Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2013) 21: 527 – 547

doi:10.3906/elk-1107-30

A unified approach to speech enhancement and voice activity detection

Ceyhan KASAP*, Mustafa Levent ARSLAN
Department of Electrical and Electronics Engineering, Boğaziçi University, İstanbul, Turkey

Received: 25.07.2011 • **Accepted:** 09.02.2012 • **Published Online:** 22.03.2013 • **Printed:** 22.04.2013

Abstract: In this paper, a unified system for voice activity detection (VAD) and speech enhancement is proposed. In the proposed system, there is mutual exchange of information between VAD and speech enhancement blocks. A new and robust VAD algorithm is implemented for the VAD block of the unified system. The newly proposed VAD algorithm uses a periodicity measure and an energy measure obtained from spectral energy distribution and spectral energy difference of the input speech data. For the speech enhancement block, the modified Wiener filtering (MWF) approach is utilized. It has been shown that the utilization of information exchange between the VAD and MWF algorithms in the unified system increases the performance of both algorithms and the proposed unified system improves the robustness of a speech recognition system significantly. Both of the enhanced algorithms are noniterative. Therefore, the proposed unified system is computationally attractive for real-time applications.

Key words: Speech enhancement, voice activity detection, noise suppression, modified Wiener filtering

1. Introduction

Voice activity detection (VAD) and speech enhancement systems have been extensively studied by the speech processing community since the 1970s because of their importance in many different applications like wireless communications, speech coding, speech recognition, hands-free conferences, and so on.

The main objective of a VAD system is to correctly decide if a given audio signal portion is speech or nonspeech. Determination of speech segments in a given signal can be considered a statistical hypothesis problem for VAD systems where the challenge is the determination of to which category (either speech or nonspeech) the given signal belongs. As indicated in [1], earlier algorithms for VAD were mostly based on various features like linear prediction coding (LPC) parameters [2], energy levels, formant shape [3], zero crossing rate (ZCR), cepstral coefficients [4], and the periodicity measure [5]. More recently, VAD systems that make use of various statistical models have been proposed [1, 6]. Speech enhancement systems, on the other hand, aim to minimize the effects of noise in audio signals. Speech enhancement can be either single-channel or multi-channel. In single-channel enhancement, speech is available from a single microphone, whereas multi-channel systems make use of more than one microphone [7]. Multi-channel speech enhancement techniques have the advantage of multiple signal inputs to the system and this enables better noise characterization and therefore better noise suppression. However, these systems are inherently more complex and this imposes constraints in terms of the algorithmic complexity and cost. In addition, there may be applications where multiple microphone input is not possible due to hardware constraints. For these reasons, single-channel speech enhancement techniques became more popular. Single channel speech enhancement algorithms may be classified under 3 main classes [8]: (i)

*Correspondence: ceyhankasap@gmail.com

Spectral-subtraction algorithms [9, 10, 11], (ii) Wiener filtering-based algorithms [12, 13, 14, 15, 16, 17], (iii) Subspace algorithms [18, 19].

This work is concentrated on implementing a unified system for VAD and single-channel speech enhancement. The following section describes the system model for the proposed unified system.

1.1. Unified system for VAD and speech enhancement

Most of the speech applications that incorporate VAD and speech enhancement blocks typically utilize these blocks separately and independently. One example might be speech recognition systems. Before processing the audio input for recognition, these systems typically require a VAD module in order to determine the utterances in the input. After the VAD module, a noise removal system is generally applied to reduce the background noise for efficient recognition. Our proposed unified system aims to provide a single and efficient framework for these seemingly separated functionalities.

Our basic motivation for the proposed unified system is primarily based on the assumption that VAD and speech enhancement problems are closely related. We expect that although VAD and speech enhancement systems operate for distinct purposes, the 2 systems might operate better if they mutually exchange information in a unified framework. Suppression of acoustic noise in the audio input signal would improve VAD performance. Similarly, the discrimination of the speech/nonspeech character of the input frames, which would be obtained from a VAD algorithm, would enable better noise attenuation. We propose a novel VAD algorithm to be used for the VAD block of the unified system. The newly proposed VAD algorithm uses a periodicity measure and an energy measure obtained from spectral energy distribution and spectral energy difference of the input speech data. Utilization of speech enhancement enables us to implement the so-called hybrid VAD algorithm to be used in the proposed unified system. Availability of noised suppressed frames in the hybrid VAD algorithm allows the extraction of more reliable VAD features and therefore improves the speech/nonspeech detection performance. The speech enhancement block of the unified system relies on the modified Wiener filtering (MWF) approach proposed in [7]. By making the VAD decisions explicitly available for the MWF algorithm, the algorithm is modified to implement the so-called enhanced modified Wiener filtering (EMWF) algorithm with increased noise suppression performance. Figure 1 depicts a schematic representation of our proposed unified system for VAD and speech enhancement. As can be seen in Figure 1, the proposed unified system incorporates the capabilities of speech/nonspeech detection and noise suppression in a single framework. The input speech signal to the system is partitioned into multiple outputs where each output contains separate speech segments from the input signal (VAD functionality). The outputs of the system comprise noise suppressed speech (speech enhancement functionality).

In this paper, we describe the components of our proposed system and present our evaluations to demonstrate the increased performances of hybrid VAD and EMWF algorithms that are implemented in the unified framework. The outline of the paper will be as follows. First, the new VAD algorithm and its hybrid version, which we use for the VAD block of the proposed unified system, are presented in Section 2. Then, in Section 3, the principles of the EMWF algorithm, which we use for the speech enhancement block of the proposed unified system, are explained. In Section 4, evaluations by both subjective and objective measures are demonstrated. Finally, Section 5 presents the discussion and conclusions.

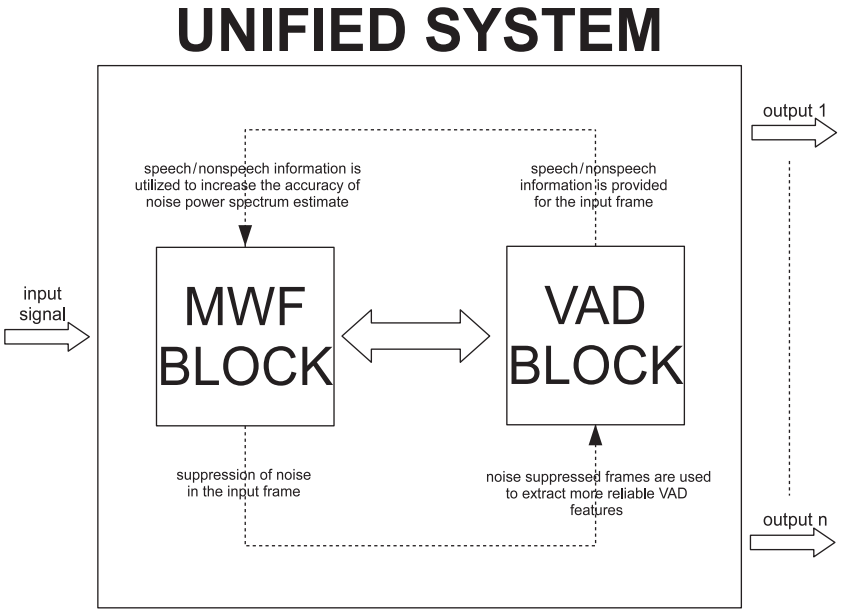

Figure 1. Unified system for VAD and speech enhancement.

