Multiband Modulation Energy Tracking for Noisy Speech Detection

Georgios Evangelopoulos, *Student Member, IEEE*, and Petros Maragos, *Fellow, IEEE*

***Abstract*—The ability to accurately locate the boundaries of speech activity is an important attribute of any modern speech recognition, processing, or transmission system. The effort in this paper is the development of efficient, sophisticated features for speech detection in noisy environments, using ideas and techniques from recent advances in speech modeling and analysis, like presence of modulations in speech formants, energy separation and multiband filtering. First we present a method, conceptually based on a classic speech–silence discrimination procedure, that uses some newly developed, short-time signal analysis tools and provide for it a detection theoretic motivation. The new energy and spectral content representations are derived through filtering the signal in various frequency bands, estimating the Teager–Kaiser Energy for each and demodulating the most active one in order to derive the signal's dominant AM–FM components. This modulation approach demonstrated an improved robustness in noise over the classic algorithm, reaching an average error reduction of 33.5% under 5–30-dB noise. Second, by incorporating alternative modulation energy features in voice activity detection, improvement in overall misclassification error of a high hit rate detector reached 7.5% and 9.5% on different benchmarks.**

***Index Terms*—Detector evaluation, energy separation algorithm (ESA), modulations, multiband demodulation, speech analysis, speech endpoint detection, Teager energy, voice activity detection (VAD).**

I. Introduction

DETECTING speech endpoints, the location of the beginning and ending time instances of speech in an acoustic background of silence, has been an important research problem with many interesting practical applications. This can be either a direct problem of labeling the boundaries of speech segments in silence and noise or an indirect one of speech versus silence classification [known as voice activity detection (VAD)]. Separation of speech from silence is considered a specific case of the more general problems of speech segmentation and event detection.

Accurate detection of speech endpoints and robust automatic segmentation, especially under noisy conditions, has come to be of importance in tasks regarding speech recognition, coding, processing, and transmission. Generally, it is critical to reduce all processing computations by selecting only the useful speech segments of a recorded signal. In word recognition or speaker verification systems, locating the speech segments is a crucial issue for the formation of speech patterns that will provide the highest recognition accuracy [1]. Estimation of speech boundaries can be helpful in segmentation of large speech databases, usually consisting of phrases, where the instances that define various speech events (e.g., isolated words) are estimated manually. It is also incorporated in various speech enhancement and manipulation techniques like noise spectrum estimation [2] and frame dropping for efficient front-ends [3], noise reduction [4], echo cancelation, energy normalization [5], and silence compression. Detecting speech in telecommunications is used for real-time speech transmission over networks, serving more customers per transmission channel, by assigning it to a new user only when silence is detected (time-assignment speech interpolation technique) [6], while in modern cellular systems (GSM, UMTS, 3GPP) voice activity detectors are used for selective power-reserving transmission [7], [8].

A broad variety of endpoint detection algorithms have been proposed by researchers in the past based on a classic algorithm developed by Rabiner and Sambur [9]. Through a procedure that involved time-domain representations of a signal and statistical characterization of a small silence segment from the beginning of an utterance, the algorithm used threshold checks to classify between speech and silence. The method displayed accurate endpoint decisions in environments where the signal-to-noise ratio (SNR) was of the order of 30 dB or better. Several approaches to improve the ideas of that basic algorithm were made aiming at increasing accuracy especially in extreme noise conditions, depending on specific applications. The major trends toward that direction focus either on the development of sophisticated features or the decision logic complexity.

In the latter, direction approaches included incorporation of semantic and syntactic constraints on the detection procedure [1], use of multiple potential endpoint ordering and recognition accuracy of the detected words [10], or three-state models (speech, silence, *and* transition state) and knowledge-based heuristics [11]. On a rather different approach, the voice detection problem was dealt through a pattern recognition framework in [12], classifying voiced or unvoiced segments using multiple features, and in [13] through a network training and classification process. To avoid empirically determined thresholds, a self-organizing neural fuzzy network was applied in [14] to label speech or nonspeech frames. A geometrically adaptive energy threshold in a fusion scheme was proposed in [15] for separating speech from silence in noise, analysis in four subbands and adaptive thresholds to improve voice

Manuscript received April 2, 2004; revised September 1, 2005. This work was supported in part by the European Network of Excellence MUSCLE, in part by the European research project HIWIRE, and in part by the National Technical University of Athens research program "Protagoras." The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Thierry Dutoit.

The authors are with the School of Electrical and Computer Engineering, National Technical University of Athens, 10682 Athens, Greece (e-mail: gevag@cs.ntua.gr; maragos@cs.ntua.gr).

Digital Object Identifier 10.1109/TASL.2006.872625

activity detectors was suggested in [16], while combinations of multiple features were tested through a CART algorithm in [17] and an HMM/TRAPS model in [18]. Finally, in [5] and [19], in the spirit of image edge detection, development of a one-dimensional optimum filter as energy edge-endpoint detector and a three-state decision logic was considered with the estimated endpoints providing reference for energy normalization and speech recognition.

Novel features for improved noisy speech detection are inspired by exploring alternative signal properties. Apart from energy and zero-crossings rate, literature includes "periodicity" and jitter [20], pitch stability [21], spatial signal correlation [4], spectral entropy [22], cepstral features [23], LPC residual [24], alternative energy measures [25], autocorrelation measures [18], temporal power envelope [2], spectral divergence [3], [26]. A time-frequency parameter was introduced in [21] and modified through multiband analysis in [14]. Recently, the statistical framework gains interest as properties of higher order statistics of speech are used in [24] to classify short segments, endpoint detection for nonstationary non-Gaussian noise is explored in [27] by means of bispectral-based energy functions, and optimized likelihood ratio rules are proposed in [28].

Motivated by recent advances in the field of nonlinear speech modeling and analysis, we approach the basic problem focusing on alternative, more sophisticated features. Our approach involves the development of new time-domain signal representations derived using demodulation of the signal in AM–FM components through the *energy separation algorithm* (ESA) [29]. The demodulation takes place in multiple frequency bands, following a multiband analysis scheme [30], to isolate the strongest modulation components in each band. Finally, a maximum average energy tracking process over the various frequency bands is used to yield short-time measurements of multiband signal modulation energy and demodulated instant amplitude and frequency.

On speech detection, the terms endpoint, voice activity, speech pause, word boundary, and silence have been used to refer to events and detection procedures depending on application priorities and specifications. Usually, voice detectors give crude estimates of speech activity, while endpoint detection involves refinements for exact boundary estimation. To verify the effectiveness of the new features, we incorporate them, in place of classic time-domain signal representations, in an endpoint locating, threshold-based algorithm. The new algorithm, combining benefits from multiband analysis and modulation modeling, displayed improved behavior regarding average detection accuracy under artificial and real noise conditions. The modulation energy was also considered as a feature for voice activity detection based on adaptive optimum thresholds for noisy speech detection [3]. Systematically evaluated under various noise levels and conditions on large databases, it consistently improved speech detection performance.

In Section II, we give a brief theoretical background on the ideas and methods mentioned, and then in Section III, we highlight the motivations for this paper and provide a theoretical interpretation using ideas from detection theory and hypothesis testing before we describe the developed new features and algorithms. Experimental results on speech endpoint detection in noise are presented in Section IV, where a method for evaluating detection performance is also proposed in the framework of the receiver operating characteristic (ROC) curves. Finally, in Section V, voice activity detection testing is presented modularly with systematic evaluation and comparisons.

II. Background

A. Classic Speech Endpoint Detection

The algorithm proposed by Rabiner and Sambur [9] incorporates the use of two short-time signal representations, namely energy, expressed via the mean square amplitude, and average zero-crossings rate. These measurements involve processing in small frames, and yield information both for the envelope variations of a signal and its frequency content. This classic endpoint detector discriminates between speech and silence by comparing these features with precomputed thresholds based on background noise statistics.

For isolated words in silence, an initial part of the recorded interval is assumed not to contain speech. Either the mean absolute amplitude (mAA) or the mean square amplitude (mSA),[1] and the average zero-crossings rate (ZR) measurements are computed for the whole signal duration while estimated statistics from these representations define activity thresholds. A double energy-threshold check performs a first discrimination between unvoiced and voiced regions. Results are then refined by a ZR check in the unvoiced-labeled regions. If a certain threshold is exceeded a specific number of times, a boundary is chosen at the first instance that value was crossed. This follows from the fact that a high zero-crossings rate, prior or after a voiced region, is strong indication for the presence of unvoiced speech energy. The algorithm's main advantages are: 1) the low computational complexity; 2) the simple structure; and 3) the ability to adjust to different, though stationary, recording conditions, as inference of background noise statistics depends only on the recording interval of an utterance.

B. Energy Operators and Multiband AM–FM Demodulation

The underlying assumption behind any short-time processing is that the speech signal possesses some kind of local stationarity for small periods (10–20 ms). The linear model for speech, according to which each speech formant is modeled through a damped cosine [6], [29], is also based on that assumption. However, experimental and theoretical indications about modulations in various scales during speech production, led to the proposal of an AM–FM modulation model for speech in [29].

Demodulation of a real-valued AM–FM signal

$$x(t) = a(t)\cos\left(\int_0^t \omega(\tau)d\tau\right) \tag{1}$$

with time varying amplitude envelope $a(t)$ and instantaneous frequency $\omega(t)$ signals, can be approached via the use of a nonlinear differential energy operator developed by Teager and

[1]In literature the term "short-time energy" is commonly used for mSA. We prefer "mean square amplitude" as more indicative of its derivation process.

Teager [31] and systematically introduced by Kaiser [32]. For continuous-time signals $x(t)$, this operator is

$$\Psi\left[x(t)\right] \equiv \left[\dot{x}(t)\right]^2 - x(t)\ddot{x}(t) \tag{2}$$

where $\dot{x}(t) = dx(t)/dt$. The Teager–Kaiser energy operator Ψ can track the instantaneous energy of a source producing an oscillation. Applied to an AM–FM signal of the form (1), Ψ yields the instantaneous source energy, i.e., $\Psi[x(t)] \approx a^2(t)\omega^2(t)$, where the approximation error becomes negligible [29], if the instantaneous amplitude $a(t)$ and instantaneous frequency $\omega(t)$ do not vary too fast or too much with respect to the average value of $\omega(t)$.

An AM–FM demodulation scheme was developed by Maragos *et al.* in [29] by separating the instantaneous energy into its amplitude and frequency components. Ψ is the main ingredient of the first ESA

$$\sqrt{\frac{\Psi\left[\dot{x}(t)\right]}{\Psi\left[x(t)\right]}} \approx \omega(t) \qquad \frac{\Psi\left[x(t)\right]}{\sqrt{\Psi\left[\dot{x}(t)\right]}} \approx |a(t)| \tag{3}$$

which can be used for signal and speech AM–FM demodulation. The instantaneous energy separation methodology has led to several classes of algorithms for demodulating discrete-time AM–FM signals

$$x[n] = x(nT) = A[n]\cos\left(\int_0^n \Omega[k]dk\right) \tag{4}$$

where the integer time indexes k and n are symbolically treated by integration as continuous variables, $A[n] = a(nT)$ and $\Omega[n] = T\omega(nT)$. A direct approach is to apply the discrete-time Teager–Kaiser operator $\Psi_d[x_n] \equiv x_n^2 - x_{n-1}x_{n+1}$, where $x_n = x[n]$, to the discrete AM–FM signal (4) and derive discrete energy equations of the form $\Psi_d[x_n] \approx A^2[n]\sin^2(\Omega[n])$. This is the basis for the *Discrete ESA* (DESA) demodulation algorithm [29]

$$\arccos\left(1 - \frac{\Psi_d[x_n - x_{n-1}] + \Psi_d[x_{n+1} - x_n]}{4\Psi_d[x_n]}\right) \approx \Omega[n] \tag{5}$$

$$\sqrt{\frac{\Psi_d[x_n]}{\sin^2\left(\Omega[n]\right)}} \approx |A[n]|\,. \tag{6}$$

The DESA is simple, computationally efficient, and has an excellent, almost instantaneous, time resolution.

