise reduction for practical applications (see above), both the hit rate and false-alarm rate are evaluated for a large range of SNRs and different types of noise and compared to a voice activity detector (VAD) algorithm recommended by the International Telecommunication Union [25].

II. Algorithm

The speech pause detection algorithm calculates the signal's temporal power envelope $E(p)$ by summing up the squares of the spectral components of the input signal in each short-time frame p

$$E(p) = \sum_k |X(p, \omega_k)|^2. \tag{1}$$

Here, $X(p, \omega_k)$ denotes the spectral component of the noisy input signal at frequency ω_k at time frame p. In addition, a low-pass band power envelope and a high-pass band power envelope are calculated:

$$E_{\mathrm{LP}}(p) = \sum_l |X(p, \omega_l)|^2 \tag{2}$$

$$E_{\mathrm{HP}}(p) = \sum_m |X(p, \omega_m)|^2 \tag{3}$$

where l runs over all spectral components up to the cut-off frequency, and m runs over the remaining spectral components. In order to slightly smooth the envelopes, $E(p)$, $E_{\mathrm{LP}}(p)$ and $E_{\mathrm{HP}}(p)$ are averaged over a few frames by a recursive low-pass filter of first order with a release time constant τ_E; no smoothing is performed in case of an increase in energy (i.e., attack time zero) to avoid smearing over onsets. The algorithm tracks the minimum value and the maximum value of each envelope and uses these for the speech pause decision as described by the following scheme.

1) After an assumed 200 ms initial phase of noise only the minimum and maximum values are set as follows:

$$\begin{aligned} E_{\min}(p) &\equiv E(p) & E_{\max}(p) &\equiv E(p) \\ E_{\mathrm{LP},\min}(p) &\equiv E_{\mathrm{LP}}(p) & E_{\mathrm{LP},\max}(p) &\equiv E_{\mathrm{LP}}(p) \\ E_{\mathrm{HP},\min}(p) &\equiv E_{\mathrm{HP}}(p) & E_{\mathrm{HP},\max}(p) &\equiv E_{\mathrm{HP}}(p). \end{aligned} \quad (4)$$

   This guarantees that the minimum envelope values correspond roughly with the noise energy at the beginning.
2) The minimum and maximum values are updated for each of the three envelopes in the following manner.
   - If the current envelope value is larger than the maximum value for the corresponding envelope, then the maximum value is set to the current value. Otherwise, the maximum value slowly decays. This is done by a recursive low-pass filter of first order with a release time constant τ_{decay}, which takes as input the current envelope value.
   - If the current envelope value is smaller than the minimum value for the corresponding envelope, then the minimum value is set to the current value. Otherwise, the minimum value is slowly raised. This is done by a recursive low-pass filter of first order with attack time constant τ_{raise}, which takes as input the current envelope value.
3) The differences between the maximum and the minimum values are calculated for each envelope

$$\begin{aligned} \Delta(p) &= E_{\max}(p) - E_{\min}(p) \\ \Delta_{\mathrm{LP}}(p) &= E_{\mathrm{LP},\max}(p) - E_{\mathrm{LP},\min}(p) \\ \Delta_{\mathrm{HP}}(p) &= E_{\mathrm{HP},\max}(p) - E_{\mathrm{HP},\min}(p). \end{aligned} \quad (5)$$

4) Three different criteria are introduced of which only one has to be true for making the decision that target speech is not present in the actual frame: a) the speech pause decision can be made because of a low signal dynamics in both the low-pass and the high-pass band (*Dyn Speech Pause*); b) the decision can be based on the low-pass band information (*LP Speech Pause*); and c) it can be made upon the high-band information (*HP Speech Pause*). These decision criteria are derived as follows.
   a) If Δ_{LP} is smaller than some threshold η and also $\Delta_{\mathrm{HP}} < \eta$ then it is assumed that only noise is present due to the very small dynamic range of the signal ($\Rightarrow$ *Dyn Speech Pause*).
   b) If a) is not true, it is checked whether Δ_{LP} is bigger than η (otherwise the dynamic range in the low-pass band is very small and it should not receive too much attention $\Rightarrow$ *no LP Speech Pause*). Now, if the difference between the current $E_{\mathrm{LP}}(p)$ and $E_{\mathrm{LP},\min}(p)$ of the low-pass band envelope is smaller than some fraction pc of Δ_{LP} (which means that the actual envelope is near its minimum), a closer look at the high-pass band is necessary to support a speech pause decision.

      Case 1) Δ_{HP} *of the high-pass band is smaller than threshold* η.

      In this case no additional information can be obtained from the high-pass band because of its small dynamic range. Now, if at least $E(p)$ (the signal's envelope) lies in the lower half of its dynamic range [i.e., in the lower half between $E_{\min}(p)$ and $E_{\max}(p)$] the current frame can be assumed to be a speech pause because of the closeness of the low-pass band energy to its minimum value ($\Rightarrow$ *LP Speech Pause*) otherwise, however, there is not enough support for a speech pause decision ($\Rightarrow$ *no LP Speech Pause*).

      Case 2) Δ_{HP} *is bigger than two times the threshold* η.

      In this case, there is enough dynamic range to pay attention to the high-pass band. Thus, it is demanded that the difference between the current $E_{\mathrm{HP}}(p)$ and $E_{\mathrm{HP},\min}(p)$ of the high-pass envelope is smaller than two times the fraction pc of Δ_{HP} to support the small envelope value in the low-pass band. Then a noise-only frame is assumed ($\Rightarrow$ *LP Speech Pause*). This demand is not as strict as that for the low-pass band, to account for the case that the disturbing noise has a rather high-frequency characteristic. But if this condition is not fulfilled, speech may be present in the actual frame ($\Rightarrow$ *no LP Speech Pause*).

      Case 3) Δ_{HP} *is smaller than two times the threshold* η*, but bigger than* η.