2. VAD algorithm

The newly proposed VAD algorithm is specifically implemented to be used for the proposed unified system. In that sense, we do not aim to implement a VAD algorithm with extreme speech/nonspeech detection performance. For our proposed unified system, the error of treating nonspeech as speech at the beginning or end of speech regions is considered less harmful than classifying speech frames as nonspeech due to the information loss in the latter case. For this reason, our proposed VAD algorithm has a relaxed condition to find the exact talkspurt boundaries and small silence margins at speech boundaries are tolerated. Computational speed and robustness to noise are chosen as the most important criteria for the newly proposed VAD algorithm. Since the inclusion of every separate feature brings about the additional cost of computation, the features used for the proposed VAD algorithm are carefully chosen according to their computational complexity and their discriminative contribution to the final speech/nonspeech decision. Features used for the proposed VAD algorithm are periodicity measure and an energy measure obtained from spectral energy distribution and spectral energy difference of the input frames.

2.1. Periodicity measure

Periodicity in the input frame is an indication of speech character rather than nonspeech character because of the fact that voiced sounds in human speech are generated by periodic vibrations of the vocal cords. The periodicity measure used for the proposed VAD algorithm is the probability of voicing parameter, which is defined as the ratio of the peak autocorrelation lag to the signal energy (zeroth autocorrelation lag) [20]. The peak autocorrelation lag of the input frame is searched in a range between 3 ms (333 Hz) and 18 ms (55.5 Hz), since the typical fundamental frequency range of human speech tends to range from 200 Hz to 125 Hz. Probability of the voicing parameter for speech frames typically ranges from 0.3 to 0.7. Therefore, frames that have a periodicity measure greater than 0.8 are considered as noise.

2.2. Energy measure

It is expected that a rise in the energy level above a threshold indicates the presence of speech in the input frame. In order to compensate for the varying background noise energy, a varying energy threshold for speech is used in the proposed VAD. The energy threshold for speech is calculated from time varying minimum and maximum energy levels that adaptively define the extreme energy thresholds for speech frames. The energy parameter for a frame is computed as the weighted sum of 2 measures, namely the spectrally weighted energy measure and the spectrally weighted energy difference measure.

2.2.1. Spectrally weighted energy measure

Peterson and Barney [21] demonstrated that the first 3 formants, which carry much of the energy content of vowels in English, are located at frequencies lower than 3 kHz. Although unvoiced sounds display spectral concentration at higher frequencies, a spectrum range up to 4 kHz contains much of the energy content of human speech. Moreover, average energy distribution up to 4 kHz range is not uniform for human speech. Lower frequency bands generally carry more energy than do upper frequency bands [22].

In the light of the above arguments, the proposed VAD extracts the energy content of the input frame in 5 equal length frequency subbands in order to compute the spectrally weighted energy measure. The energy content of the frame is considered to be located in the 300 Hz to 3400 Hz range (i.e. each subband has a length of 620 Hz). FFT of the input frame is used to calculate the total power for each subband and the log_{10} value of the total power in each frequency subband is interpreted as the energy measure of that specific subband. Heuristically determined weighting coefficients, which are shown in Figure 2, are applied to the energy measures calculated for each subband. These weighted energy values are then summed to obtain the spectrally weighted energy measure for the frame. The algorithm used to compute the spectrally weighted energy parameter, *m_energyWeighted*, is shown in Figure 3.

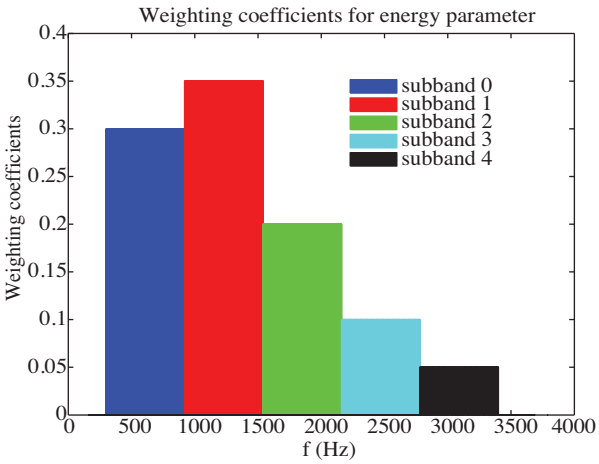

Figure 2. Weighting coefficients of subbands for spectrally weighted energy measure computation.

2.2.2. Spectrally weighted energy difference measure

Since the average energy distribution over frequency is not uniform for human speech, large energy differences over the 0 to 4 kHz frequency range are supposed to be an indication of the presence of actual speech in the input signal. The proposed VAD utilizes this fact by calculating a spectrally weighted energy difference measure.

Algorithm to compute spectrally weighted energy parameter, *m_energyWeighted*

i) Compute subband energies in 5 bands

$$m\_subBandEnergy[i] \text{ for } i = 0 \ldots 4$$

ii) Compute weighted sum of subband energies

$$\begin{aligned} m\_energyWeighted \quad = \quad & 0.30 * m\_subBandEnergy[0] + 0.35 * m\_subBandEnergy[1] \\ & +0.20 * m\_subBandEnergy[2] + 0.10 * m\_subBandEnergy[3] \\ & +0.05 * m\_subBandEnergy[4] \end{aligned}$$

Figure 3. Algorithm to compute the spectrally weighted energy parameter, *m_energyWeighted*.

The same subbands that are used to calculate the spectrally weighted energy measure are also utilized for the calculation of the spectrally weighted energy difference measure. First, an average subband energy value for each of the 5 subbands is calculated for the current frame. If it is the first frame, the average subband energy is simply set to the total energy of that subband. If it is not the first frame, 90% of the previous average subband energy is summed with 10% of the current total subband energy to calculate the average energy. The difference between the average subband energy and the total subband energy gives the subband energy difference value for that subband. Lastly, heuristically determined weighting coefficients, which are shown in Figure 4, are applied to the energy difference values calculated for each subband. The sum of these weighted energy difference values gives the spectrally weighted energy difference measure for the frame. The algorithm used to compute the spectrally weighted energy difference parameter, *m_energyDifference*, is shown in Figure 5.

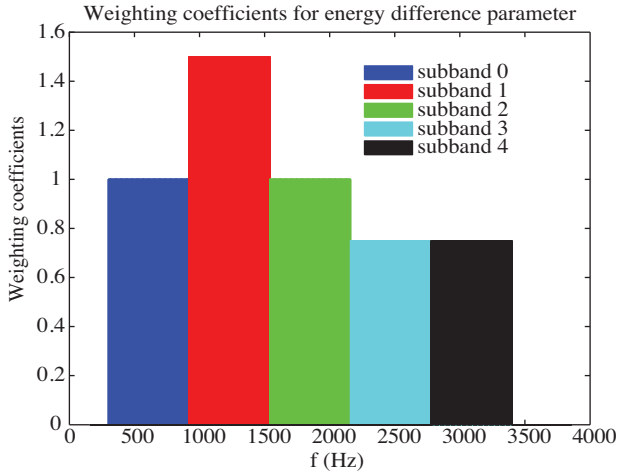

Figure 4. Weighting coefficients of subbands for spectrally weighted energy difference measure computation.