In order to apply demodulation through ESA in speech or any wideband signal that can be modeled as a sum of AM–FM components, it is necessary to filter the signal and isolate specific frequency bands. After applying a bank of bandpass filters, one can either retain information from every channel or choose one for demodulation. The *multiband demodulation analysis* (MDA) scheme was introduced by Bovik *et al.* [30] as a way of capturing modulations in the presence of noise. It has been refined and extended in [33] for purposes of formant frequency and bandwidth tracking.

In this paper, MDA is applied through a filterbank of linearly-spaced Gabor filters and demodulation of the most active channel based on a decision rule. Gabor filters [34], whose impulse response $h(t)$ and frequency response $H(\omega)$ are given by

$$h(t) = \exp(-\alpha^2 t^2)\cos(\omega_c t) \tag{7}$$

$$H(\omega) = \frac{\sqrt{\pi}}{2\alpha}\left[\exp\left(-\frac{(\omega-\omega_c)^2}{4\alpha^2}\right)\exp\left(-\frac{(\omega+\omega_c)^2}{4\alpha^2}\right)\right] \tag{8}$$

with ω_c the central filter frequency and α its rms bandwidth, are chosen as an optimum candidate for being compact and smooth and attaining a minimum joint time-frequency spread [29], [30], [34].

III. Modified Speech Endpoint Detection

A. Motivations and Hypothesis Testing

Short-time features like energy, absolute amplitude, zero-crossings rate, pitch, cepstrum, and autocorrelation to name some, are tools frequently used for analysis, coding, processing, and detection of speech as a means of retaining slowly-varying, lowpass signal information. In order to effectively capture speech activity, however, one must take into account both the energy level of the excitation and its spectral content. For example, low amplitude level consonants like weak fricatives or plosives are harder to discriminate than vowels by simple energy checks.

A scheme that would treat fairly speech versus nonspeech events should attribute such low energy but high frequency-level sounds the same amount of importance and sensitivity as to stronger ones. In that framework, a recently proposed parameter in [14] called *adaptive time-frequency*, based on a feature in [21], takes into account both the time and frequency content of a signal. In our work we choose to adopt Teager's definition of the energy of a signal as the energy produced by its generating source, counting both its spectral and its magnitude level. Moreover, demodulation-decomposition of this energy can produce alternative time-domain analysis features or complements to long-established ones.

Let us first consider a model-based statistical detector of speech and explore the role of modulations in the detection process. For that, we consider a sum of AM–FM sines as a model for speech, as proposed in [29]

$$s[n] = \sum_{k=1}^{K} A_k[n]\cos\left(\Omega_{ck}\cdot n + \Phi_k[n]\right) \tag{9}$$

where k is the resonance index and K the number of speech formants. The instantaneous varying amplitude A_k and phase Φ_k (or frequency $\Omega_k = \Omega_{ck} + \partial\Phi_k/\partial n$) signals are to be estimated and detected.

In a statistical detection scheme, the aim is to detect the presence of speech in the ideal case of background stationary i.i.d. Gaussian noise $W(0, \sigma^2)$. Suppose now that a single AM–FM signal is present $(K = 1)$. That can be the case of capturing a

single modulation with a sufficiently narrowband Gabor filter. The carrier frequency Ω_{ck} of the modulated signal can then be assumed known and approximated via the central frequency of the Gabor filter. For simplicity, we also assume that the amplitude and phase signals are deterministic and stationary within each analysis frame, i.e., $A[n] \approx A$ and $\Phi[n] \approx \Phi$. With the above considerations, the unknown formant of the speech signal takes the form $s[n] = A\cos(\Omega_c n + \Phi) + B$, where B is a dc offset. The binary hypothesis problem is then formulated as sinusoid detection with unknown nonrandom parameters in Gaussian noise of unknown variance [35]

$$\begin{aligned} H_0: &\quad X[n] = W[n] \\ H_1: &\quad X[n] = W[n] + A\cos(\Omega_c n + \Phi) + B \end{aligned} \tag{10}$$

for each frame of length N, with $n = 0 \ldots N-1$, and $\{A, \Phi, B, \sigma^2\}$ is the set of deterministic parameters estimated via *Maximum Likelihood* (ML). We will use θ_i to represent the unknown set of parameters

$$\theta_0 = \{\sigma^2\} \qquad \theta_1 = \{A, \Phi, B, \sigma^2\} \tag{11}$$

under hypothesis $i = 0$, 1 and $\hat{\theta}_i$ for their ML estimates.

The conditioned probability distributions are Gaussian under both hypotheses, due to the noise assumptions, with density $p(x) = (\sqrt{2\pi}\sigma)^{-1}\exp(-x^2/2\sigma^2)$. The distribution conditioned on H_1 is

$$p(X;\theta_1|H_1) = \frac{1}{(\sqrt{2\pi}\sigma)^N} \times \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\left(X[n] - A\cos(\Omega_c n + \Phi) - B\right)^2\right]. \tag{12}$$

By maximizing (12) with respect to each of the unknown parameters, we have that $\hat{A}, \hat{\Phi}$ are the quadrature pair matched filter estimation of amplitude and phase, $\hat{B} = (1/N)\sum_{n=0}^{N-1}(X[n] - \hat{A}\cos(\Omega_c n + \hat{\Phi}))$ is an approximation of the mean of the observed data, and $\hat{\sigma}_1^2 = (1/N)\sum_{n=0}^{N-1}(X[n] - \hat{A}\cos(\Omega_c n + \hat{\Phi}) - \hat{B})^2$ is the noise variance estimated under H_1. Using these estimates, the sum in the argument of the exponential in (12) becomes

$$\begin{aligned} &\sum_{n=0}^{N-1} X[n]^2 + \sum_{n=0}^{N-1}\Big[(\hat{A}\cos(\Omega_c n + \hat{\Phi}) + \hat{B})^2 \\ &\qquad - 2X[n]\left(\hat{A}\cos(\Omega_c n + \hat{\Phi}) + \hat{B}\right)\Big] \\ &\quad \approx N\hat{\sigma}_0^2 - N\frac{\hat{A}^2}{2} \\ &\qquad + \hat{B}\sum_{n=0}^{N-1}\left(\hat{B} + 2\hat{A}\cos(\Omega_c n + \hat{\Phi}) - 2X[n]\right) \\ &\quad \approx N\left(\hat{\sigma}_0^2 - \frac{\hat{A}^2}{2} - \hat{B}^2\right) \end{aligned} \tag{13}$$

where we used the approximation from [35, pp. 265], and $\hat{\sigma}_0^2 = (1/N)\sum_{n=0}^{N-1} X[n]^2$ is the variance estimated under H_0. By (13), the maximum log-likelihood of (12) is approximated by

$$\ln p(X;\hat{\theta}_1|H_1) \approx N\frac{\hat{A}^2}{4\hat{\sigma}_1^2} + \frac{N}{2\hat{\sigma}_1^2}\left(\hat{B}^2 - \hat{\sigma}_0^2\right) - \frac{N}{2}\ln 2\pi\hat{\sigma}_1^2. \tag{14}$$

We decide in favor of hypothesis H_i that maximizes the log-likelihood function $\ln p(X;\hat{\theta}_i|H_i) = \max_{\theta_i}\ln p(X;\theta_i|H_i)$, which is the hypothesis that best models the observed data in each frame. Because the parameters estimated under the two hypotheses are coupled, as $\theta_0 \subset \theta_1$, the modeling accuracy is always better for H_1. To balance the difference in the number of estimated parameters, and thus prediction accuracy, some sort of penalization needs to be imposed on H_1. For that, we use the *generalized ML rule* and its approximation the *Minimum Description Length* (MDL) criterion [35]. In detail, we choose H_i that maximizes

$$\mathrm{MDL}(i) = \ln p(X;\hat{\theta}_i|H_i) - \frac{n_i}{2}\ln N, \qquad i = 1, 2 \tag{15}$$

with n_i the cardinality of $\hat{\theta}_i$. The higher the accuracy on the likelihood estimation, the stronger the penalty. As MDL is analogous to the log-likelihood of the hypotheses conditioned on the data, $\max_i \ln p(H_i|X) = \max_i \mathrm{MDL(i)}$. For H_1 where $n_1 = 4$, we can make the fair generalization that $N \approx p/\Omega_c$, where p is some constant and Ω_c the sinusoid carrier frequency. This rationale stems from time-frequency uncertainty which we will briefly explain.

Estimation of Ω_c depends on maximization of the spectrogram using a window of N samples. A number of properties and relationships regarding resolution of the transform can be found in [36]. For a potential sinusoid and a Gaussian or rectangular window, it is straightforward to prove that the average frequency at a given instance is always $\langle\omega\rangle_t = \Omega_c$. The frequency spread at any instance, i.e., the conditional standard deviation, depends on the length of the window; hence, in the case of a $\sigma_g^{-2}/2$ spread Gaussian, $\sigma_{\omega|t}^2 = \sigma_g^2$. The index t in these quantities refers to the given instance t. By applying the uncertainty principle for any t, we derive for the product of the average bandwidth B_t and duration T_t of the signal at any instance

$$\langle\omega^2\rangle_t T_t^2 \geqslant B_t^2 T_t^2 \geqslant \frac{1}{4}. \tag{16}$$

If we approximate the mean duration of the signal T_t by the length of the short-time window or the number of samples N available for the estimation and rewrite the mean square frequency as $\langle\omega^2\rangle_t = \sigma_{\omega|t}^2 + \langle\omega\rangle_t^2$ then

$$\left(\sigma_{\omega|t}^2 + \langle\omega\rangle_t^2\right)T_t^2 \geqslant \frac{1}{4} \Rightarrow \Omega_c^2 + \sigma_g^2 \geqslant \frac{1}{4N^2}. \tag{17}$$

The lower uncertainty bound is a fair approximation of the frequency-window link in the estimation process, something that

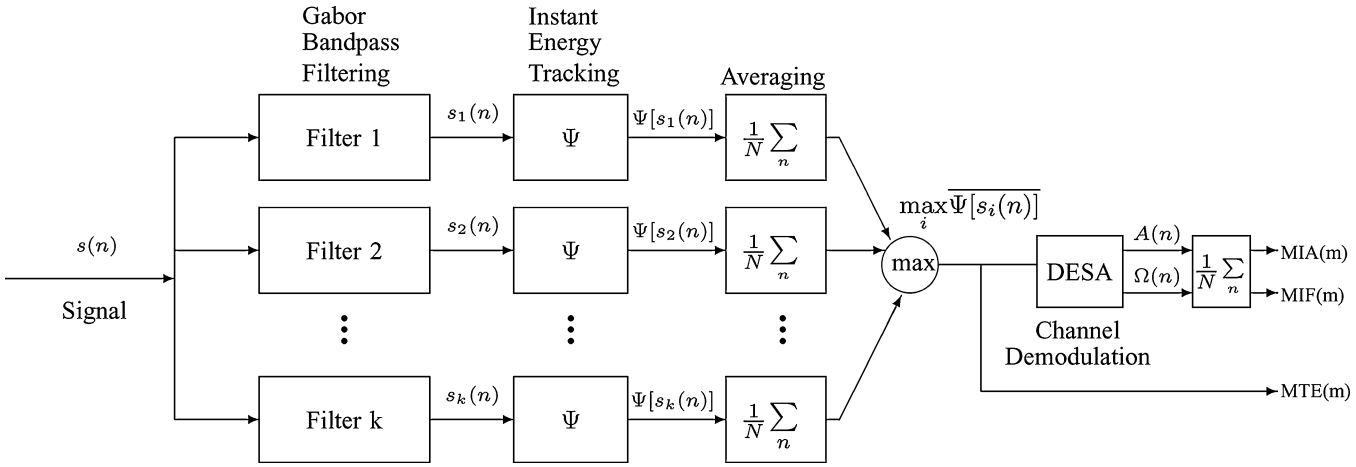


Fig. 1. Multiband filtering, modulation energy tracking, and demodulation of the filter with the *maximum average Teager energy* operator Ψ response. Averaging of the demodulated selected filter output gives values for the mean *instant amplitude* and *frequency* features. Local averaging takes place in frames of N samples.

agrees with physical intuition that locating a sinusoid of a certain frequency requires a window of size inversely proportional to the carrier frequency.