      In this case, which is not as clear as Case 2, it is only demanded that $E_{\mathrm{HP}}(p)$ (the high-pass envelope) lies in the lower half of its dynamic range to support the small envelope value in the low-pass band. Then it is assumed that target speech is absent ($\Rightarrow$ *LP Speech Pause*). However, if this condition is not fulfilled, speech may be present in the actual frame ($\Rightarrow$ *no LP Speech Pause*).

   c) Condition b) accounts for the case that the disturbing noise has a rather high-frequency characteristic, hence the speech pause decision should mainly be made upon the information in the

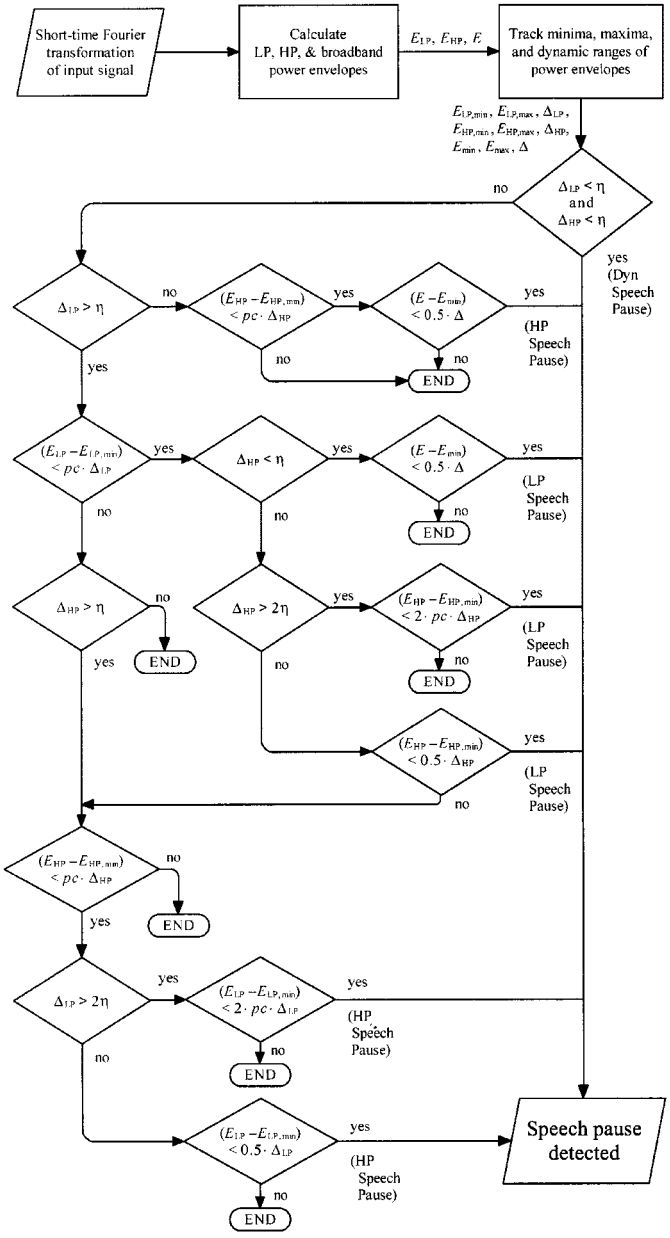

Fig. 1. Flowchart of the proposed speech pause detection algorithm operating on a single time frame. See text for details.

low-pass band. To account also for the case that it has a rather low-frequency characteristic, the same conditions as under condition b) have to be checked but now with reverse roles of the low-pass and the high-pass bands to determine whether target speech is absent (*HP Speech Pause*).

Fig. 1 gives a flowchart of the proposed speech pause detection algorithm. The flowchart is not fully symmetrical with respect to LP and HP speech pause detection since several redundant tests are omitted.

Due to its flexible design this novel approach for speech pause detection can easily be adjusted to obtain a rather low false-alarm rate by adapting the main parameters η and pc. Generally, a low false-alarm rate is desirable to reduce speech distortions in the subsequent noise reduction process. However, this also results in a reduced hit rate.

During the development of the algorithm noisy signals generated from various different noise types and speech signals at several SNRs were used for performance verification. Finally, the following values were chosen for the free parameters: The input signal was digitized with a sampling frequency of 22 050 Hz and partitioned in Hann-windowed segments of length 8 ms with 4 ms overlap. These segments were padded with zeros and a 256-point FFT was performed. This framework is compatible with most single-microphone noise reduction algorithms which can thus easily be integrated. Such short segments are motivated by the fact that then the same signal analysis and synthesis as necessary for a real-time noise reduction environment can be used. Due to the longer signal delay, longer window lengths in real-time signal processing applications would cause problems with lip reading and would cause stuttering when speaking. The cut-off frequency between low-pass and high-pass band was set to 2 kHz, motivated by the fact that excluding speech frequencies above 1.9 kHz has a roughly similar effect on speech intelligibility as excluding those below this value [26]. The time constant τ_E for the envelope smoothing was set to 32 ms. The time constants τ_{decay} and τ_{raise} were both set to 3 s. These constants were determined by examination of the envelopes from several speech samples. With these settings a good approximation to the actual dynamic range of the signal and of its "placement" in the level area under a variety of conditions was achieved. However, systematic variations of these parameters were not investigated. The threshold η was set to 5 dB and the fraction pc was set to 0.1.

III. Examples

To illustrate the speech pause detection scheme, Figs. 3–5 show some detection examples using a target sentence of approximately 5 s length mixed with different noises (digitally added).

Fig. 3 shows an example with car noise. This type of noise was recorded in the cabin of a driving car and has dominant parts in the low frequency range. The bar at the bottom of the panels shows the real speech pauses which were determined manually. [For comparison, the waveform of the clean sentence is displayed in Fig. 2 (upper panel); the lower panel shows the mixed signal with a SNR of -5 dB.] The speech pause decisions of the algorithm are displayed in the other bottom three bars. The distinct bars give additional information about the reason for the speech pause decision. The first bar shows a symbol whenever a speech pause is detected due to a small dynamic range of the signal in the low-pass band as well as in the high-pass band, and generally in the initial noise estimation phase (the first 200 ms). The second bar shows a symbol whenever a speech pause is detected on the basis of the low-pass band information. Finally, a symbol in the third bar means that the decision was based on the high-pass band information.

The car noise example shows that it is worthwhile to consider band-limited envelopes. In this case, the signal's low-pass band envelope (as well as its broadband envelope) are strongly disturbed by the noise. However, the high-pass envelope is "clean enough" for obtaining reliable speech pause decisions (Fig. 3). Actually, the third bar in the figure panels shows that the decision is mainly based on the high-pass information.