2.3. Soft decision assignment and decision smoothing

After the features are extracted from the input frame, the proposed VAD associates a soft decision value, rather than a strict speech/nonspeech decision, with the frame. The soft decision value, *soft_decision*, assigned to a frame is nonnegative and increases as the speech likeliness of the frame increases. It is calculated as a function of 3 parameters: (i) energy measure for the current frame, *m_energy*, (ii) speech threshold for the current frame, *speech_threshold*, and (iii) probability of voicing measure for the current frame, *prob_voice*.

Algorithm to compute spectrally weighted energy difference parameter, *m_energyDifference*

i) Compute subband average energy for each band, *m_subBandAvg[i]* for $i = 0 \ldots 4$

if input frame is the first frame **then**

$$m\_subBandAvg[i] = m\_subBandEnergy[i]$$

end if

else

$$m\_subBandAvg[i] = m\_subBandAvg[i] * 0.9 + m\_subbandEnergy[i] * 0.1$$

end else

ii) Compute subband energy difference for each band, *m_subBandEnergyDifference* for $i = 0 \ldots 4$

$$m\_subBandEnergyDifference[i] = m\_subBandEnergy[i] - m\_subBandAvg[i]$$

iii) Compute weighted sum of subband energy differences

$$\begin{aligned} m\_energyDifference \;=\; & 1.00 * m\_subBandEnergyDifference[0] \\ & +1.50 * m\_subBandEnergyDifference[1] \\ & +1.00 * m\_subBandEnergyDifference[2] \\ & +0.75 * m\_subBandEnergyDifference[3] \\ & +0.75 * m\_subBandEnergyDifference[4] \end{aligned}$$

Figure 5. Algorithm to compute the spectrally weighted energy difference parameter, *m_energyDifference*.

The energy measure for a frame, *m_energy*, is computed as

$$\begin{aligned} m\_energy \;=\; & \Big(1.10 * m\_energyWeighted\Big) + \Big(0.25 * min(m\_energyDifference, 2)\Big) \\ & +\Big(min(1.0, 0.5 * prob\_voice)\Big) \end{aligned} \tag{1}$$

where *m_energyWeighted* is the spectrally weighted energy measure computed according to the algorithm shown in Figure 3, *m_energyDifference* is the spectrally weighted energy difference measure computed according to the algorithm shown in Figure 5, and $min(x, y)$ represents the minimum function that returns the minimum of x and y.

Speech threshold for a frame, *speech_threshold* is computed as

$$\begin{aligned} speech\_threshold \;=\; & \Big[0.01 * \big(m\_maxEnergy - m\_minEnergy\big) * \\ & \big(40 + 5 * (10 - sensitivity)\big)\Big] + m\_minEnergy \end{aligned} \tag{2}$$

where *m_maxEnergy* is the expected maximum energy level for a speech frame, *m_minEnergy* is the expected minimum energy level for a speech frame, and *sensitivity* is a configurable input parameter with a default value of 3. Note that *m_maxEnergy* and *m_minEnergy* are adaptive during the operation of VAD.

Lastly, the soft decision value for a frame, *soft_decision*, is computed as

$$soft\_decision = \begin{cases} 0.75 + m\_energy - speech\_threshold & \text{if } m\_energy \geq threshold \text{ and } prob\_voice > 0.4 \\ 0.5 + m\_energy - speech\_threshold & \text{if } m\_energy \geq threshold \text{ and } prob\_voice \leq 0.4 \\ 0 & \text{if } m\_energy < threshold \end{cases} \tag{3}$$

where $threshold = speech\_threshold - 0.5$.

The proposed VAD implements decision smoothing for better speech/nonspeech characterization. The implemented decision smoothing approach relies on the heuristic rule of increasing the soft decision value for the current frame if the previous frame is tagged as a speech frame. Speech/nonspeech determination for the processed frame is based on the history of soft decisions for the last 20 frames. The sum of the soft decision values of the last 20 frames is used to make the final speech/nonspeech determination. The reason for using a history of soft decisions is the fact that human speech may show very large energy variations even in very short time intervals and speech/nonspeech determination based on a single soft decision may cause low energy speech frames to be misevaluated as nonspeech.

2.4. Operation of the proposed VAD

The proposed VAD system analyzes the audio input and separates the input into numerous portions, where each portion contains actual speech in a distinct output stream (i.e. the nonspeech portions between the utterances are removed). A circular speech buffer continuously reads the input speech signal frame by frame. A frame length of 20 ms and a skip length of 10 ms are used for the system. Based on the speech/nonspeech decisions made for the frames, the system is either in a *speech started* or *speech ended* state. The system is initially at *speech ended* state. Detection of speech in the current frame makes the VAD system switch its state from *speech ended* to *speech started*. This in turn triggers the system to start writing the input signal into a separate output stream. After a while, when the utterance ends, detection of nonspeech in the incoming frames triggers the VAD system to switch to the *speech ended* state and this stops the writing operation. Transitions between the 2 states enable the partitioning of the input signal into separate output streams and these transitions continue until the whole audio input stream is processed by the system.

A common problem in all VAD systems is the difficulty in locating the beginning and end of an utterance if there are: a) weak fricatives at the beginning or end, b) weak plosive bursts at the beginning or end, c) nasals at the end, d) voiced fricatives that become devoiced at the end of words, and e) trailing off of vowel sounds at the end of an utterance [23]. The proposed VAD algorithm uses configurable *prespeech buffer* and *postspeech buffer* values in order to overcome these problems. These parameters extend the duration of speech decisions and enable the detection of actual speech instances that are very much likely to be missed by the VAD algorithm.

Other than *prespeech buffer* and *postspeech buffer*, there are 3 additional configurable parameters, namely *speech trigger*, *silence trigger*, and *sensitivity*, which are used to adjust the sensitivity of the proposed VAD. *Speech trigger* has a default value of 8 and it determines the threshold for the speech started decision. When the sum of the objective soft decision values of the last 20 frames exceeds the *speech trigger* value, the current frame is considered an actual speech instance and the system transitions to *speech started* state (if the current state is already *speech started*, the current state is preserved). *Silence trigger* has a default value of 700 ms and it determines the total required signal length to trigger the speech ended decision. The system keeps track of the total length of successive frames, where the sum of the objective soft decision values in the previous 20 frames is smaller than the value of *speech trigger* for each frame. When this length exceeds the *silence trigger*, the current state of the system is set to *speech ended*. *Sensitivity* has a default value of 3 and it is restricted to be between 0 and 12. As can be deduced from Eq. (3), a decreased *sensitivity* value increases the signal energy threshold required to achieve a positive soft decision value. Therefore, under low SNR conditions, lower *sensitivity* values must be used to avoid false triggering.

The default values for the configurable parameters used for the proposed VAD algorithm are aimed to be optimized for a wide range of audio signals under different SNR conditions. They are finalized after an extensive number of observations. However, if there is a priori knowledge about the characteristics of the audio signal to be processed, these parameters may be altered for better performance. The flowchart of the proposed VAD algorithm is shown in Figure 6.