By setting this result in (15) and $n_i = 4$, we have $-\ln N^{n_i/2} = \ln 4 + \ln(\Omega_c^2 + \sigma_g^2)$ and with approximation (14)

$$\mathrm{MDL}(1) \approx N\frac{\hat{A}^2}{4\hat{\sigma}_1^2} + \ln\left(\Omega_c^2 + \sigma_g^2\right) + \frac{N}{2\hat{\sigma}_1^2}\left(\hat{B}^2 - \hat{\sigma}_0^2\right) - \frac{N}{2}\ln 2\pi\hat{\sigma}_1^2. \quad (18)$$

We can then construct a rule for the test (10) on the sinusoid-speech component detection

$$\mathrm{MDL}(1) \underset{H_0}{\overset{H_1}{\gtrless}} \mathrm{MDL}(0) \Rightarrow N\frac{\hat{A}^2}{4\hat{\sigma}_1^2} + \ln\left(\Omega_c^2 + \sigma_g^2\right) \underset{H_0}{\overset{H_1}{\gtrless}} \mathcal{O}\left(\hat{B}, \hat{\sigma}_1^2, \hat{\sigma}_0^2, N\right) \quad (19)$$

where $\mathcal{O}$ a function of statistics on the analysis frame, and $\hat{A}^2/2\hat{\sigma}_1^2$ is the SNR. The aforementioned rationale also applies for detecting one out of K sinusoids with different carrier frequencies that may stand for the different speech formants or the various passbands imposed by K Gabor filters. We then have to test $K+1$ hypothesis by maximizing the MDL criterion (15) with $n_0 = 1, n_{1:K} = 4$. We label a frame as noise if $\mathrm{MDL}(0) > \mathrm{MDL}(\mathrm{i}), \forall i \neq 0$.

This analysis states that in order to detect whether various sinusoidal components of a signal are present or not, we need to maximize a quantity that includes squares of signal amplitude and frequency estimates. From [30], the expected value of the energy operator on a filtered AM–FM signal-plus-noise can be approximated by $\Psi(X[n]) \approx A[n]^2|H(\Omega_c)|^2[(\Omega_c + \partial\Phi[n]/\partial n)^2 + \Gamma_c]$, where Ω_c and H are the filter's central frequency and frequency response, respectively, and Γ_c a constant standing for the averaged filtered noise power. In our case, this approximation yields $\Psi(X[n]) \approx A^2(\Omega_c^2 + \Gamma_c) \cdot |H(\Omega_c)|^2$ and by taking logarithms we have

$$\ln\Psi\left(X[n]\right) \approx \ln A^2 + \ln\left(\Omega_c^2 + \Gamma_c\right) + \mathrm{const.} \quad (20)$$

Comparing (18) and (20), we notice the amplitude–frequency product components and the constants depending on the average bandpass noise inside the logarithms. These similarities despite the simplifications posed on the problem, give an insight on the role of the energy operator and the DESA estimates on a channel decision-speech detection process. We interpret our data either in terms of the sinusoid of estimated amplitude $\hat{A}$ and carrier frequency Ω_c that maximizes (15) or as a background noise process. In the latter case, a quantity similar to a maximum log-Teager energy is below some estimated threshold (19). Either the instant amplitude and frequency estimates through DESA (5), (6) or the Teager Energy estimation (20) can serve as speech energy tracking and detection features.

Thus, motivations for developing features for speech detection that involve Teager energy, ESA, and MDA include 1) the optimality of a threshold detector that simultaneously compares mean amplitude and frequency estimates, 2) the dual amplitude–frequency variation information captured by the Teager–Kaiser energy operator, 3) the ability of ESA to reveal and estimate modulations in small time scales, 4) the expected noise reduction due to multiband demodulation [30].

B. Modulation-Based Analysis

We propose three new time-domain signal representations as alternatives to the common mean square amplitude (mSA), mean absolute amplitude (mAA), and average zero-crossings rate (ZR). These multiband features are the maximum average Teager Energy (MTE), the mean instant amplitude (MIA), and mean instant frequency (MIF) through a max-energy output filter selection and demodulation scheme.[2] The block diagram

[2]Analogous to the common ones, the new measurements are also based on an underlying averaging-lowpassing procedure per signal frame. For simplicity though, we exclude the term "mean" from the abbreviations. "M" will stand for multiband.

in Fig. 1 shows how these new representations are derived. The signal is processed in small analysis frames varying between 10–25 ms. For speech signals, a choice of window length in that range is aimed at covering all pitch duration diversities between different speakers. Every frame is filtered through a bank of K Gabor bandpass filters with linearly spaced center frequencies. Frequency steps between adjacent filters may vary between 200–400 Hz, yielding filterbanks consisting of 20–40 filters. The discrete Teager–Kaiser energy operator is then applied to the output of every filter, and the average frame Teager energy is computed.

For each frame m, the maximum average energy is tracked on the multidimensional filter response feature space. The filter considered most active in this energy sense is submitted to demodulation via the DESA. The instant demodulated amplitude and frequency derived from the energy separation algorithm are also averaged over the frame duration

$$i = \arg\max_{1\le k\le K} \left(\overline{\Psi_d\left[(s * h_k)(n)\right]}\right)$$

$$\mathrm{MTE}(m) = \left(\overline{\Psi_d\left[(s * h_i)(n)\right]}\right) \tag{21}$$

$$\mathrm{MIA}(m) = \left(\overline{|A_i(n)|}\right) \tag{22a}$$

$$\mathrm{MIF}(m) = \left(\overline{\Omega_i(n)}\right) \tag{22b}$$

where $*$ denotes convolution and h_k the impulse response of the k_{th} filter. Averaging takes place over N samples of frame m and n is the sample index with $(m-1)N+1 \le n \le mN$. Each frame yields average measurements for the Teager energy, instant amplitude and frequency based on the filter that captures the strongest modulation component of the signal. The maximum average Teager energy may be thought of as the dominant signal *modulation energy*.

The classic and new short-time speech analysis features are depicted in Fig. 2, for the word /idiosyncrasies/ in low-level background noise. All plots are normalized with respect to the ratio of maximum signal value to maximum feature value and post-smoothed. The new features were derived through a linear filterbank of 25 Gabor filters with 160-Hz rms effective bandwidth. To obtain smooth frequency representations, the demodulated instant frequency from the dominant channel is smoothed, prior to averaging, using a 13-point median filter, in order to reduce effects of demodulation errors.

Regarding the captured signal information, especially for tasks of speech versus silence discrimination:

1) Both the classic mAA (or mSA) and the new MTE and MIA, provide information about signal envelope variations. The new MTE assigns greater values to fricatives, stops, and nasals compared to silence than the common mSA.
2) The classic ZR and the proposed MIF relate to the signal's spectral content. MIF has a block-like variance with time, due to the max-select procedure, an attribute that may be used for speech labeling and discrimination. Furthermore it attributes to some speech sounds, like nasals or voiced fricatives, high frequency components, and in some cases decreases the level distance between speech and silence.

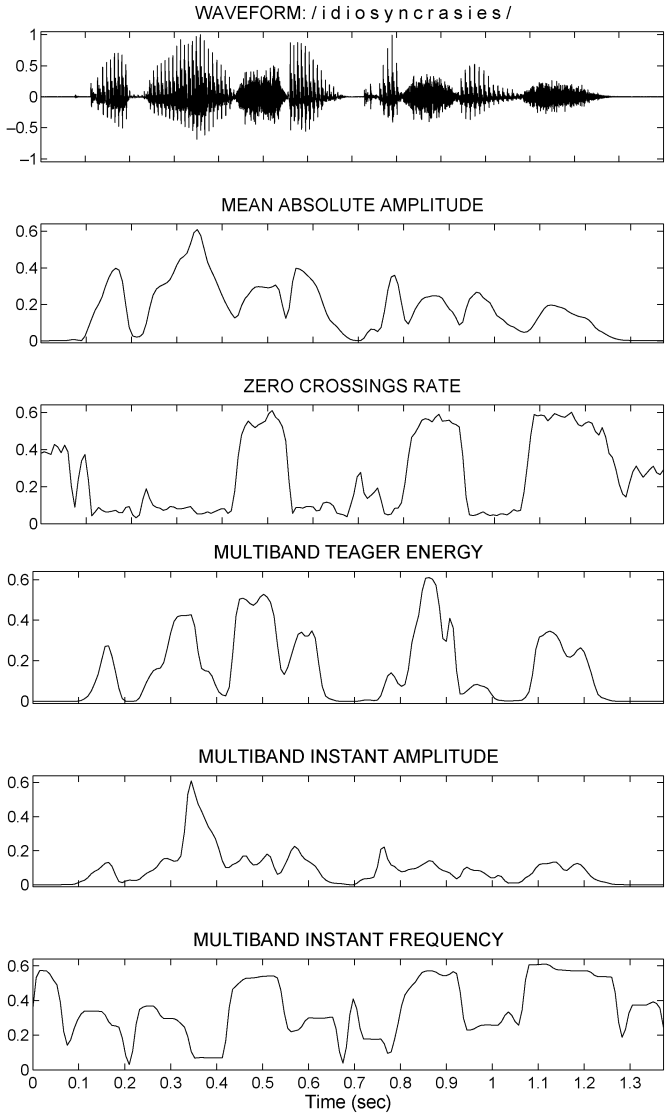


Fig. 2. Short-time features for signal analysis using 15-ms Hamming frames, updated every 1/2 of frame duration, at a 16-kHz sampling rate. Signal waveform, classic features (mAA and ZR) and the proposed new features of multiband Teager energy (MTE), instant amplitude (MIA) and frequency (MIF) are depicted. Plots are all normalized and smoothed by combined linear-nonlinear post-filtering (three-point median, three-point Hanning).