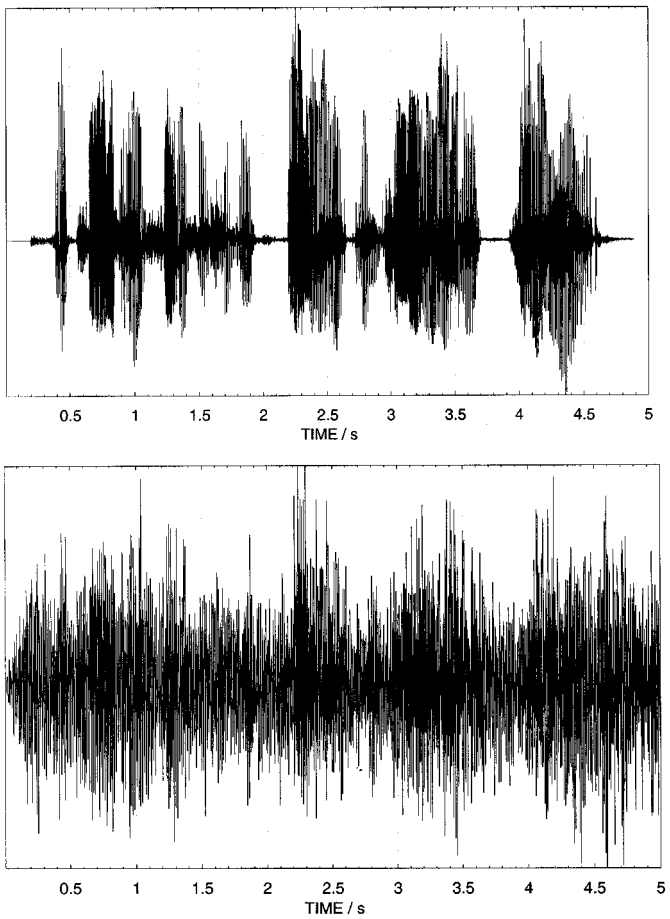

Fig. 2. Upper panel: Waveform of the sentence "I played in a theater festival, honoring the German writer Heiner Müller." Lower panel: Sentence mixed with car noise at -5 dB SNR.

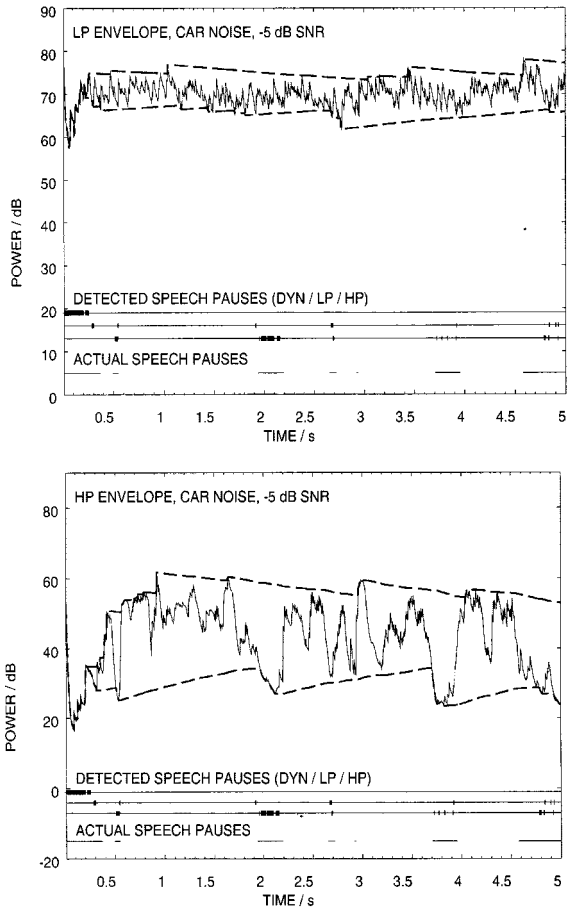

Fig. 3. Low-pass band power envelope (upper panel) and high-pass band power envelope (lower panel) of the sentence displayed in Fig. 2 when mixed with car noise at -5 dB SNR (solid curves). The dashed curves display $E_{\mathrm{HP},\min}$ and $E_{\mathrm{HP},\max}$, respectively. The detected as well as the actual speech pauses are displayed in the additional bars (see text for details).

Fig. 4 shows an example, where the sentence is mixed with the noise of a drilling machine at $+5$ dB SNR. This noise makes it impossible to get reliable speech pause information from the high-pass channel, but in this case the low-pass band information can be used. Comparison with the lowest bar in the figures (the "true" speech pauses) shows that a good speech pause detection is obtained. Although the algorithm wrongly considers the time frames around 0.6 s ("p" from "played"), 1.2 s ("th" from "theater") and around 1.5 s ("f" from "festival") as noise, these speech parts actually sound very similar to equally short segments of the drill noise. Hence, these wrong decisions are assumed to have no adverse effects on the speech quality when used for noise estimation in a noise reduction algorithm.

Fig. 5 shows an example with restaurant noise, which is neither mainly low-frequency nor high-frequency in its characteristics. As can be seen at the second and third bar in the figures, the speech pause detection, indeed, is sometimes based on the low-pass band information and sometimes on the high-pass information. In combination, a good speech pause detection performance is obtained.

IV. Comparison With G.729 VAD Algorithm

In 1996 the International Telecommunication Union (ITU) "standardized" a voice activity detector (VAD) algorithm for a speech coding scheme as its Recommendation G.729 Annex B [25]. The VAD algorithm makes a voice activity decision every 10 ms based on differential parameters of the full-band energy, the low-pass band energy, the zero-crossing rate and a spectral distortion measure. These are obtained at each frame as differences between each parameter and its respective long-term average. The output of the VAD module is either 1 or 0, indicating the presence or absence of voice activity, respectively. Several publications compared their own algorithms with the G.729 VAD so far [27], [28].

Using the G.729 algorithm here as a competitor is motivated by the fact that it has proven being successful in a wide range of conditions and that it is available from the ITU. Comparing a novel algorithm with this "standard" makes it also comparable to other algorithms, if these are tested against this "standard." Of course, the G.729 algorithm was intended to be used in less noisy environments, originally.