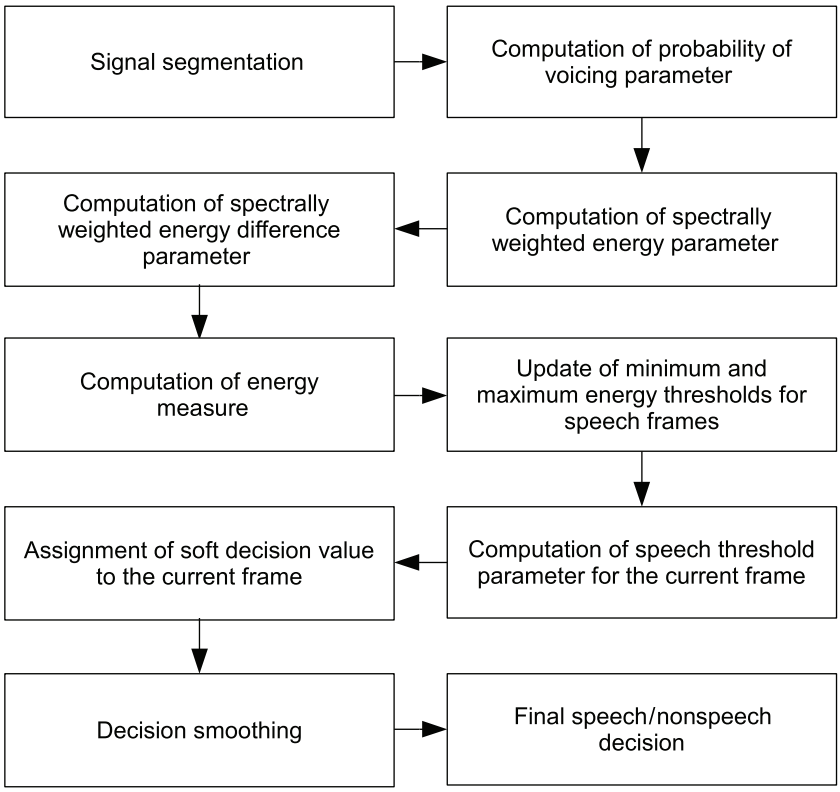

Figure 6. Flowchart for the proposed VAD algorithm.

2.5. Hybrid VAD

As expected, performance of the proposed VAD is reduced under low SNR conditions. After experiments, 2 main problems were observed under low SNR conditions. Firstly, noise frames with high energy could easily be detected as actual speech. Secondly, detection of actual speech frames that contain low energy was problematic when the SNR was low.

The hybrid VAD algorithm, which we implement as the VAD block of our proposed unified system for VAD and speech enhancement, tries to eliminate the above problems by utilizing speech enhancement. By making use of speech enhancement in the form of MWF, performance of the proposed VAD is improved. Two main modifications are done in the proposed VAD algorithm to implement the hybrid VAD algorithm. Firstly, noise suppressed frames are used to extract features for the hybrid VAD algorithm. This results in better periodicity characterization. Secondly, since spectral energy variation is more distinct for noise suppressed frames, contribution of the spectrally weighted energy difference measure in the final speech/nonspeech decision is emphasized by modifying Eq. (1) to

$$m\_energy = \Big(1.10 * m\_energyWeighted\Big) + \Big(0.375 * min(m\_energyDifference, 2)\Big) + \Big(min(1.0, 0.5 * prob\_voice)\Big) \quad (4)$$

for the hybrid VAD. The net effect of these modifications is the increased accuracy of soft decision values assigned to frames.

3. Speech enhancement algorithm

The speech enhancement block of the unified system relies on the MWF algorithm proposed in [7]. The MWF algorithm proposed in [7] uses an SNR dependent noise suppression factor with the aim of employing an aggressive enhancement at nonspeech intervals and a milder filtering at the speech segments. By increasing the noise suppression factor for regions where speech is not likely to be present, the algorithm reduces the Wiener filter gain in order to employ aggressive filtering. That way, the degraded parts of the signal are suppressed. Conversely, for regions where speech is likely to be present, the algorithm increases the Wiener filter gain in order to employ only mild filtering. That way, the distortion/suppression of the speech segments is prevented. The bottleneck of the MWF algorithm is the absence of an explicit speech/nonspeech decision for the processed frame. Such a decision would help in adjusting the level of filtering. Our new algorithm improves the MWF from this perspective. By making the VAD decisions explicitly available for the MWF algorithm, the algorithm is modified to implement the so-called EMWF algorithm with increased performance. The following section presents a preliminary explanation of the basics of the MWF algorithm.

3.1. Modified Wiener filtering

For an additive noise signal model of $y(t) = s(t) + n(t)$, where $y(t)$ is noisy speech, $s(t)$ is noise-free speech, and $n(t)$ is noise signal, a generalized Wiener filter can be formulated as

$$H(\omega) = \left(\frac{\hat{P}_s(\omega)}{\hat{P}_s(\omega) + \alpha P_n(\omega)} \right)^{\beta} \quad (5)$$

where $\hat{P}_s(\omega)$ is the clean speech power spectrum estimate, $P_n(\omega)$ is the noise power spectrum, α is the noise suppression factor, and β is the power of the filter. Application of this filter to the noisy input speech signal produces an estimate for the noise-free speech signal. It is assumed that the noisy speech and noise-free speech have the same phase and so the filter just alters the amplitude at each frequency. Thus, we have

$$\hat{S}(\omega) = H(\omega)Y(\omega) \quad (6)$$

$$\hat{s}(t) = F^{-1}\{\hat{S}(\omega)\} \quad (7)$$

where $Y(\omega)$ is the Fourier transform of the noisy speech, $F^{-1}\{.\}$ is the inverse Fourier transform operation, and $\hat{S}(\omega)$ is the estimate of the Fourier transform of the clean speech signal. In this formulation, it is assumed that we have an estimate of the clean speech power spectrum, $\hat{P}_s(\omega)$. This estimate is calculated from the

Fourier transform of the LPC coefficients of the noisy speech, $P_y(\omega)$, by only a DC gain modification of $P_y(\omega)$ as

$$\hat{P}_s(\omega) = \frac{\hat{g_s}^2}{g_y{}^2} P_y(\omega) \tag{8}$$

where $\hat{g_s}$ and g_y are the DC gains of the noise-free speech signal and the noisy speech signal, respectively. The MWF algorithm assumes that noise and speech are uncorrelated and power spectra of the noisy speech signal, noise-free speech signal, and noise signal are related as

$$P_y(\omega) = \hat{P}_s(\omega) + P_n(\omega) \tag{9}$$

If we integrate both sides of the equation over ω and use the expression for $\hat{P}_s(\omega)$ stated in Eq. (8) we have

$$\int_{-\pi}^{\pi} P_y(\omega) d\omega = \int_{-\pi}^{\pi} \frac{\hat{g_s}^2}{g_y{}^2} P_y(\omega) d\omega + \int_{-\pi}^{\pi} P_n(\omega) d\omega \tag{10}$$