3) The MTE energy can be considered as an alternative energy measurement, that even as a stand-alone feature can easily indicate existence of speech, due to its joint amplitude–frequency information. However, because it is based on differential operators it is very sensitive to abrupt noise artifacts and nonspeech events, e.g., pops, mouth clicks.

C. Endpoint Detection With Modulation Features

We based our endpoint detection approach on ideas similar to the classic one, only with different signal processing methods and measurements, aiming to improve isolated word detection and explore the effectiveness of the developed features for speech analysis. We used the MTE measurement instead of mAA for a first voiced region approximation and MIF instead of ZR for a refinement of the initial estimates. A modified Teager energy measurement was also used in [25], as a feature for

endpoint detection, however, not in a multiband demodulation framework. Multiband processes have been applied to voice detection in [16] and word boundary detection in [14] as useful band selection mechanisms.

The MTE and MIF representations are computed for the whole signal. From the first 100 ms, which are assumed to contain no speech and *a priori* labeled as silence, we estimate the mean μ_{sif} and standard deviation σ_{sif} of the "silent" instant frequency measurement. The maximum MTE values for silence, $W_{\max}$, and for the whole signal, $S_{\max}$, are also computed. Finally, three thresholds, one for MIF and two for MTE, are estimated according to the rules

$$\gamma_{\text{f}} = \mu_{\text{sif}} + \kappa\sigma_{\text{sif}} \quad \gamma_{\text{d}} = \min(T_1, T_2) \quad \gamma_{\text{u}} = 5 \cdot \gamma_{\text{d}} \tag{23}$$

$$T_1 = \lambda S_{\max} + (1 - \lambda) W_{\max} \quad T_2 = 3 \cdot W_{\max} \tag{24}$$

where κ, λ are weighting constants. The energy thresholds are in essence decided by comparing the ratio $S_{\max}/W_{\max}$ to a fixed constant.

The double energy-threshold check, searching for the point where the higher one γ_{u} is exceeded and then moving backward (in search for the beginning) and forward (if searching for the end of the word) until the lower-stricter γ_{d} threshold is reached, detects the main duration of the speech signal. Because of the frequency content inherent in the Teager energy, this energy check alone may often be adequate for accurate endpoint detection. In order to increase accuracy, a refinement check is made using the γ_{f} frequency threshold. Weak unvoiced fricatives, stops, or nasals are searched for in the previous (in the beginning) or following (in the end) 250 ms. If the threshold is exceeded a fixed number of times, depending on the frame refresh rate, a starting or ending point for the word is chosen at the first crossing instance. In our tests, we used $\lambda = 0.02$, $\kappa = 1$. The frequency threshold cannot be made stricter without subsequent increase in the rate of false alarms.

Apart from the prominent pair MTE and MIF, which gave the best word detection results, we tested various combinations of the classic (mSA, mAA, ZR) and the new modulation (MTE, MIA, MIF) short-time tools. The efficacy of the new features depends equally on the advantages of both multiband analysis and modulation energy features. To demonstrate that we tested two other forms of Teager energy analysis (STE, PTE) for their detection ability and a multiband version of the classic measurements (MAA, MZR).

IV. Experimental Results

The results in this section refer to endpoint detection of isolated words or short phrases, performance evaluation and robustness in noise. Generally, there are three categories of errors encountered in detecting speech boundaries, which are: 1) lost phonemes in the beginning or end of a word; 2) overestimation of the "silent" period prior or after a word, known as "spoke too soon" errors; and 3) misclassification of nonspeech, acoustic events (clicks, pops, breathy speech) as speech endpoints. Here, we are primarily concerned with the first type, that leads to cropped or entirely lost phonemes, like stops, unvoiced fricatives, and whispers of low energy but high spectral components. The other two categories are jointly dealt with, as the presence of an unknown acoustic event may often lead to boundary overestimation. Note that below 30-dB human noises that lead to errors of the third type are obscured by the additive noise and thus do not deteriorate the detection process.

A. Test Databases

To evaluate detection performance and the advantages of the new short-time features for speech-signal analysis, we used three datasets of utterances from the English DARPA-TIMIT, NYNEX-NTIMIT and Spanish SDC-Aurora database. From Aurora, due to the lack of labeled data, we used a fairly small number of isolated digits, chosen randomly on three different noise conditions, labeled manually through auditory and visual inspection (50 utterances). These serve for indications, qualitative evaluation, and examples only, but such results will nevertheless be mentioned. SDC-Aurora consists of digits, isolated or sequences, in three noise conditions (quiet, low, and high road-traffic noise) [37].

For the actual quantitative evaluation of our methods, we used the TIMIT and NTIMIT databases. The manually labeled data from these databases consist of phrases, in different dialects and various male and female speakers. The task was to detect the phrase boundaries and compare results with the labels specified in documentations. For that we used the first and last word of each phrase and treated them as a single utterance, ignoring activity in between.[3] Our demand was that at least 100 ms at the beginning of the phrases are silence and that no extreme or long-duration nonsilence nonspeech noise artifacts existed.[4] This candidate scan to the TIMIT test set led to 1570 (out of a 1680 total) utterances which formed our dataset. Detection was evaluated under artificial noise added at five different SNRs. For real noise conditions testing, we used the NTIMIT database, a telephone bandwidth adjunct to TIMIT, developed by transmitting TIMIT through a telephone network.

On some rough databases statistics, the beginning or ending phonemes were on the average: 19% vocals, 21.1% unvoiced stops, 13.3% voiced stops, 1% voiced fricatives, 22.2% unvoiced fricatives, 5.4% nasals, and 18% semivowels. We will refer to the different sets as Aurora, TIMIT, and NTIMIT, corresponding to the previously described databases.

B. Indications—Examples

In Table I, we present experimental results that indicate improvement over the classic method by use of the combination of Teager energy (MTE) and demodulated instant frequency (MIF) features. The tests were performed on the Aurora set and results refer to percent of correctly detected endpoints. An error in detection was considered when a first or last phoneme of a word was cropped more than 30 ms or when the boundaries were overestimated for more than 60 ms.[5] The percentages refer to the total number of endpoints (beginning and end). We used

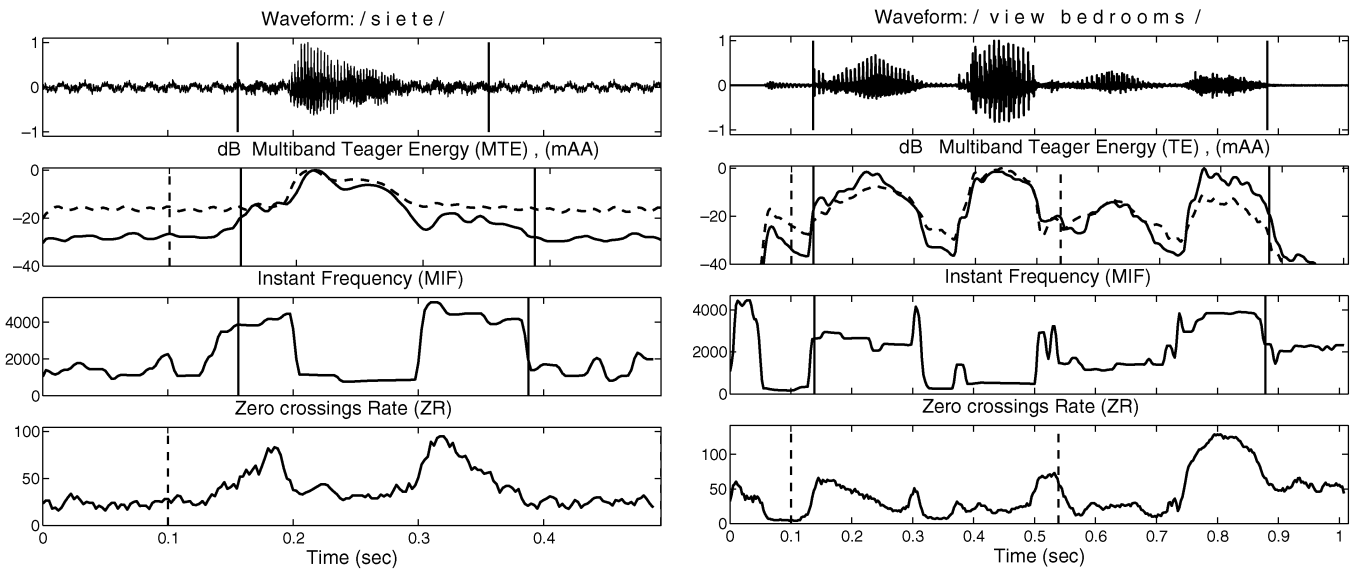
[3]This is actually what the classic endpoint algorithm performs by finding two high energy-level instances and considering everything in-between as speech.

[4]These artifacts usually affect detection under large SNRs.

[5]For frames of 15 ms, with 50% overlap between adjacent frames, 30 and 60 ms are four and eight signal frames, respectively.

TABLE I
PERCENTAGE (%) OF DETECTED SPEECH ENDPOINTS WITH VARIOUS FEATURES ON AURORA SAMPLES

Classic		Multiband Modulation					Teager	
mAA, ZR	mSA, ZR	**MTE, MIF**	MTE, ZR	MTE	MIA, MIF	MIA×MIF	STE	PTE
84.2	77.6	86.8	84.2	77.6	75.0	78.9	81.6	80.3



(a) (b)

Fig. 3. Endpoint detection using classic and proposed modulation features in utterances. (a) /siete/, from Spanish SDC-Aurora. (b) /view-bedrooms/, from TIMIT phrase (on a 30-dB SNR level). Top figures are the signal waveforms with vertical lines marking the actual endpoints. Following are the MTE (continuous curve) and the classic mAA (dashed curve) superimposed, on a decibel scale. The proposed MIF and the conventional ZR are presented below (after median filtering). All were derived using windows of 15 ms updated every 5 ms. The markers for the endpoints detected with the classic features are dashed, while for the proposed are solid vertical segments.

various combinations of conventional and proposed features for the most efficient choices in terms of detection. Column MTE refers to detection using only the new energy feature, while MIA $\times$ MIF is an energy detector based on the two feature product. The last two columns of single-feature detection refer to Teager energy alternatives that will be later discussed.

All short-time measurements were made in 15-ms Hamming windows, updated every 2.5 ms, in 16-kHz sampling frequency using 16-bit signal representations. The new representations were derived by choosing the maximum mean Teager energy from a filterbank of 25 Gabor filters of 160 Hz. Both ZR and MIF were post-smoothed using median filtering. The number of times that ZR or MIF threshold (23) was exceeded, thus signalling unvoiced speech, was empirically set to N/s, where N the window length and s the shift in samples. Apart from results in Table I referring to shifts of 5 ms $(\mathrm{s} = 80)$, the algorithms were also tested for shifts of 2.5 ms $(s = 40)$ and 7.5 ms $(s = 120)$. Smaller shifts generally behaved better toward improvement of detection.

On the noise-free TIMIT set, the results were in favor of the classic features, in terms of detection error previously defined, with the classic algorithm achieving 76.3% and the algorithm with MTE detecting 72.2% of the endpoints. Refinement with the MIF feature did not improve overall performance. This is probably due to the nature of the TIMIT data (i.e., manual labeling), noise artifacts on silent intervals (to whom the Teager operator, and any differential operator, is sensitive) which are obscured at lower SNRs and our measure of performance (by empirical error thresholds).