A. Procedure

A female reading of a short story (41 s length) from the German PhonDat database [29] was used to test the performance of the proposed algorithm versus the G.729 algorithm. The speech signal was mixed with a car noise, a multi-talker babble noise, an aircraft engine noise, and a factory noise,

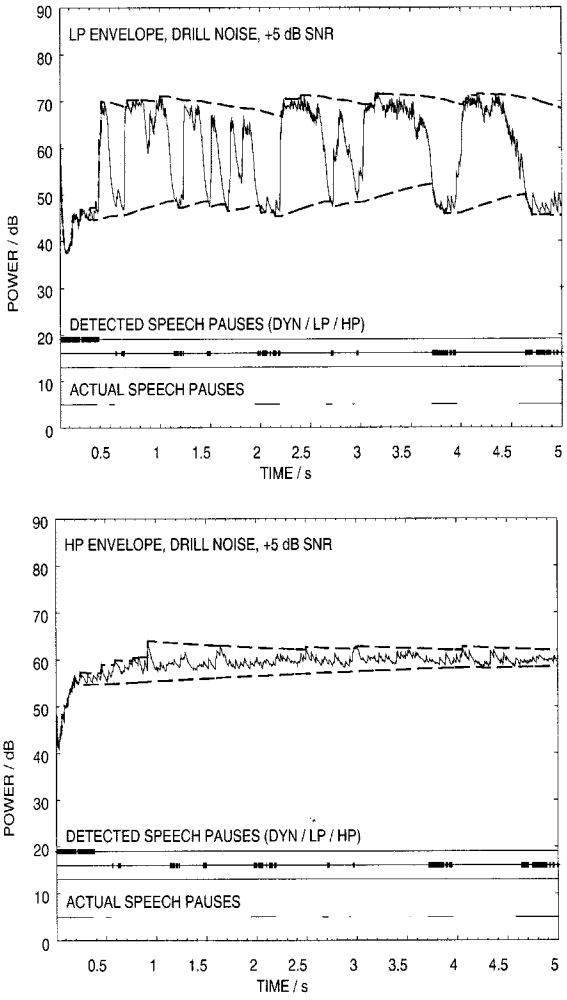

Fig. 4. Low-pass band power envelope (upper panel) and high-pass band power envelope (lower panel) of the sentence displayed in Fig. 2 when mixed with drilling machine noise at +5 dB SNR (solid curves). The dashed curves display $E_{\mathrm{HP},\min}$ and $E_{\mathrm{HP},\max}$, respectively. The detected as well as the actual speech pauses are displayed in the additional bars (see text for details).

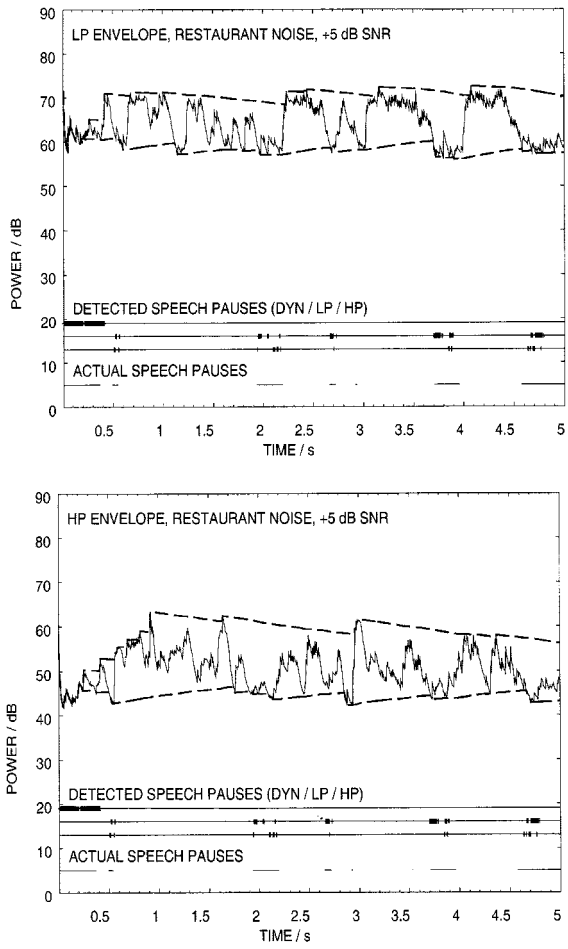

Fig. 5. Low-pass band power envelope (upper panel) and high-pass band power envelope (lower panel) of the sentence displayed in Fig. 2 when mixed with restaurant noise at +5 dB SNR (solid curves). The dashed curves display $E_{\mathrm{HP},\min}$ and $E_{\mathrm{HP},\max}$, respectively. The detected as well as the actual speech pauses are displayed in the additional bars. See text for details.

respectively, which were taken from the NOISEX-92 database [30]. SNRs from −10 dB to +20 dB were employed. Negative SNRs do often occur in real-life situations and especially hearing-impaired persons have enormous problems to have conversations in noisy environments. Of course, the frequency shape of a noise signal has a strong influence on its masking effect. While the speech reception threshold (i.e., SNR where 50% of the speech are intelligible) for some machinery noises can be very low (for drill noise it is about −20 dB; [31]), for cafeteria noise, e.g., it may be much higher (about −4 dB; [31]) but still negative.

False-alarm rates (i.e., the fraction of all real speech frames that were erroneously detected as speech pauses) and hit rates (i.e., the fraction of all real speech pauses that were correctly detected as speech pauses) were determined in each noise condition for both the proposed algorithm and the G.729 algorithm. For the calculation of the false-alarm rate as well as the hit rate, the "real" speech frames and "real" speech pauses were determined using the G.729 VAD algorithm on the clean speech signal. Using the G.729 itself as reference takes into consideration that no simple rule exists even for determining pauses in clean speech. Since the G.729 algorithm is recommended by the ITU, it can be taken for granted that it works well for clean speech. Note, that in the comparative test with the proposed new algorithm this may give an advantage for the G.729 algorithm, as it defines the "clean" standard. Hand-labeling of the real speech pauses was not considered since an automatic procedure was much more economical for determination of even very short pauses.

Finally, both algorithms are compared in terms of receiver operating characteristics (ROC).[1]

B. Results

The detection results are shown in Figs. 6 and 7. The upper panels show the false-alarm rate, the lower panels present the hit rate of both algorithms.

The comparison with the G.729 Annex B algorithm shows that the proposed speech pause detection algorithm yields a clearly lower false-alarm rate in each of the four different noises

[1]According to Egan [32], the receiver operating characteristic (ROC) is a function which summarizes the possible performances of an observer faced with the task of detecting a signal in noise. In general, the ROC is given as a plot of the hit rate versus the false-alarm rate which is obtained by modifying the decision criterion. In the present study, the signal to be detected is a "speech pause" occurring in a noisy speech signal.