Using Parseval's relation, the above equation can be simplified to

$$\frac{\hat{g_s}^2}{g_y{}^2} = \begin{cases} \frac{E_y - E_n}{E_y} & \text{if } E_y > E_n, \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where E_n is the noise energy and E_y is the noisy speech energy. If we substitute the expression for $\hat{g_s}^2/g_y{}^2$ in the dc gain modification equation, the clean speech spectrum estimate becomes

$$\hat{P}_s(\omega) = \frac{E_y - E_n}{E_y} P_y(\omega) \tag{12}$$

Using the above expression in Eq. (5) and introducing a time-dependent noise suppression factor α_t we obtain

$$H(\omega) = \left(\frac{[(E_y - E_n)/E_y] P_y(\omega)}{[(E_y - E_n)/E_y] P_y(\omega) + \alpha_t P_n(\omega)} \right)^{\beta} \tag{13}$$

The above equation can be simplified to

$$H(\omega) = \left(\frac{P_y(\omega)}{P_y(\omega) + [E_y/(E_y - E_n)] \alpha_t P_n(\omega)} \right)^{\beta} \tag{14}$$

Eq. (14) indicates that more aggressive filtering is applied for increasing values of α_t. For proper speech enhancement, the value of α_t must be high for noise only frames and low for speech only frames; i.e. an inverse relation between the SNR value of the frame (E_s/E_n) and α_t must be introduced. This inverse relation is simply obtained by replacing α_t with $E_n/E_y\ \alpha'$ where α' is a constant. With this modification Eq. (14) becomes

$$H(\omega) = \left(\frac{P_y(\omega)}{P_y(\omega) + [E_n/(E_y - E_n)] \alpha' P_n(\omega)} \right)^{\beta} \tag{15}$$

Let us denote the time dependent multiplication factor that scales the noise spectrum, the $[E_n/(E_y - E_n)]\alpha'$ term, by λ_t. Then the above equation is equivalent to

$$H(\omega) = \left(\frac{P_y(\omega)}{P_y(\omega) + \lambda_t P_n(\omega)} \right)^{\beta} \tag{16}$$

3.2. Enhanced modified Wiener filtering

The novelty of the EMWF algorithm lies in the utilization of VAD decisions for frames. Compared with MWF, which employs "hard" noise spectrum update rules, usage of VAD decisions enables the EMWF algorithm to apply more flexible updates for noise spectrum estimation. The EMWF algorithm employs more aggressive updates for increasing the estimated noise spectrum during speech regions (EMWF uses a higher static upconstant value (10 dB/s for 10 ms skip length) compared to MWF (3.5 dB/s for 10 ms skip length)). This results in more rapid convergence of the actual and estimated noise spectra. During speech regions, a variable upconstant value with an upper limit of 7 dB/s for 10 ms skip length is used to increase the noise power spectrum estimate. This variable upconstant value with an upper limit is required to prevent erroneous upwards updates. The net result of these enhancements is the increased accuracy of the noise spectrum estimate. Increased accuracy of the noise spectrum estimate enables better noise attenuation performance. The EMWF algorithm is noniterative like the original MWF and hence it is also attractive for real-time implementation.

Step-by-step algorithm description of the new EMWF algorithm is as follows:

Step 1) A frame length of 20 ms with a skip length of 10 ms is provided as the input to the algorithm.

Step 2) Hanning window is applied on the frame.

Step 3) Autocorrelation lags of order 18 are calculated for the input frame. Let us denote the index of the successive input frames by k. If this is not the first frame ($k \neq 0$), an interpolation factor of $\gamma = 0.7$ is applied on the autocorrelation lags of the k^{th} frame (i^{th} autocorrelation lag for the k^{th} frame, $R[i]_k$, is set to $R[i]_k = \gamma R[i]_k + (1-\gamma)R[i]_{k-1}$). If this is the first frame ($k = 0$), autocorrelation lags are unchanged. Then, 18^{th} order LPC coefficients are calculated from the autocorrelation lags using Durbin's recursive procedure [24]. Finally, DFT of the LPC coefficients are calculated in order to find $P_y(\omega)_k$, power spectrum of the k^{th} noisy input frame.

Step 4) The noise power spectrum estimate for the k^{th} frame, $P_n(\omega)_k$, is calculated for each frequency. If it is the first frame ($k = 0$), it is assumed that the first frame purely denotes noise and the initial noise power spectrum estimate, $P_n(\omega)_0$, is calculated by taking DFT of 8th order LPC coefficients of the input frame. The reason for not directly setting $P_n(\omega)_0 = P_y(\omega)_0$ and using an order of 8 for LPC coefficients instead is that LPC order of 8 results in a smoother spectrum compared to 18. Test simulations verified that this initial smoothed spectrum resulted in less distortion.

If it is not the first frame ($k \neq 0$), the noise power spectrum estimate for the k^{th} frame, $P_n(\omega)_k$, is found by an update of the previous value $P_n(\omega)_{k-1}$. The equation for $P_n(\omega)_k$ is

$$P_n(\omega)_k = \begin{cases} 1.023P_n(\omega)_{k-1} & \text{if } P_y(\omega)_k > P_n(\omega)_{k-1} \text{ and signal is in nonspeech region} \\ (1+\delta)P_n(\omega)_{k-1} & \text{if } P_y(\omega)_k > P_n(\omega)_{k-1} \text{ and signal is in speech region} \\ P_n(\omega)_{k-1} & \text{if } P_y(\omega)_k = P_n(\omega)_{k-1} \\ 0.933P_n(\omega)_{k-1} & \text{if } P_y(\omega)_k < P_n(\omega)_{k-1} \end{cases} \tag{17}$$

In the above equation, VAD outputs are used to decide whether the signal is in the speech region or not. Soft decision values for the last 20 frames, which are provided by the VAD algorithm, are summed. If the sum is found to be smaller than 10, then the current frame is tagged as nonspeech. Conversely, if the sum is higher than 10, this is interpreted as showing that the signal is currently in the speech region.

The $1+\delta$ term in Eq. (17) represents the variable upconstant that is inversely related to the log energy difference between the current frame energy, $(E_y)_k$, and average noise energy estimate, $\hat{E}_n$. The inverse relation

is achieved by setting δ to $\delta = 1 \ / \ 1000(\log_{10}(E_y)_k - \log_{10}\hat{E_n})$. As previously mentioned, we prefer to restrict the upwards noise update within certain limits when we are in the actual speech region in order to minimize errors. Therefore, an upper limit of 1.016 (7 dB/s for 10 ms skip length) is set for $1 + \delta$.

Step 5) The time dependent λ_t factor that scales the noise spectrum Eq. (16) is calculated. After test simulations, it has been determined that an exponential relation, rather than a linear relation, between the SNR and the scaling factor λ_t results in less distorted speech and the time dependent scaling factor is found as

$$\lambda_t = \left(\frac{(E_n)_k}{max\left[((E_y)_k - (E_n)_k), ((E_n)_k/50)\right]} \right)^{\mu} \nu \tag{18}$$

where $\mu = 0.4$ and $\nu = 63.01$ (18 dB) are heuristically determined constants.

Step 6) The Wiener filter gain for each frequency is calculated from Eq. (16), where $\beta = 0.5$ and the value of λ_t is determined from Step 5.