Following, in Fig. 3 we show two examples in which the proposed method succeeded in capturing phonemes that the classic algorithm failed to include in the speech region. The utterance in (a) is the word /siete/ from the Aurora set at quiet noise conditions (stopped car-motor running). The beginning /s/ and final /te/ are embedded in the signal of the running engine. As a result, the low-level mAA completely misses the final endpoint, while the proposed method accurately marks both and even improves the manual labeling of the ending instance. In (b), the results on the utterance consisting of the words /view-bedrooms/ from the TIMIT set. Here the classic features signify the ending point much too early, missing the largest part of the second word. Note that the final /s/ on the MTE is of the same or even higher level as the vowels of the utterance. Both endpoints are marked in accordance with manual labeling.

Generally, improved detection was observed mostly in cases of weak stops, both voiced and unvoiced, weak, low-energy, unvoiced fricatives and parts of unvoiced-turned phonemes (e.g., nasals). Also, in cases of long stops in the middle of words [e.g., the /d/ in Fig. 3(b)], the classic algorithm failed to capture the whole syllable after the stop. The power of the new features though is the improved robustness in noise.

C. Detection in Noise

To test robustness of both algorithms under noisy conditions, we used the dataset of TIMIT utterances, with randomly gener-

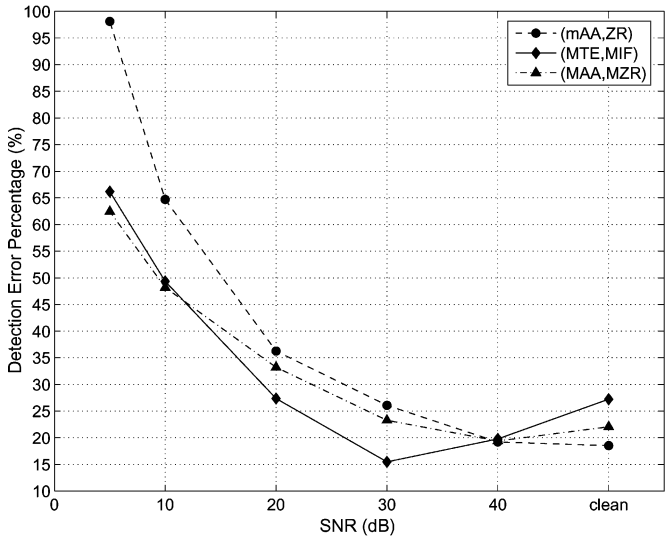


Fig. 4. Error in endpoint detection (%) under different SNR levels (decibels) for the classic and the proposed features on the TIMIT dataset. Solid line is for the modulation features, dashed line for the classic features, and dash-dotted for multiband versions of amplitude and zero-crossings rate.

ated additive white noise. An amplitude parameter, according to the utterance rms value, was set to provide five SNR levels 40, 30, 20, 10, and 5 dB. As before, we used two critical values to quantify error in detection (60 ms for a "spoke too soon" error, 60 ms for a "lost phoneme" duration error), with the actual endpoints dictated by the TIMIT files. Fig. 4 shows the deterioration of both algorithms as the noise level increases by means of absolute % error in detection. Results, averaged over the whole set, refer to 5-ms shifts of 15-ms frames, retaining the filterbanks and sampling frequency configurations.

To separately validate the effect of multiband analysis and the use of modulation Teager operator-based measurements, we also implemented multiband versions of the classic features. Specifically, using the same frequency bands as the new algorithm, we choose for the amplitude and zero-crossings rate the channel that maximizes the absolute amplitude measurement, analogous to (21) and (22). This leads to multiband max average amplitude (MAA) and zero-crossings rate (MZR). This MZR is the rate of crossings at the band with maximum mean filtered envelope. Notice from Fig. 4 that these features performed transiently between the two. Better than the proposed on the clean set but worse than the classic one, equally at 40 and 10 dB, almost the same at -5, where multiband analysis is most beneficial and worst between 40–10 dB, where improvement is due to the use of modulation energy features.

As can be seen in Fig. 4, the proposed method is more robust in detection at noise levels below 30 dB with similar performance at 40 dB. The relative decrease in detection error is $\approx$41% (30 dB), 24.6% (20 dB), 23.7% (10 dB), and 32.5% at high noise level (5 dB), where the classic algorithm is extremely unreliable. While the classic algorithm is responsible for the most "lost phonemes" errors, the new one gives more "spoke too soon" false indications. This may be due to the frequency component included to the Teager energy measurement. Note that these results may vary by changing the detection error definitions depending on task expectations for endpoint accuracy.

In order to evaluate results independently of the empirical thresholds that define error in detection, we produced curves in the philosophy of the ROCs [35], [38]. An ROC curve is a detector evaluation measure depicting the change of probability of detection (PD) with the increase of probability of false alarm (PF). For the fixed decision thresholds of (23), we use a simple convention to produce curves that approximate the ROCs, by varying the error interval. We set a rather tight "lost-phoneme" error threshold at 30 ms and let the "spoke too soon" tolerance vary from 2 to 150 ms. Any endpoint falling between these error limits is considered correctly detected. The final detection percentage over all data defines the PD, which is plotted against the estimation tolerance interval. This interval, normalized to its maximum, yields a measure related to the probability of false alarm. The two quantities are connected by a one-two-one, monotonically increasing, unknown function, as increase in the error interval increases the PF by *some* amount.[6] These curves are not always convex or above the diagonal, like the actual ROCs, but they serve as an evaluation measure of the detection process.

In Fig. 5, we present such *detection-tolerance* curves for the TIMIT set, on the noise-free case and the five additive noise levels for the classic, multiband classic, and proposed algorithms. Above 20 dB, the detection probabilities are increasing for all three methods. This translates to most errors belonging to the "spoke-too-soon" category. The estimated endpoints that fall within the detection interval increase with the tolerance in "spoke-too-soon" error. In contrast for lower SNRs, the piecewise flat curves reveal that the errors are mainly due to lost phoneme durations, where the error threshold is fixed. A performance around 0.6 at 20 dB, or even lower for increasing noise may be meaningful only by relaxing the strict constraints for lost-phoneme duration more than four frames (30 ms), resulting in overall detection improvement. Comparatively, detection below 30 dB with the modulation features is superior compared to the classic one. At 40 dB, there is a transition in performance around the tolerance interval of 50 ms. Note that the best performance for the classic algorithm was achieved at noise-free conditions $(\infty$ dB$)$, whereas for the new algorithm at 30 dB, conditions closer to real-life practical applications. The area under these curves, expressing average detector performance in ROC analysis, is increased with the modulation features by 19% at 30 dB over both versions of the classic features.

The same testing and comparisons were performed on the NTIMIT noisy-telephone speech dataset. Results for the whole set, again for various approaches, can be seen in Table II. Decrease in the average detection error was 38.7% with the sole use of the MTE feature and 40% after refinement by the MIF. In Fig. 6, the detection-tolerance curves are illustrated, where again the improved robustness of the multiband modulation approach is highlighted, against the classic and the multiband-classic features, under realistic noise conditions. Notice how the

[6]PF for speech is complementary to PD for nonspeech events. By the same logic of error intervals for nonspeech detection, this PD decreases by increasing the tolerance interval for speech detection.

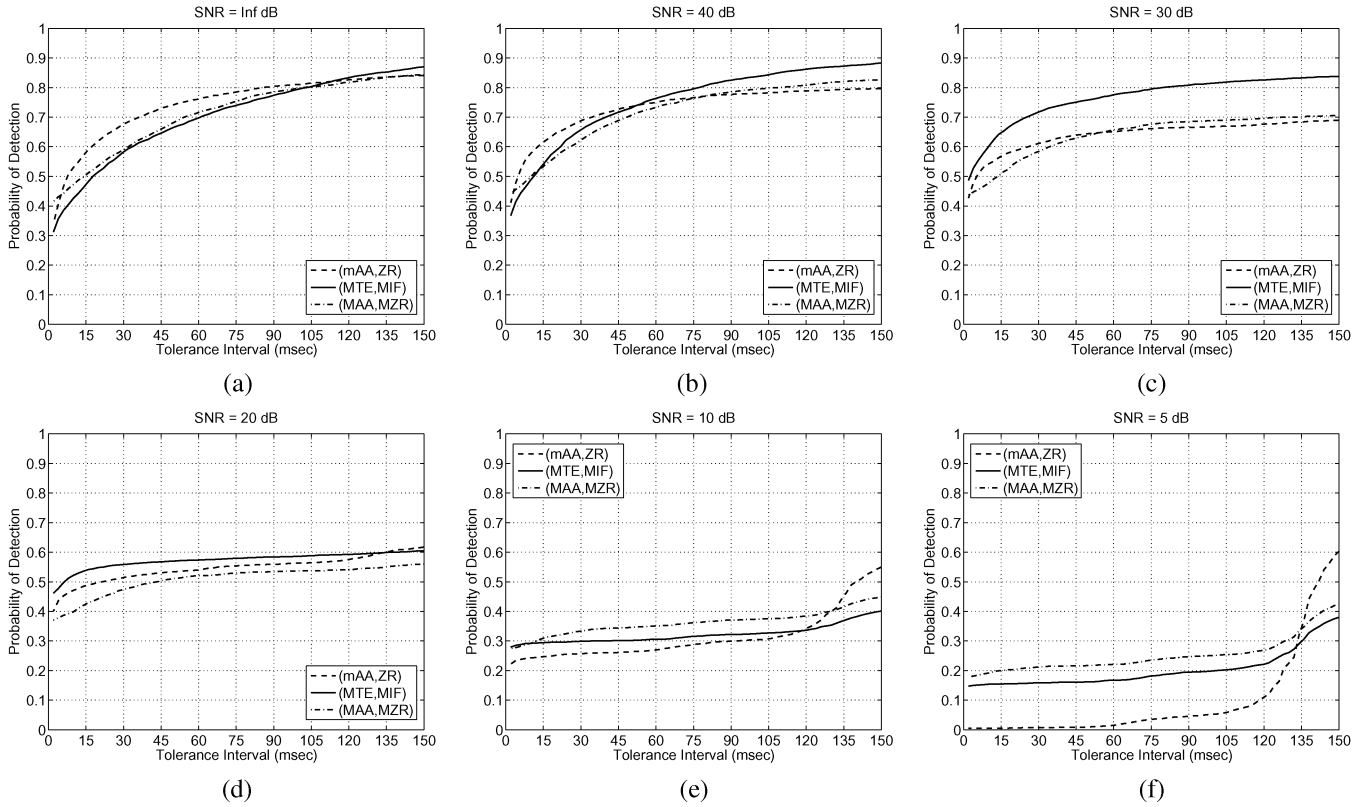


Fig. 5. Detection-tolerance curves, in the spirit of ROC, for the TIMIT dataset in various additive noise levels. (a) ∞ dB (clean data). (b) 40 dB. (c) 30 dB. (d) 20 dB. (e) 10 dB. (f) 5 dB. Dashed line corresponds to results from the classic algorithm, dashed–dotted line to its multiband version, and solid line to the proposed one. The time interval is the tolerance in endpoint estimation and is related to probability of false alarm.