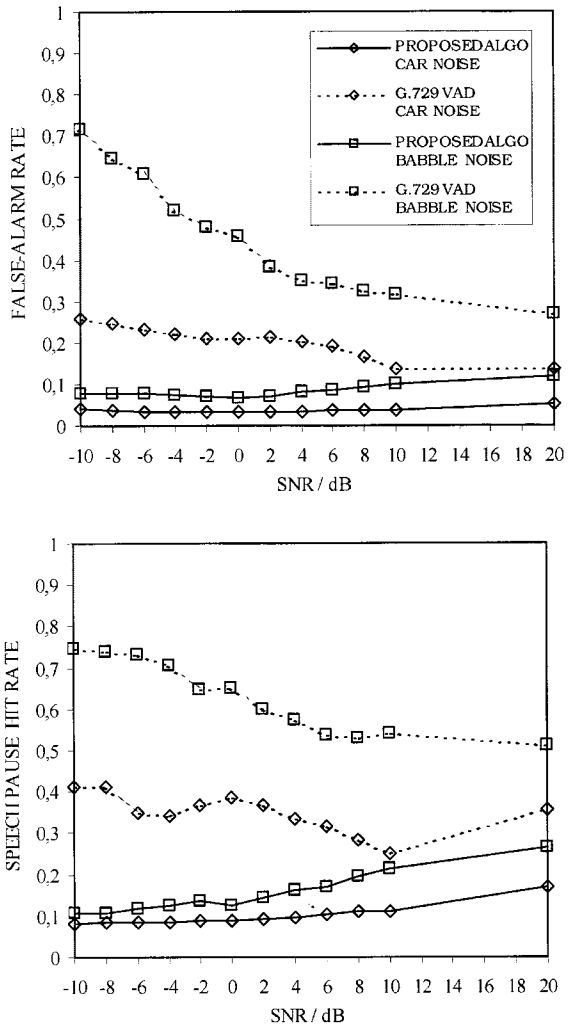

Fig. 6. Speech pause detection performance of the proposed algorithm and the G.729 VAD algorithm in car noise and multi-talker babble noise with SNRs ranging from -10 to $+20$ dB. The upper panel shows the false-alarm rates and the lower panel shows the hit rates with the respective algorithms.

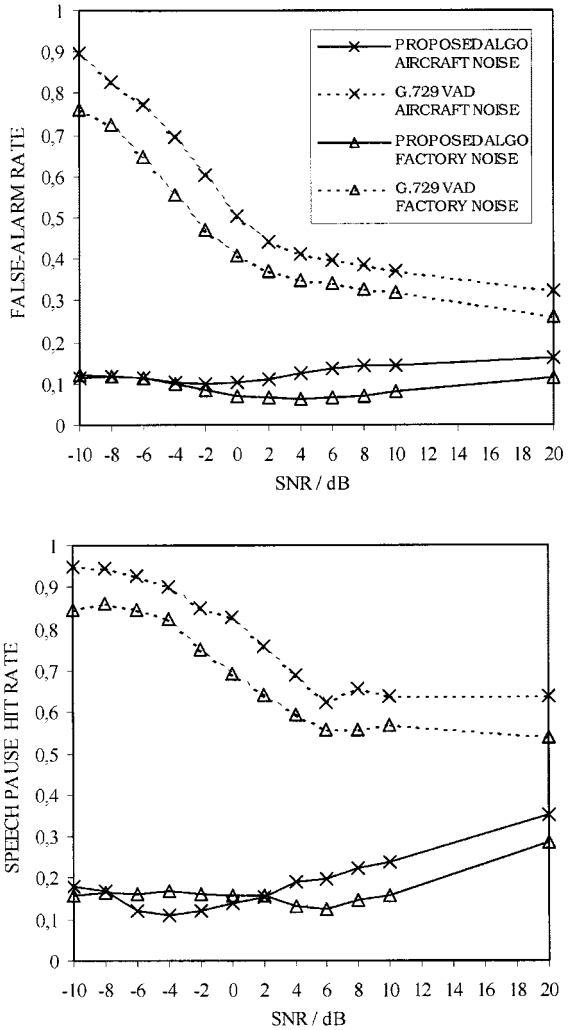

Fig. 7. Speech pause detection performance of the proposed algorithm and the G.729 VAD algorithm in aircraft engine and factory noise with SNRs ranging from -10 to $+20$ dB. The upper panel shows the false-alarm rates and the lower panel shows the hit rates with the respective algorithms.

over the entire range of SNRs that were tested (cf., Figs. 6 and 7). On the other hand, fewer speech pauses are actually detected than with the G.729 algorithm.

The false-alarm rates are lowest in car noise, followed by the multi-talker babble noise, the factory noise, and the aircraft engine noise. However, a principal difference between the algorithms is observed: While the proposed algorithm keeps the false-alarm rate and the hit rate almost constant with changing SNR, the performance of the G.729 algorithm strongly depends on the SNR—the lower the SNR, the larger the false-alarm rate as well as the hit rate. It is striking that the performance of the G.729 algorithm in car noise is rather poor even at moderate noise levels of $+20$ dB.

In terms of receiver operating characteristics (ROC), the working point of the G.729 algorithm shifts up and to the right in ROC space with decreasing SNR, while the working point of the proposed algorithm stays nearly at the same place in ROC space. In general, the false-alarm rates can be decreased by changing threshold criteria in the algorithm's decision rules. This is, of course, connected with a decrease of the hit rates. Whether the proposed algorithm is generally "better" than the G.729 algorithm can be examined by comparing them in ROC space (in terms of discriminability, i.e., the area under the ROC curve). Figs. 8–10 show ROC curves of the proposed algorithm using car noise, babble noise, and aircraft noise, respectively. The upper panels were obtained at SNRs of -10 dB; for the lower panels SNRs of $+10$ dB were used. The curves were generated by varying the threshold η in the decision rule of the proposed algorithm (cf., Section II) from 1 to 25 dB in 1-dB steps.

Since in all noise conditions the G.729 algorithm falls below the ROC curve of the proposed algorithm, it may be concluded that the discriminability is better with the proposed speech pause detection algorithm.