Step 7) DFT of the k^{th} noisy-speech frame, $Y(\omega)_k$, is calculated.

Step 8) The spectrum of the k^{th} noise-free speech frame, $\hat{S}(\omega)_k$, is found by multiplying $Y(\omega)_k$ with $H(\omega)_k$ at each frequency.

Step 9) The real part of the inverse DFT of $\hat{S}(\omega)_k$ is calculated to obtain the k^{th} noise-free speech frame $\hat{s}_k$.

Step 10) The overlap-add method is used for combining the filtered frames to form the overall enhanced signal output.

4. Evaluations

4.1. Evaluations for the proposed VAD

Although implementing a VAD algorithm with superior speech/nonspeech detection performance was not the main focus of this study and our proposed VAD algorithm has a relaxed condition to find the exact talkspurt boundaries, we still wanted to have a quantitative measure for the performance of the newly proposed VAD algorithm. To evaluate the performance of the newly proposed VAD algorithm, we compared our algorithm with 2 standard VAD algorithms, namely ITU-T G.729 Annex B VAD [25] and ETSI Adaptive Multi-Rate (AMR) VAD option 2 [26]. The database for the VAD performance comparison experiment was prepared by adding white Gaussian noise to clean speech signals (5 TIMIT database sentences) at SNR levels of 5, 10, 15, 20, and 25 dB. We used a configuration of *prespeech buffer* = 10 ms, *postspeech buffer* = 15 ms, *speech trigger* = 8, *silence trigger* = 700 ms, and *sensitivity* = 3, for our VAD algorithm. Figure 7 graphically shows the total error for all algorithms at varying SNR levels. As can be seen in the figure, both of the standard VADs have better performances than the proposed VAD at high SNR levels. However, at low SNR conditions, the performance of the proposed VAD is better than that of G.729 VAD and slightly worse than that of AMR VAD.

4.2. Hybrid VAD improvements

In order to demonstrate the increased performance of the hybrid VAD, we used 10 recordings of actual noisy speech data in Turkish that are collected in a car under different conditions. The properties of the recordings are provided in Table 1. These recordings contain a total of 644 words, phrases, or sentences spoken by 4 male speakers. The recordings were presented as inputs both to the newly proposed VAD algorithm and to its hybrid

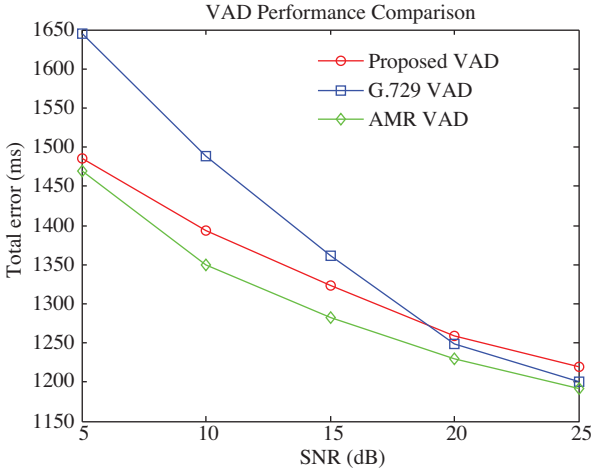

Figure 7. Performance comparison of the proposed VAD with G.729 VAD and AMR VAD under additive white Gaussian noise.

version. Among the 644 words, phrases, or sentences, 484 of them were successfully detected by the proposed VAD, whereas the detected instance number increased to 584 for the hybrid VAD. For the common detected words, phrases, or sentences, a comparative error analysis was performed. Table 2 demonstrates the error analysis performance comparison. As can be seen in Table 2, the total errors in speech/nonspeech boundary decisions are smaller in the hybrid VAD compared to the standard VAD. The detection rate, which is computed as the ratio of the number of successfully detected utterances to the number of actual utterances, is also higher for the hybrid VAD. The false alarm rate, which is computed as the ratio of the number of faulty detected noise only utterances to the number of actual utterances, is smaller for the hybrid VAD too.

Table 1. Properties of the recordings used for VAD performance measurements.

Sample	Sample properties	Sample length (min:s)	Number of utterances
Sample 1	air conditioner on at level 2, windows open, noisy traffic	6:02	182
Sample 2	air conditioner off, windows open, traffic	2:30	46
Sample 3	air conditioner on at level 1, windows closed, traffic	2:43	47
Sample 4	air conditioner on at level 2, windows closed, car at idle	1:53	46
Sample 5	air conditioner on at level 2, windows closed, traffic	2:58	51
Sample 6	air conditioner on at level 3, windows closed, car at idle	1:51	46
Sample 7	air conditioner off, windows open, noisy traffic	3:47	90
Sample 8	air conditioner off, windows open, noisy traffic	1:26	45
Sample 9	air conditioner off, windows open, traffic	1:35	44
Sample 10	air conditioner off, windows open, traffic	1:49	47

Table 2. Performance comparison of proposed VAD and Hybrid VAD.

Sample	Total error in proposed VAD (ms)	Total error in Hybrid VAD (ms)	Detection rate in proposed VAD (%)	Detection rate in Hybrid VAD (%)	False alarm rate in proposed VAD (%)	False alarm rate in Hybrid VAD (%)
Sample 1	20,235	19,165	62.6%	90.1%	2.7%	1.6%
Sample 2	9194	8316	95.7%	100.0%	0	0
Sample 3	7828	7679	95.7%	100.0%	0	0
Sample 4	12,044	7118	84.8%	100.0%	0	0
Sample 5	7117	6353	98.0%	100.0%	0	0
Sample 6	8964	6308	95.7%	100.0%	0	0
Sample 7	8431	8999	25.6%	60.0%	4.4%	2.2%
Sample 8	9831	6657	80.0%	88.9%	2.2%	2.2%
Sample 9	6435	6924	97.7%	100.0%	0	0
Sample 10	6252	5131	97.9%	97.9%	0	0
Total	96,331	82,650	75.2%	90.7%	0.015%	0.009%

4.3. Enhanced modified Wiener filtering improvements

The accuracy of the noise power spectrum estimate is the key factor for increased performance of the Wiener filtering-based speech enhancement algorithms. In order to compare the accuracy of the noise power spectrum estimates of EMWF and MWF, we first added 0.5 s of silence to 5 different TIMIT database utterances. Then car noise from the NOISEX database was added to the clean speech signals at 2 different SNR levels, 5 dB and 20 dB. Noisy signals were filtered using both methods separately. At every 10th frame, starting from the first frame, estimated noise power spectra of the algorithms were compared with real noise power spectra.

Figures 8 and 9 demonstrate noise power spectra comparisons for a TIMIT database utterance with 20 dB car noise at several frame indices. As can be seen in these figures, during nonspeech intervals, the EMWF algorithm employs more aggressive updates for the estimated noise spectrum. This enables faster convergence of the estimated noise power spectrum to the real noise power spectrum. During speech intervals, noise power spectrum estimate update is restricted within smaller ranges in order to minimize errors. The net effect is the increased accuracy of noise power spectrum estimation for EMWF compared to MWF.