TABLE II
PERCENTAGE (%) OF DETECTED SPEECH ENDPOINTS FOR VARIOUS APPROACHES ON NTIMIT

Classic		Multiband	Multiband Modulation		Teager	
mAA, ZR	mSA, ZR	MAA, MZR	**MTE, MIF**	MTE	STE	PTE
56.1	66.6	51.5	73.5	73.1	71.6	49.5

multiband-classic features give worse results compared to their full-band counterparts, something that was also noted for SNRs of 20–30 dB at the TIMIT tests.

The aforementioned experiments demonstrate the improved noise robustness of the modulation features and the proposed algorithms. However, a stronger criterion on detection accuracy would be incorporation of the detected words in a speech recognition task. Results presented here were based on the thresholds that were set to define accuracy in detection. For task-independent comparisons, results are repeated in Table III in the form of absolute deviations in milliseconds from the true manually defined TIMIT word boundaries, for various SNRs and 15-ms window length, updated every 2.5 and 7.5 ms. For both sets, the absolute difference, averaged over all data, is smaller for the proposed detector under 40 dB.

D. Related Teager Operator Work

A modified Teager energy has been previously used for endpoint detection [25]. This feature called *frame-based Teager energy* computed as the sum of the squared frequency-weighted power spectrum per analysis frame (PTE) performed better than simple rms energy. Using only one feature for detection, we tested our MTE against PTE and a short-time version of Teager's energy, the average of Ψ per frame (STE). The percentages under STE, PTE in Tables I and II refer to these

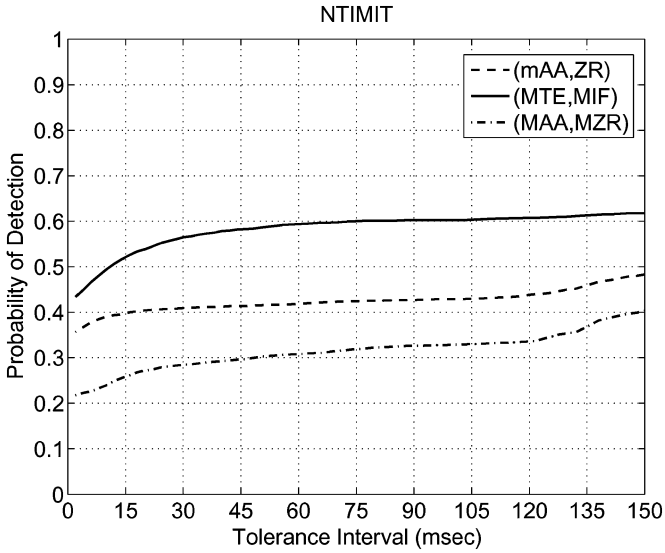


Fig. 6. Detection-tolerance curves for the NTIMIT dataset. Dashed and dash-dotted lines corresponds to results from the classic and multiband classic features and solid line from the proposed modulation ones. The time interval is the tolerance in endpoint estimation and is related to probability of false alarm.

TABLE III
AVERAGE DEVIATIONS FROM "TRUE" TIMIT ENDPOINTS

Shift (N=window length)	Method	Absolute Deviation in milliseconds						
		TIMIT						NTIMIT
		∞ dB	40 dB	30 dB	20 dB	10 dB	5 dB	
$N/6$	mAA, ZR	36.1	37.1	53.3	78	134.1	178.7	94.8
	MTE, MIF	53.3	39.4	30.5	55.5	106.9	137.7	58.5
$N/2$	mAA, ZR	37.4	38.1	53.1	77.2	135.5	180.1	92
	MTE, MIF	59.1	43.7	33.7	54.6	107.5	138.3	60.2

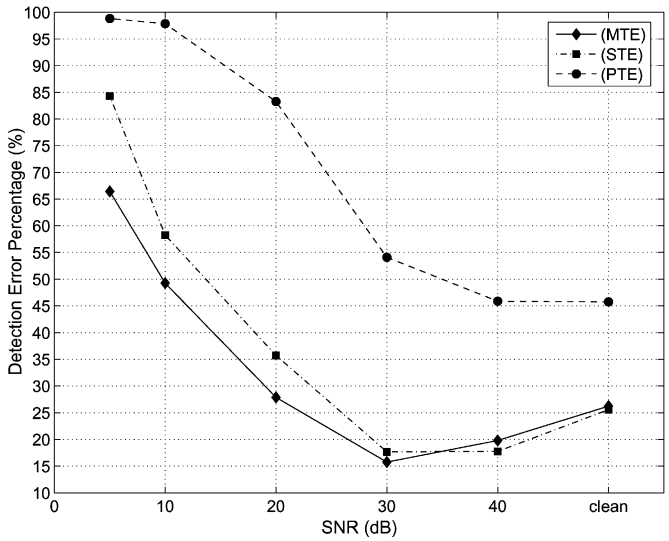


Fig. 7. Endpoint detection error (%) on the TIMIT dataset under various SNR levels (decibels), using three Teager energy features: The proposed multiband max average Teager energy (MTE), short-time Teager (STE) , and the frame-based Teager (PTE).

measurements. For the TIMIT dataset, with and without additive noise, the results are shown in Fig. 7. MTE is more robust in noise as a consequence of the multiband filtering process, with an average 10.1% error decrease over STE and 29.6% over PTE. In comparing STE either with the classic combination in Table II, or its multiband alternative, we conclude that improved performance of the proposed MTE does not stem only from multiband analysis but also from the explicit use of the modulation energy representation.

V. VOICE ACTIVITY DETECTION (VAD)

The problem of speech detection is formally known in telecom systems as VAD and is an essential part of most modern speech recognition and transmission front-ends. Any VAD performs a rough classification of incoming signal frames based on feature estimation in two classes: speech activity and nonspeech events (pauses, silence, or background noise). The output of such a process is a logical flag pointing at the speech-classified signal frames.

A. Feature-Based Detector

A recently developed and highly accurate VAD system, based on short-time features, was proposed in [3] and [26] for noise reduction and improvement of speech recognition by enhancement and frame dropping. The algorithm is based on adaptive thresholds and noise parameter updating and decides on speech/nonspeech activity by estimating a feature termed *long-term spectral divergence* (LTSD). For each frame m under consideration, the LTSD is defined as

$$\mathrm{LTSD}(m) = 10\log_{10}\left(\frac{1}{N_F}\sum_{k=0}^{N_F-1}\frac{\mathrm{LTSE}^2(k,m)}{|W(k)|^2}\right) \quad (25)$$

where LTSE is the long-term spectral envelope, the maximum spectrum amplitude in a neighborhood of frames, at each discrete frequency $k = 0, \ldots, N_F - 1$, and $|W(k)|^2$ is the average noise spectrum magnitude. The LTSD quantifies the divergence of speech from background noise/silence and is in essence an energy measure that retains spectral information by keeping the strong spectral components on neighboring frames.

The algorithm in [3] required estimation of the bounds for the adaptive voice-triggering threshold and included updating of the average noise spectrum measurement every time a pause was detected, for adaptation to nonstationary noise environments and a controlled hang-over scheme to avoid abrupt state transitions. The VAD algorithm and LTSD were extensively tested in large databases and varying noise conditions against standard VADs such as the ITU-G.729 or ETSI-AMR [7], [8]. The LTSD-based VAD performance was evaluated using both common VAD evaluation methods and recognition accuracy.

B. Modulation Energy Detector

To evaluate the modulation-based features, and especially MTE as efficient speech/nonspeech discriminators in a VAD system, we chose the LTSD-based VAD for: 1) its improved performance and slow degradation with SNR; 2) the favorable comparisons in [3] against standardized detectors; and 3) the extensive experimental evaluation on appropriate databases. We adapted the aforementioned algorithm by changing the core feature with the proposed modulation MTE feature in two alternative expressions.

The signal is frame-processed and during a short initialization period the initial noise characteristics are learned. After feature computation, the level difference in decibels from the respective background noise feature is compared to an adaptive threshold $\gamma \in [\gamma_0, \gamma_1]$

$$\gamma = \gamma_0 + (\gamma_1 - \gamma_0)\frac{(E - E_0)}{(E_1 - E_0)} \quad (26)$$

where E the background noise energy. The threshold interval boundaries depend on the cleanest E_0 and noisiest E_1 energies, computed during the initialization period from the database under consideration. The noise feature is estimated during

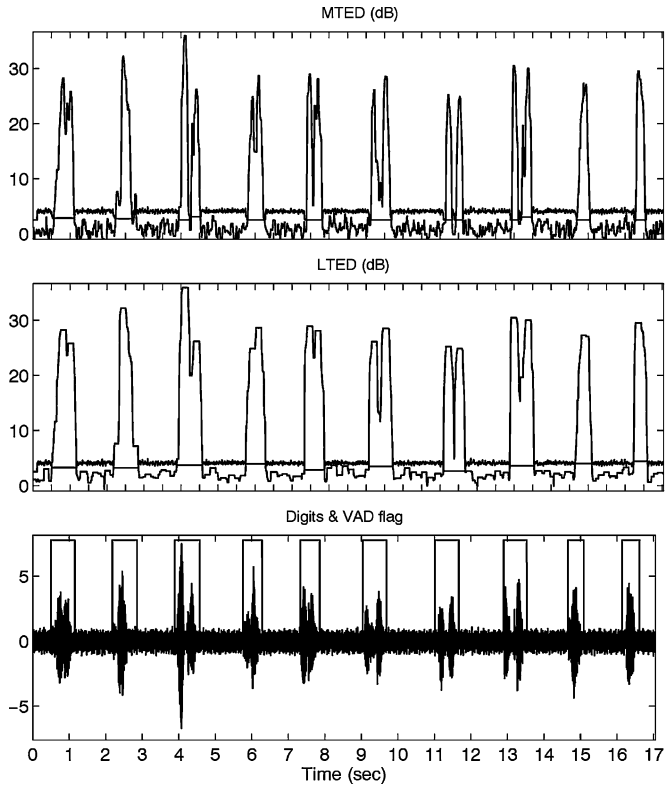


Fig. 8. Features based on maximum average Teager energy (MTE) for voice activity detection (digit sequence in 12-dB SNR). Both were derived for frames of 25 ms with 10-ms shifts and a bank of 25 Gabor filters. Top figure is the MTE divergence and middle the long-term MTE divergence. The VAD flag in the waveform was identically derived by both MTED and LTED features.

the initialization period and adapted whenever silence or pause is detected, by averaging in a small neighborhood of frames.

To measure modulation "divergences" in the spirit of LTSD for VAD, we use features based on MTE.

1) Multiband Teager Energy Divergence (MTED): The multiband max average Teager energy MTE as used previously, compared to the respective feature MTEW for background noise

$$\mathrm{MTED}(m) = 10\log_{10}\left(\frac{\mathrm{MTE}(m)}{\mathrm{MTEW}}\right). \tag{27}$$

The MTED is measuring the divergence between the multiband MTE of a frame and the corresponding noise feature. This is conceptually the same as the endpoint detection algorithm of Section III-C, comparing MTE level difference.

2) Long-Term Multiband Teager Energy Divergence (LTED): The MTE is locally maximized in a neighborhood of $2L$ frames resulting in a dilated and normalized, with respect to the background noise, version

$$\mathrm{LTED}(m) = 10\log_{10}\left(\frac{\max\limits_{l}\{\mathrm{MTE}(m+l)\}}{\mathrm{MTEW}}\right) \tag{28}$$

with $-L \leq l \leq L$ defining the order of the dependence.