Additionally, in Fig. 10 (upper panel) the ROC curve was determined for the proposed algorithm using a noise-reduced signal as input for the speech pause detection (by employing the single-microphone noise reduction algorithm from Ephraim and Malah [1], on a frame-by-frame basis) instead of the noisy signal. The detected speech pauses are in turn used to adjust the noise spectrum estimate for the noise reduction. Although this leads to a recursive design of the signal flow, no stability

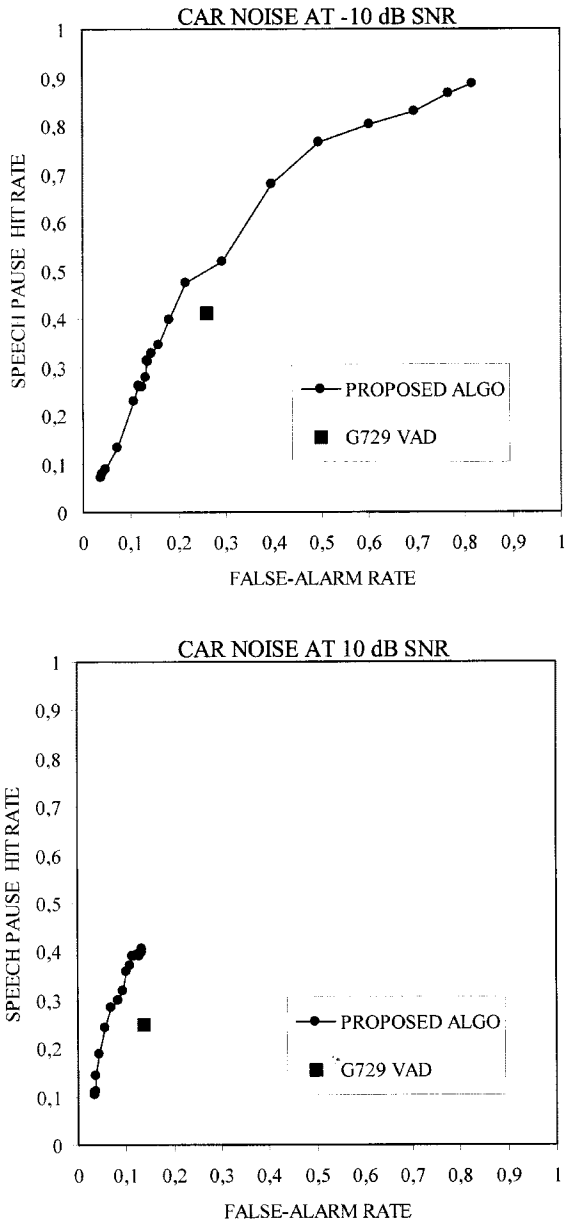

Fig. 8. ROC curve of the proposed algorithm using car noise at −10 dB SNR (upper panel) and +10 dB SNR (lower panel). The curve was generated by varying the threshold η in the decision rule from 1 to 25 dB in 1-dB steps. For comparison, the performance of the G.729 VAD algorithm is also indicated.

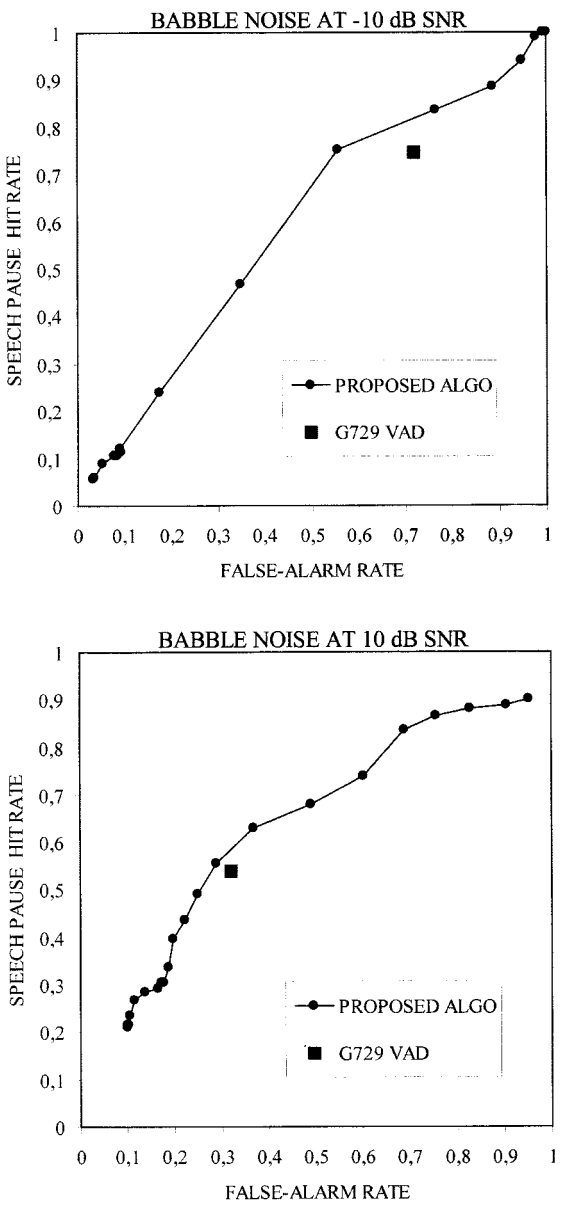

Fig. 9. ROC curve of the proposed algorithm using babble noise at −10 dB SNR (upper panel) and +10 dB SNR (lower panel). The curve was generated by varying the threshold η in the decision rule from 1 to 25 dB in 1-dB steps. For comparison, the performance of the G.729 VAD algorithm is also indicated.

problems were observed for a wide range of input signals and SNRs.

This modified algorithm is denoted as "Proposed Algo NR." Actually, the discriminability of the speech pause detection algorithm is further increased by this modification as can be seen at the larger area under the ROC curve (cf., Fig. 10, upper panel).

C. Discussion

In a noise estimation application for noise reduction algorithms it is generally proposed to operate the speech pause detection at rather low hit rates to keep the false-alarm rate low. Large false-alarm rates in the speech pause detection lead to wrong noise spectrum estimates which include significant speech parts and hence cause artifacts in a subsequent noise reduction process. In fact, the proposed speech pause detection algorithm maintains a low false-alarm rate over a wide range of SNRs while the hit rate decreases only slightly at poorer SNRs. Hence, the algorithm keeps a relatively fixed position in ROC space over a wide range of SNRs. In contrast to the proposed algorithm, the algorithm of the ITU Recommendation G.729 yields very large false-alarm rates (but also larger hit rates) at low SNRs.