In order to obtain a quantitative measure for increased noise power spectrum estimation accuracy of the EMWF algorithm, spectral distortion between estimated and real noise power spectra was computed at every 10th frame starting from the first frame for both of the algorithms. Then an overall spectral distortion measure, SD, was computed as

$$SD = \frac{10}{L} \sum_{i=0}^{L-1} \int_0^{Fs} \left[ln|P_n(\omega)| - ln|\hat{P}_n(\omega)| \right]^2 d\omega \tag{19}$$

where $P_n(\omega)$ is the real noise spectrum, $\hat{P}_n(\omega)$ is the estimated noise spectrum, and L is the number of frames used in computation. Computed spectral distortion ratios of SD_{EMWF}/SD_{MWF} for each sample are tabulated in Table 3. Spectral distortion ratios, which are smaller than 1 for all samples, demonstrate the increased noise power spectrum estimation accuracy of the EMWF algorithm compared to the MWF algorithm.

Another experiment was performed in order to compare the objective quality measures of MWF and EMWF outputs. The database for this experiment was prepared by adding white Gaussian noise to clean speech signals (20 TIMIT database utterances) at 10 dB SNR. This database was then used to compare the

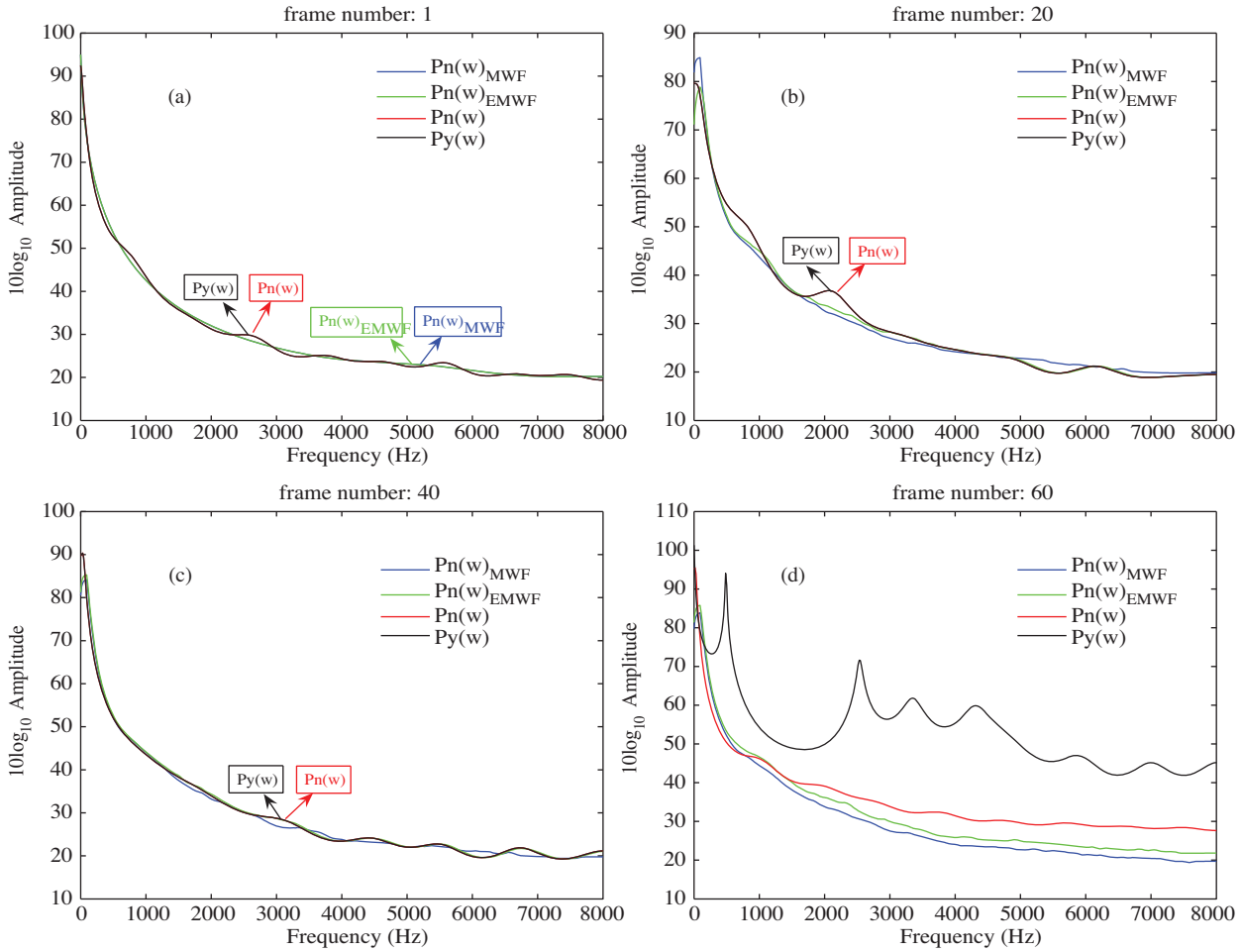

Figure 8. Comparison of noise power spectrum estimate of MWF $(Pn(w)_{\mathrm{MWF}})$, noise power spectrum estimate of EMWF $(Pn(w)_{\mathrm{EMWF}})$, real noise power spectrum $(Pn(w))$, and noisy input speech power spectrum $(Py(w))$ for TIMIT database utterance with 20 dB car noise at frames (a) 1, (b) 20, (c) 40, and (d) 60.

perceptual evaluation of speech quality (PESQ) scores of the noisy and enhanced speech signals, where the enhancement was implemented by applying MWF and EMWF algorithms to the noisy signals. PESQ scores

Table 3. Comparison of spectral distortion measures for noise power spectrum estimation.

Sample used	$\mathrm{SD}_{\mathrm{EMWF}}/\mathrm{SD}_{\mathrm{MWF}}$
Sample 1 at 5 dB SNR	0.67
Sample 1 at 20 dB SNR	0.42
Sample 2 at 5 dB SNR	0.68
Sample 2 at 20 dB SNR	0.45
Sample 3 at 5 dB SNR	0.53
Sample 3 at 20 dB SNR	0.52
Sample 4 at 5 dB SNR	0.88
Sample 4 at 20 dB SNR	0.60
Sample 5 at 5 dB SNR	0.67
Sample 5 at 20 dB SNR	0.46