In Fig. 8, the proposed VAD features are presented on a digit sequence by Aurora 3 in quiet (12-dB), hands-free microphone, recording conditions [37]. In each plot, superimposed is the adaptive threshold signalling voice activity.

C. Experimental Evaluation

The experimental framework for comparing performance of the LTSD-based and the MTE-based VADs, consists of a large number of detection experiments, under varying noise conditions on the Aurora 2 and Aurora 3 databases. The Aurora 2 [39] consists of connected digits under real-life noises of various characteristics, added to the clean TIdigits at SNRs 20, 15, 10, 5, 0, and -5 dB, reaching a total of 70 070 utterances. The Aurora 3 Spanish database [37] consists of 4914 utterances of the SDC digit sequences, in three noise conditions (quiet, low noisy, high noisy) corresponding roughly to average SNR values of 12, 9, and 5 dB and two recording conditions (close-talking and hands-free recordings).

Evaluation is based on classification errors at different SNRs [2], [3], [8] using some reference labeling of the clean digit set. In our experiments, automatic speech recognition experiments were used to segment and label the speech/nonspeech events on the databases. High recognition rates on the clean sets defined the ground truth for the digit sequences. Briefly, for Aurora 2, the training was done using 32 mixtures, 18 states. and the 39-long feature vector $[\mathrm{MFCC}, \log \mathrm{E}, \Delta, \Delta\Delta]^{\mathrm{T}}$ on the clean-train scenario. The test run on the clean data achieved a word accuracy of 99.6%. For Aurora 3, the training was done with 16 mixtures, 16 states, and the same feature vector. A 1522 subset of the 4914 utterances was used as the test set (well matched test) with a word recognition accuracy of 93.7%. Utterances with erroneously recognized digits were removed from the reference labeling to improve ground truth accuracy.

For the reference LTSD-based VAD, we used the specifications and the values reported in [3] about the decision thresholds (26) ($\gamma_0 = 6$ dB, $\gamma_1 = 2.5$ dB), with the background noise energies E_0, E_1 estimated by the averages of the first 100 ms on all utterances, in the cleanest and noisiest conditions, respectively. The hang-over mechanism was set to four frames. For the proposed VADs, we used roughly similar specifications but determined the optimum thresholds by means of ROC curves analysis [38]. In Fig. 9, these curves are presented in the cleanest and noisiest sets for the MTED and LTED-based VADs. We chose the thresholds that correspond to the points of the curves with minimum distance from the upper left, ideal working point, corner. This analysis led to $\gamma_0 = 24$ dB, $\gamma_1 = 0.5$ dB for the MTED-based VAD and $\gamma_0 = 32$ dB, $\gamma_1 = 2$ dB for the LTED-based one on the Aurora 2 set. The tests on Aurora 3 were conducted for all three features with the same pair of thresholds for reference.

The recognition-labeled speech, pause and silence durations were used to define the actual speech and nonspeech intervals. Performance of the VADs was evaluated with respect to the speech hit rate HR1, defined as the ratio of the detected speech frames to the total number of speech frames and the nonspeech hit rate HR0, defined, respectively, as the detected percentage of the nonspeech frames in the dataset. Complementary to these quantities, $\mathrm{FAR1} = 100 - \mathrm{HR0}$ and $\mathrm{FAR0} = 100 - \mathrm{HR1}$, are the false alarm rates (FARs) of the decision for speech or nonspeech.

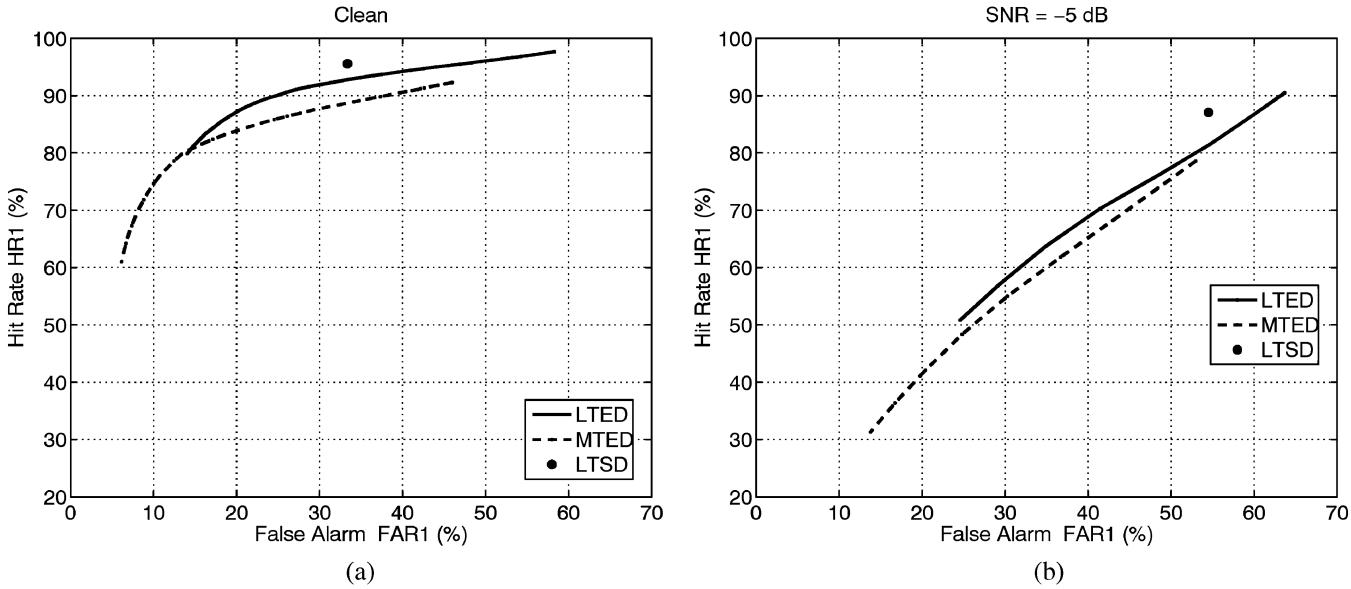


Fig. 9. ROC curves for speech detection performance in (a) clean and (b) noisiest (-5 dB) Aurora 2 set for the MTE-based VADs. Decision thresholds varied as $4 \leq \gamma \leq 38$ in the clean and $-1 \leq \gamma \leq 3$ in the noise case. The operating point of the LTSD-based VAD is also depicted.

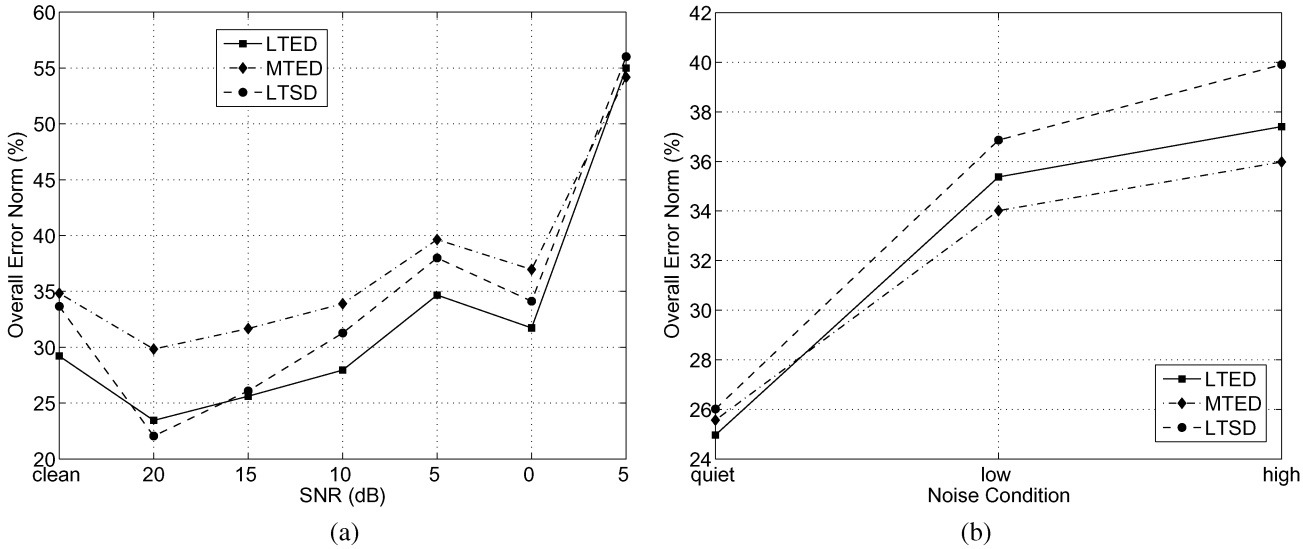


Fig. 10. Overall false alarm error for speech/nonspeech detection in various SNR levels on (a) Aurora 2 and (b) Aurora 3.

In Tables IV and V, we present detection performance results for the various VADs on the reference datasets. Results are presented in terms of both hit rates per noise level and averages over all data. The rates HR1 and HR0 are considered of equal importance since misclassification errors take place both in speech and silent signal periods. This may be quantified through the L_2 norm of false alarm rates. In effect, we aim to minimize the *overall false alarm error norm*:

$$\mathrm{E_{FAR}} = \|(\mathrm{FAR0}, \mathrm{FAR1})\| = \left[(1-\mathrm{HR1})^2 + (1-\mathrm{HR0})^2\right]^{\frac{1}{2}}. \tag{29}$$

Statistically, this measure expresses the average performance of the detector as $\sqrt{2}$ times the rms error norm of the false alarms, while geometrically it is the shortest Euclidian distance from the ideal operating point (upper left corner) on the ROC plot of a detector ($\mathrm{HR} = 100$, $\mathrm{FAR} = 0$) (see also Fig. 9). In Fig. 10, that error norm is presented for the two datasets and the three VADs as a function of decreasing SNR.

The LTSD-based detector, as proposed in [3], is quite conservative with respect to the actual silence percentages that are being detected, with high speech hit rates in return. On the Aurora 2 tests, where the thresholds were optimally set, the MTE-based algorithms equally weight both percentages giving average hit rates above 70% on both speech and silence periods. The LTED achieved the minimum false alarm error norm, with a 7.6% decrease of the overall error over the LTSD-based VAD. In Fig. 10(a), the LTED detector minimizes the error, except on 20-dB SNR, where all three features follow analogous degradations in performance. On the Aurora 3 results in Table V, where the detection thresholds were the same, the LTED achieves

TABLE IV
DETECTION ACCURACY, AURORA 2

Noise (dB)	VAD feature					
	LTSD		LTED		MTED	
	HR1	HR0	HR1	HR0	HR1	HR0
clean	95.6	66.6	89.7	72.7	88.1	67.3
20	89.9	80.4	83.2	83.6	79.4	78.4
15	90.1	75.8	80.2	83.8	77.1	78.2
10	90.3	70.3	77.2	83.9	73.9	78.4
5	89.8	63.4	71.9	79.7	75.6	68.8
0	89.9	67.4	74.3	81.4	75.1	72.6
-5	87.1	45.5	51.0	75.0	70.7	54.4
Average (%)	90.4	67.1	75.4	80.0	77.1	71.1
Error norm (%)	34.3		31.7		36.8	

HR1: speech hit rate, HR0: nonspeech hit rate.