Obviously, the G.729 was not designed to detect the true speech pauses in adverse noise conditions. In conditions where the speech is hardly noticeable, the G.729 VAD algorithm rather decides to classify this situation as speech-free (i.e., a kind of extended speech pause). Since this behavior is inherent in the algorithmic design of the G.729 scheme, it cannot be overcome by global changes of its threshold parameters. In a noise reduc-

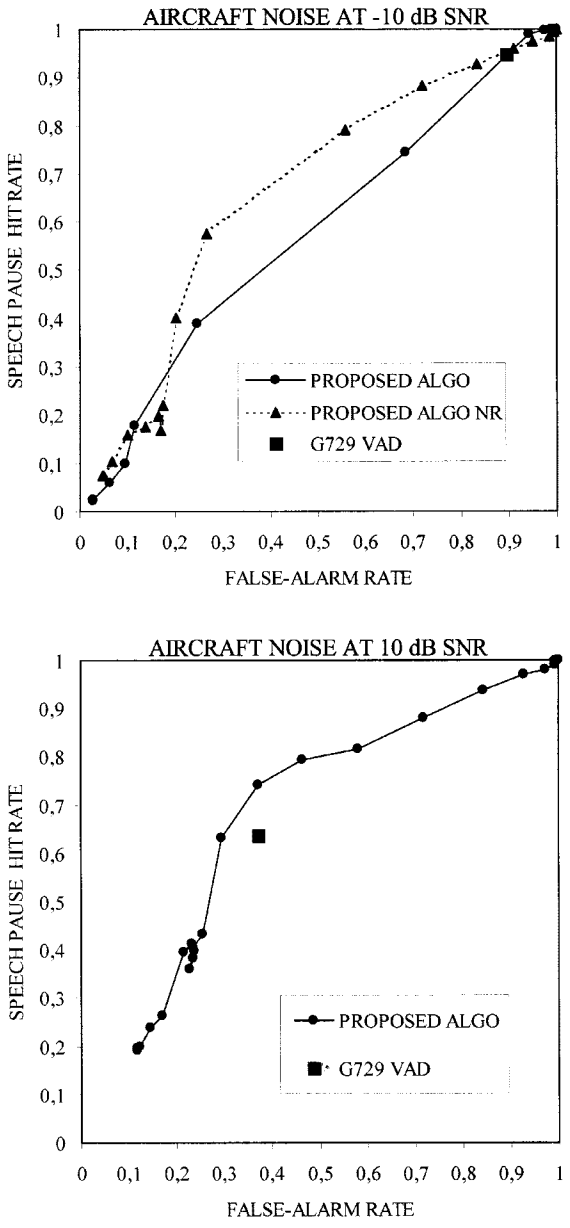

Fig. 10. ROC curve of the proposed algorithm using aircraft noise at -10 dB SNR (upper panel) and $+10$ dB SNR (lower panel). The curve was generated by varying the threshold η in the decision rule from 1 to 25 dB in 1-dB steps. For comparison, the performance of the G.729 VAD algorithm is also indicated.

tion application, this behavior probably makes it impossible for a noise reduction algorithm to "retrieve" the speech signal, if the whole signal is classified as noise. As the proposed algorithm detects speech pauses by tracking envelope minima, its behavior at very poor SNRs differs here. It still decides for speech pauses only when energy minima occur.

The threshold parameters in the proposed speech pause detection algorithm were determined empirically to obtain low false-alarm rates for a wide range of input signals and SNRs. By this, speech deteriorations due to wrong noise spectrum estimates (i.e., including speech energy) in any subsequent noise reduction processing are minimized. However, low false-alarm rates are connected with lower hit rates which could also lead to signal deteriorations for certain types of strongly fluctuating noises. If the noise is strongly fluctuating in its characteristics between speech pauses, a noise estimate determined only when speech is absent is not sufficient to ensure effective noise reduction. For such conditions, noise reduction schemes have to be employed which exploit other features (for example separation in space between noise and target source [33]), or a running noise estimate has to be determined from the noisy signal and not only during speech pauses.

Apart from that, low hit rates in the proposed algorithm do not necessarily mean that some speech pause intervals are not detected at all, but rather that several frames *during* speech pauses are not detected as such (see for example Fig. 3). For the adjustment of a noise spectrum estimate, the proposed algorithm can hence be employed at rather low hit rates to obtain low false-alarm rates and still detects at least some frames during most speech pauses. The proposed algorithm has successfully been employed in several experiments with single-microphone noise reduction algorithms [31].

It might seem strange that the false-alarm rates of the proposed algorithm increase slightly for *better* SNRs, but this is due to the fact that the G.729 defines the clean reference. Very soft consonant parts (with insignificant low energy) are classified as speech pause by the proposed algorithm. However, these parts are classified as speech by the G.729 algorithm.

V. Conclusions

The proposed speech pause detection algorithm maintains a low and approximately constant false-alarm rate over a wide range of SNRs. The hit rate decreases only slightly at poorer SNRs.

Since the proposed speech pause detection algorithm was shown to be superior to the G.729 VAD algorithm in terms of discriminability (area under the ROC curve) in speech with noise, it should be preferred in applications where noise disturbances may occur.

The performance can be further enhanced if the algorithm is combined with the single-microphone noise reduction algorithm proposed by Ephraim and Malah [1] and the noise reduced signal is employed for the speech pause detection.

The relatively low complexity of the algorithm should allow an immediate application in, for example, digital hearing aids or cellular phones. The delay time due to the signal processing is below 10 ms.

Acknowledgment

The authors thank the anonymous reviewers for critical reading of the manuscript and for their helpful comments.

References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 1109–1121, June 1984.

[2] R. Martin, "An efficient algorithm to estimate the instantaneous SNR of speech signals," in *Proc. EUROSPEECH'93*, vol. 1, 1993.

[3] ——, "Spectral subtraction based on minimum statistics," in *Signal Processing VII, Theories and Applications. Proceedings of EUSIPCO-94*, vol. 1, M. J. J. Holt, C. F. N. Cowan, P. M. Grant, and W. A. Sandham, Eds. Lausanne, Switzerland, 1994.