TABLE V
DETECTION ACCURACY, AURORA 3

Noise (dB)	VAD feature					
	LTSD		LTED		MTED	
	HR1	HR0	HR1	HR0	HR1	HR0
quiet	85.2	66.4	87.4	70.7	73.2	81.1
low	85.3	66.2	67.8	66.8	74.1	78
high	86.3	62.5	87.3	64.8	71.0	78.7
Average (%)	85.6	65.1	87.5	67.4	79.3	72.8
Error norm (%)	37.8		34.9		34.2	

HR1: speech hit rate, HR0: nonspeech hit rate.

higher individual hit rate performance than LTSD and an overall decrease in error of 7.7%. Minimum false alarm error is given by the MTED feature with a relative decrease of 9.5%, while both modulation feature-based algorithms outperform the LTSD in terms of the overall error under all three noise conditions, as can be seen in Fig. 10(b). Note that LTED is consistently best on both sets.

VI. CONCLUSION

The existence of modulations in speech, energy operators, and multiband analysis have been applied to problems regarding endpoint detection of isolated words in a "silent" background, speech analysis, and voice activity detection. We proposed speech detection algorithms based on new short-time signal features, derived through multiband filtering and modulation energy tracking. These features seem efficient in capturing slowly varying amplitude and frequency information, inherent in the modulation model for speech signals. Motivation also comes from a decision-theoretic analysis where, using multiple hypothesis testing, we derived a close link between the energy operator and an optimum detector of multiple sinusoids in noise.

Experimental results showed that an algorithm for endpoint detection with the developed features decreased the detection error of conventional time-domain features. It demonstrated improved robustness in noise: 32.5% detection error reduction at a 5-dB noise level on the TIMIT set, compared to the classic one and 40% reduction on the NTIMIT set. Additive noise, up to 30 dB, seemingly contributed beneficially as the proposed method at 30-dB SNR gave better results by a 45.3% compared to the noise-free case. The effectiveness of these features, especially of the multiband maximum average Teager energy, stems equally from their multiband nature and the duality of amplitude–frequency analysis.

In the direction of performance evaluation for speech detectors, a convenient method was proposed and used, following ideas from typical ROC analysis. Curves of detection probability versus a time parameter dictating certainty in endpoint estimation, can be applied to evaluate average performance even for nonstatistical endpoint detectors without explicit measurement of detected speech recognition accuracy.

Modulation energy tracking was also applied for voice activity detection. Based on a recent speech/noise divergence feature of high detection accuracy, we proposed a multiband Teager energy divergence and a long-term alternative for speech detection in alternating sequences of speech and nonspeech events. Through extensive evaluation on large databases, we aimed at minimizing an overall false alarm error norm. The proposed modulation features, although sporadically behaved worse compared to the reference detector, consistently decreased the overall error by 7.5% on Aurora 2 and 9.5% on Aurora 3 sets under varying noise levels and conditions while demonstrating a robust degradation response.

Labeling of speech events was approached through detection of speech modulations, tracking through multiple frequency bands their dominant structures, and measuring slowly-varying amplitude and frequency information. Modulation features were systematically verified to improve noisy speech detection on different benchmarks. Development of "smarter" detection algorithms based on these signal representations may lead to increased accuracy in speech boundary localization. Incorporation of these detectors in speech recognition and noise suppression schemes will dictate performance in terms of applications. As a side effect, the developed signal analysis methods and time-domain features may be further applied apart from detection, to speech analysis, recognition, segmentation, phoneme classification, or event detection.

REFERENCES

[1] J. G. Wilpon, L. R. Rabiner, and T. B. Martin, "An improved word-detection algorithm for telephone-quality speech incorporating both syntactic and semantic constraints," *AT&T Tech. J.*, vol. 63, no. 3, pp. 479–498, Mar. 1984.

[2] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 2, pp. 109–118, Feb. 2002.

[3] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Commun.*, vol. 42, no. 3–4, pp. 271–287, Apr. 2004.

[4] R. L. Bouquin-Jeannes and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech Commun.*, vol. 16, no. 3, pp. 245–254, Apr. 1995.

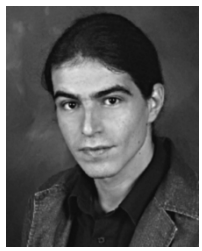
[5] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 3, pp. 146–157, Mar. 2002.

[6] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[7] Digital Cellular Telecommunications system (Phase 2+); Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels; General Description 1998 [Online]. Available: http://www.etsi.org, ETSI EN 301 708 v7.1.1

[8] F. Beritelli, S. Casale, G. Ruggeri, and S. Serrano, "Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 85–88, Mar. 2002.

[9] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.*, vol. 54, no. 2, pp. 297–315, Feb. 1975.
[10] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 4, pp. 777–785, Aug. 1981.
[11] M. H. Savoji, "A robust algorithm for accurate endpointing of speech," *Speech Commun.*, vol. 8, no. 1, pp. 45–60, Feb. 1989.
[12] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 3, pp. 201–212, Aug. 1976.
[13] Y. Qi and B. R. Hunt, "Voiced-unvoiced-silence classification of speech using hybrid features and a network classifier," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 2, pp. 250–255, Apr. 1993.
[14] G. D. Wu and C. T. Lin, "Word boundary detection with mel-scale frequency bank in noisy environment," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 5, pp. 541–553, Sep. 2000.
[15] S. G. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 478–482, Jul. 2000.
[16] K. Srinivasan and A. Gersho, "Voice activity detection for cellular networks," in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, Quebec, QC, Canada, Jun. 1993, pp. 85–86.
[17] W. H. Shin, B. S. Lee, Y. K. Lee, and J. S. Lee, "Speech/nonspeech classification using multiple features for robust endpoint detection," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Istanbul, Turkey, Jun. 2000, pp. 1399–1402.
[18] B. Kingsbury, P. Jain, and A. Adami, "A hybrid HMM/TRAPS model for robust voice activity detection," in *Proc. Int. Conf. Spoken Lang. Process.*, Denver, CO, Sep. 2002, pp. 1073–1076.
[19] Q. Li, J. Zheng, Q. Zhou, and C. H. Lee, "A robust, real-time endpoint detector with energy normalization for ASR in adverse environments," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Salt Lake City, UT, May 2001, pp. 233–236.
[20] D. L. Thomson and R. Chengalvarayan, "Use of periodicity and jitter as speech recognition features," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Washington, DC, May 1998, pp. 21–24.
[21] J. C. Junqua, B. Mak, and B. Reaves, "A robust algorithm for word boundary detection in the presence of noise," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 3, pp. 406–412, Jul. 1994.
[22] J. L. Shen, J. W. Hung, and L. S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," in *Proc. Int. Conf. Spoken Lang. Process.*, Sydney, Australia, Nov./Dec. 1998, pp. 1015–1018.
[23] S. E. Bou-Ghazale and K. Assaleh, "A robust endpoint detection of speech for noisy environments with application to automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, May 2002, pp. 3808–3811.
[24] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 217–231, Mar. 2001.
[25] G. Ying, C. Mitchell, and L. Jamieson, "Endpoint detection of isolated utterances based on a modified Teager energy measurement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Minneapolis, MN, Apr. 1993, pp. 732–735.
[26] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. J. Rubio, "A new Kullback–Leibler VAD for speech recognition in noise," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 266–269, Feb. 2004.
[27] J. Navarro-Mesa, A. Moreno-Bilbao, and E. Leleida-Solano, "An improved speech endpoint detection system in noisy environments by means of third-order spectra," *IEEE Signal Process. Lett.*, vol. 6, no. 9, pp. 224–226, Sep. 1999.
[28] J. S. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.
[29] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Process.*, vol. 41, no. 10, pp. 3024–3051, Oct. 1993.
[30] A. C. Bovik, P. Maragos, and T. F. Quatieri, "AM–FM energy detection and separation in noise using multiband energy operators," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3245–3265, Dec. 1993.
[31] H. M. Teager and S. M. Teager, "Evidence of nonlinear sound production mechanisms in the vocal tract," in *Speech Production and Speech Modeling*. Norwell, MA: Kluwer, 1990, pp. 241–261.
[32] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Proc. Int. Conf. Speech Signal Process.*, Albuquerque, NM, Apr. 1990, pp. 381–384.
[33] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3795–3806, Jun. 1996.
[34] D. Gabor, "Theory of communication," *J. Inst. Elect. Eng.*, vol. 93, no. III, pp. 429–457, 1946.
[35] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[36] L. Cohen, *Time-Frequency Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1995.
[37] D. Macho, *Spanish SDC-Aurora Database for ETSI STQ Aurora WI008 Advanced DSR Front-End Evaluation: Description and Baseline Results*. Sophia-Antipolis Cedex, France: STQ Aurora DSR Working Group, 2000, Input Doc. AU/271/00.
[38] T. Fawcett, "ROC graphs: Notes and practical considerations for data mining researchers" HP Laboratories, Palo Alto, CA, Tech. Rep. HPL-2003-4 [Online]. Available: http://www.hpl.hcom/techreports
[39] H. G. Hirsh and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions," in *ISCA ITRW ASR2000-ASR: Challenges for the Next Millennium*, Paris, France, Sep. 2000, pp. 181–188.



Georgios Evangelopoulos (S'02) was born in Grevena, Greece. He received the Diploma degree in electrical and computer engineering from the National Technical University of Athens (NTUA), Athens, Greece, in 2001. He is currently pursuing the Ph.D. degree with the Computer Vision, Speech Communication, and Signal Processing Group, NTUA, since 2002. His degree thesis, supervised by Prof. Petros Maragos, was in nonlinear signal analysis and speech–silence-noise discrimination.

His research interests are in the areas of nonlinear signal processing, generalized signal event detection, visual texture modeling and analysis with applications in the fields of speech processing, image analysis, and computer vision.

Petros Maragos (S'81–M'85–SM'91–F'95) received the Diploma degree in electrical engineering from the National Technical University of Athens, Athens, Greece, in 1980, and the M.Sc.E.E. and Ph.D. degrees from the Georgia Institute of Technology (Georgia Tech.), Atlanta, in 1982 and 1985, respectively.

In 1985, he joined the faculty of the Division of Applied Sciences, Harvard University, Cambridge, MA, where he worked for eight years as a Professor of electrical engineering, affiliated with the interdisciplinary Harvard Robotics Laboratory. He has also been a consultant to several industry research groups including Xerox's research on image analysis. In 1993, he joined the faculty of the School of Electrical and Computer Engineering, Georgia Tech. During parts of 1996 and 1998, he was on sabbatical and academic leave working as a Senior Researcher at the Institute for Language and Speech Processing, Athens, Greece. In 1998, he joined the faculty of the National Technical University of Athens, where he is currently working as a Professor of electrical and computer engineering. His current research and teaching interests include the general areas of signal processing, systems theory, communications, pattern recognition, and their applications to image processing and computer vision, and computer speech processing and recognition.

Dr. Maragos has received several awards, including a 1987 US National Science Foundation Presidential Young Investigator Award; the 1988 IEEE Signal Processing Society's Young Author Paper Award for the paper "Morphological Filters"; the 1994 IEEE Signal Processing Society's Senior Award and the 1995 IEEE Baker Award for the paper "Energy Separation in Signal Modulations with Application to Speech Analysis"; and the 1996 Pattern Recognition Society's Honorable Mention Award for the paper "Min–Max Classifiers."