[4] D. B. Paul, "The spectral envelope estimation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 786–794, Apr. 1981.
[5] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," in *Proc. 4th Eur. Conf. Speech Communication Technology EUROSPEECH'95*. Madrid, Spain, Sept. 1995, pp. 1513–1516.
[6] H. G. Hirsch, "Estimation of noise spectrum and its application to SNR-estimation and speech enhancement," Int. Comput. Sci. Inst., Berkeley, CA, Tech. Rep. TR-93-012, 1993.
[7] H. G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing 1995*, vol. 1, 1995, pp. 153–156.
[8] A. Fischer and V. Stahl, "On improvement measures for spectral subtraction applied to robust automatic speech recognition in car environments," in *Proc. Workshop Robust Methods Speech Recognition Adverse Conditions*, Tampere, Finland, May 1999, pp. 75–78.
[9] E. Nemer, R. Goubran, and S. Mahmoud, "SNR estimation of speech signals using subbands and fourth-order statistics," *IEEE Signal Processing Lett.*, vol. 6, pp. 171–174, July 1999.
[10] M. Dendrinos and S. Bakamidis, "Voice activity detection in colored-noise environment through singular value decomposition," in *Proc. 5th Int. Conf. Signal Processing Applications and Technology*. Waltham, MA: DSP Associates, 1994, vol. 1, pp. 137–141.
[11] P. Sovka and P. Pollák, "The study of speech/pause detectors for speech enhancement methods," in *Proc. 4th Eur. Conf. Speech Communication Technology EUROSPEECH'95*. Madrid, Spain: ESCA, September 1995, pp. 1575–1578.
[12] G. S. Kang and L. J. Fransen, "Quality improvement of LPC-processed noisy speech by using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 930–942, June 1989.
[13] S. Van Gerven and F. Xie, "A comparative study of speech detection methods," in *Proc. 5th Eur. Conf. Speech Communication Technology, EUROSPEECH'97*, Rhodes, Greece, 1997.
[14] C. Elberling, C. Ludvigsen, and G. Keidser, "The design and testing of a noise reduction algorithm based on spectral subtraction," *Scand. Audiol.*, vol. Suppl. 38, pp. 39–49, 1993.
[15] H. Sheikhzadeh, R. L. Brennan, and H. Sameti, "Real-time implementation of HMM-based MMSE algorithm for speech enhancement in hearing aid applications," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing 1995*, vol. 1, 1995, pp. 808–811.
[16] K. El-Maleh and P. Kabal, "Comparison of voice activity detection algorithms for wireless personal communications systems," in *Proc. CCECE'97 Can. Conf. Electrical Computer Engineering*, vol. 2, 1997, pp. 470–473.
[17] K. Srinivasan and A. Gersho, "Voice activity detection for cellular networks," in *Proc. IEEE Speech Coding Workshop*, 1993, pp. 85–86.
[18] TIA, "Enhanced variable rate codec, speech service option 3 for wideband spread spectrum digital systems,", Document PN-3292, 1996.
[19] M. Rangoussi and G. Carayannis, "Higher order statistics based Gaussianity test applied to on-line speech processing," in *Proc. IEEE Asilomar Conf.*, 1995, pp. 303–307.
[20] V. Davídek, J. Šika, and J. Štusák, "Noise cancellation system on TMS320C31," in *Proc. 1st Eur. DSP Education Research Conf.*. Paris, France, 1996, pp. 134–138.
[21] I. Abdallah, S. Montrésor, and M. Baudry, "Speech signal detection in noisy environment using a local entropic criterion," in *Proc. 5th Eur. Conf. Speech Communication Technology, EUROSPEECH'97*, Rhodes, Greece, 1997.
[22] B. L. McKinley and G. H. Whipple, "Model based speech pause detection," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing 1997*, Los Alamitos, CA, 1997, pp. 1179–1182.
[23] K. Itoh and M. Mizushima, "Environmental noise reduction based on speech/nonspeech identification for hearing aids," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing 1997, Conference Proceedings*. Los Alamitos, CA: IEEE Comput. Soc. Press, 1997, pp. 419–422.
[24] J. M. Festen, J. N. Van Dijkhuizen, and R. Plomp, "The efficiacy of a multichannel hearing aid in which the gain is controlled by the minima in the temporal envelope," *Scand. Audiol.*, vol. Suppl. 38, pp. 101–110, 1993.
[25] *ITU-T Recommendation G.729—Annex B: A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70*, 1996.
[26] D. M. Jones, "Noise," in *Stress and Fatigue in Human Performance*, R. Hockey, Ed. New York: Wiley, 1983, ch. 3, pp. 61–95.
[27] J. Stegmann and G. Schröder, "Robust voice-activity detection based on the wavelet transform," in *Proc. 1997 IEEE Workshop Speech Coding Telecommunications*, New York, 1997, pp. 99–100.
[28] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Lett.*, vol. 6, pp. 1–3, Jan. 1999.
[29] C. Draxler, "Introduction to the Verbmobil-PhonDat Database of spoken German," in *Proc. 3rd Int. Conf. Practical Application Prolog*, Paris, France, 1995, pp. 201–212.
[30] H. J. M. Steeneken and F. W. M. Geurtsen, "Description of the RSG.10 noise database," TNO Inst. Perception, Soesterberg, The Netherlands, Tech. Rep. IZF 1988-3, 1988.
[31] M. Marzinzik, "Noise reduction schemes for digital hearing aids and their use for the hearing impaired," Ph.D. dissertation, Carl von Ossietzky Universität, Oldenburg, [Online]. Available: http://docserver.bis.uni-oldenburg.de/publikationen/dissertation/2001/marnoi00/marnoi00.html, Germany, 2000.
[32] J. P. Egan, *Signal Detection Theory and ROC Analysis*. New York: Academic, 1975.
[33] T. Wittkop, "Two-channel noise reduction algorithms motivated by models of binaural interaction," Ph.D. dissertation, Carl von Ossietzky Univ., Oldenburg, Germany, 2001.

Mark Marzinzik was born in 1970 in Bremen, Germany, and studied physics from 1990 to 1996 at the Carl von Ossietzky Universität, Oldenburg, Germany. He received the Ph.D. degree in physics (supervised by B. Kollmeier) in 2000 with a dissertation on "Noise reduction schemes for digital hearing aids and their use for the hearing impaired."

He is currently a Research Associate with the Medical Physics Department, Universität Oldenburg. His studies focus on dynamic compression and noise reduction for digital hearing aids.

Birger Kollmeier was born in 1958. He received the Diplom degree in physics in 1982, the M.D. degree in 1986, the Ph.D. degree in physics (supervised by M. R. Schroeder) in 1986 and the Ph.D. degree in medicine in 1989, all from the Universität Göttingen, Germany. He received the Fulbright Scholarship and was with Washington University and Central Institute for the Deaf in St. Louis, MO, from 1982 to 1983.

He was an Assistant Professor (1986–1991) and Associate Professor (1991–1992) at the Third Physikalisches Institut, Universität Göttingen. Since 1993, he has been Full Professor of physics and Head of the Medical Physics Department at the Universität Oldenburg, Germany. He has authored or co-authored more than 100 original papers and six books and has supervised 21 completed Ph.D. dissertations.

Dr. Kollmeier is vice president of the German Audiological Society and has received various prizes and honors.