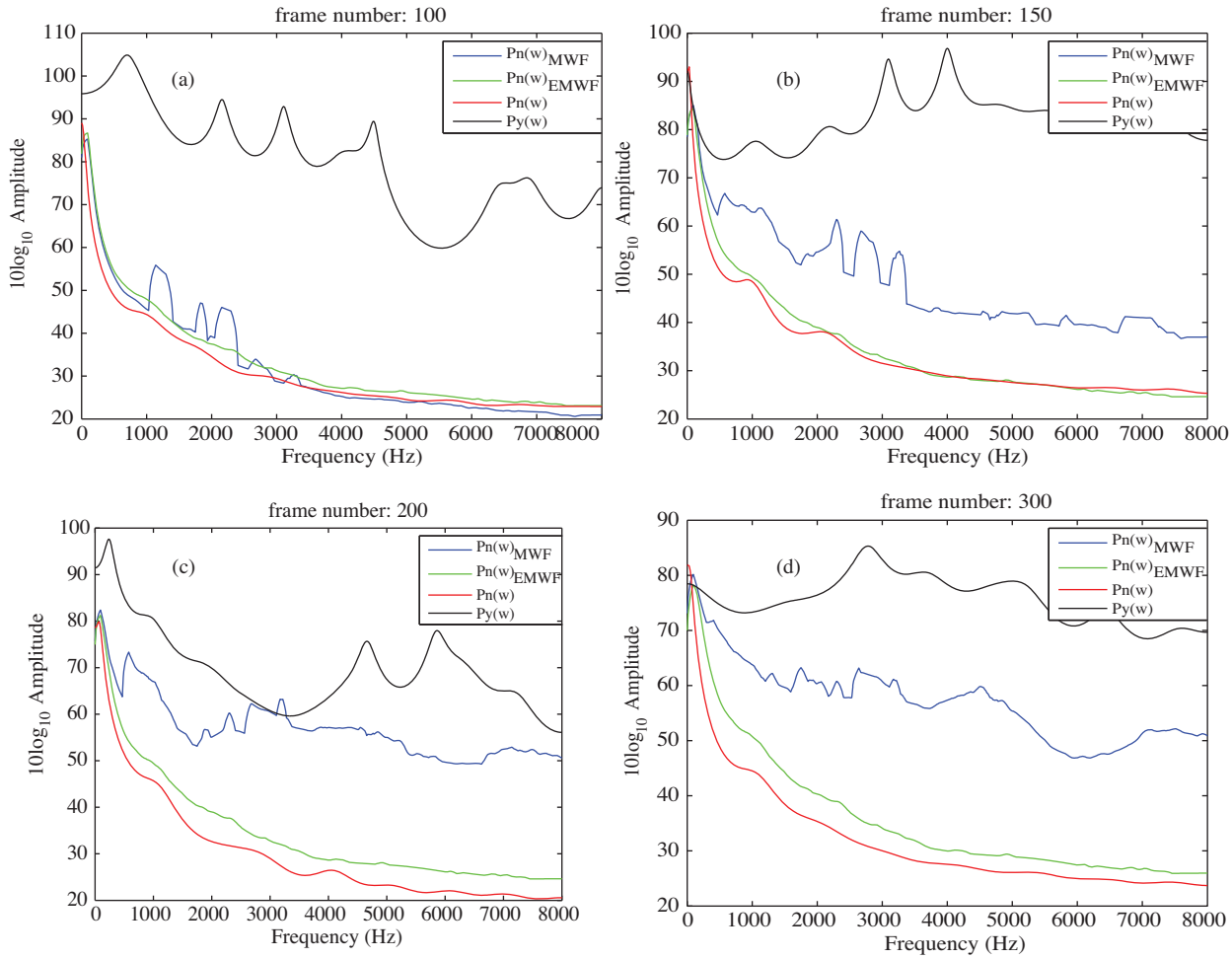

Figure 9. Comparison of noise power spectrum estimate of MWF ($Pn(w)_{\mathrm{MWF}}$), noise power spectrum estimate of EMWF ($Pn(w)_{\mathrm{EMWF}}$), real noise power spectrum ($Pn(w)$), and noisy input speech power spectrum ($Py(w)$) for TIMIT database utterance with 20 dB car noise at frames (a) 100, (b) 150, (c) 200, and (d) 300.

are shown in Table 4. As can be seen in the results, the average PESQ score was 2.163 for noisy speech, 2.276 for the MWF method, and 2.4 for the EMWF method. This indicates that both algorithms increase the speech quality and EMWF achieves better PESQ scores compared to the original MWF.

We also wanted to evaluate the performance increase in the objective quality for different SNR levels. For this purpose, we added speech babble, car, pink, and white noise to 5 TIMIT database sentences at SNR levels of 5, 10, 15, 20, 25, and 30 dB using the NOISEX database. Mean PESQ scores of noisy speech and MWF and EMWF outputs were computed for each noise type. Figure 10 demonstrates the results. As can be seen in Figure 10, the performance improvement of EMWF is more explicit at low SNR levels.

Finally, we performed another experiment to compare the subjective quality measures of MWF and EMWF outputs. As indicated in [27], MOS tests are commonly used by the speech processing community for both evaluating the effectiveness of speech coding algorithms and assessing the quality of synthesized speech. Since we needed a subjective quality comparison of MWF and EMWF outputs, we performed MOS tests on a test database that comprised 10 utterances collected in a car driven in traffic. In the MOS test, 15 subjects (3 females) were used. The subjects listened the original noisy samples and filtered samples (MWF and EMWF

Table 4. PESQ scores of noisy and enhanced speech signals for white Gaussian noise with 10 dB SNR.

Sample	Noisy speech	Enhanced speech with MWF	Enhanced speech with EMWF
Sample 1	2.224	2.271	2.285
Sample 2	2.294	2.463	2.589
Sample 3	2.297	2.364	2.391
Sample 4	1.730	2.241	2.254
Sample 5	2.568	2.530	2.862
Sample 6	1.897	2.015	2.033
Sample 7	2.275	2.342	2.440
Sample 8	2.015	2.371	2.435
Sample 9	2.145	2.412	2.498
Sample 10	2.346	2.246	2.630
Sample 11	2.033	2.121	2.221
Sample 12	2.105	2.310	2.359
Sample 13	1.975	1.948	2.373
Sample 14	2.408	2.303	2.349
Sample 15	2.357	2.376	2.504
Sample 16	2.096	2.207	2.202
Sample 17	2.182	2.154	2.203
Sample 18	2.294	2.489	2.584
Sample 19	1.928	2.299	2.389
Sample 20	2.083	2.067	2.393
Mean	2.163	2.276	2.400

outputs) using headphones. They were instructed to rate the sentences on a scale of 1–5, where 1 is very poor and 5 is excellent. Some speech samples of speech coders having different MOS scores were presented to the subjects to ensure consistency in evaluating the speech quality. Average MOS scores per subject are shown in Table 5. As can be seen in the results, the average MOS score was 2.53 for noisy speech, 2.96 for the MWF method, and 2.99 for the EMWF method. This indicates that both algorithms increase the subjective speech quality and EMWF achieves a slightly better performance compared to the original MWF.

4.4. Unified system evaluations

Speech recognition plays an important role in various health care, military, human-computer interaction, automated documentation applications etc. The voice-controlled prosthetic hand proposed in [28] can be considered a recent example where speech recognition is used to enable human–machine interaction for health care purposes. Due to its wide range of applicability, we wanted to investigate the performance of our proposed unified system as a front-end to a speech recognition engine. Speech recognition performances were measured for both actual and artificially noise added noisy speech samples. For actual noise performance comparison, the same 10 recordings of actual noisy speech data in Turkish, which we used in Section 4.2, were utilized. For artificially added noise performance comparison, 30 different TIMIT database sentences were grouped into 24 different sets, where each set was degraded with a certain type of noise at a certain SNR level. The NOISEX database was used to add speech babble, car, pink, and white noise types to the samples at SNR levels of 5, 10, 15, 20, 25, and 30 dB. Noisy signals were first processed separately by the proposed VAD algorithm and the unified system in order to eliminate the silence regions. The processed signals were then presented as inputs to the speech recognizer.

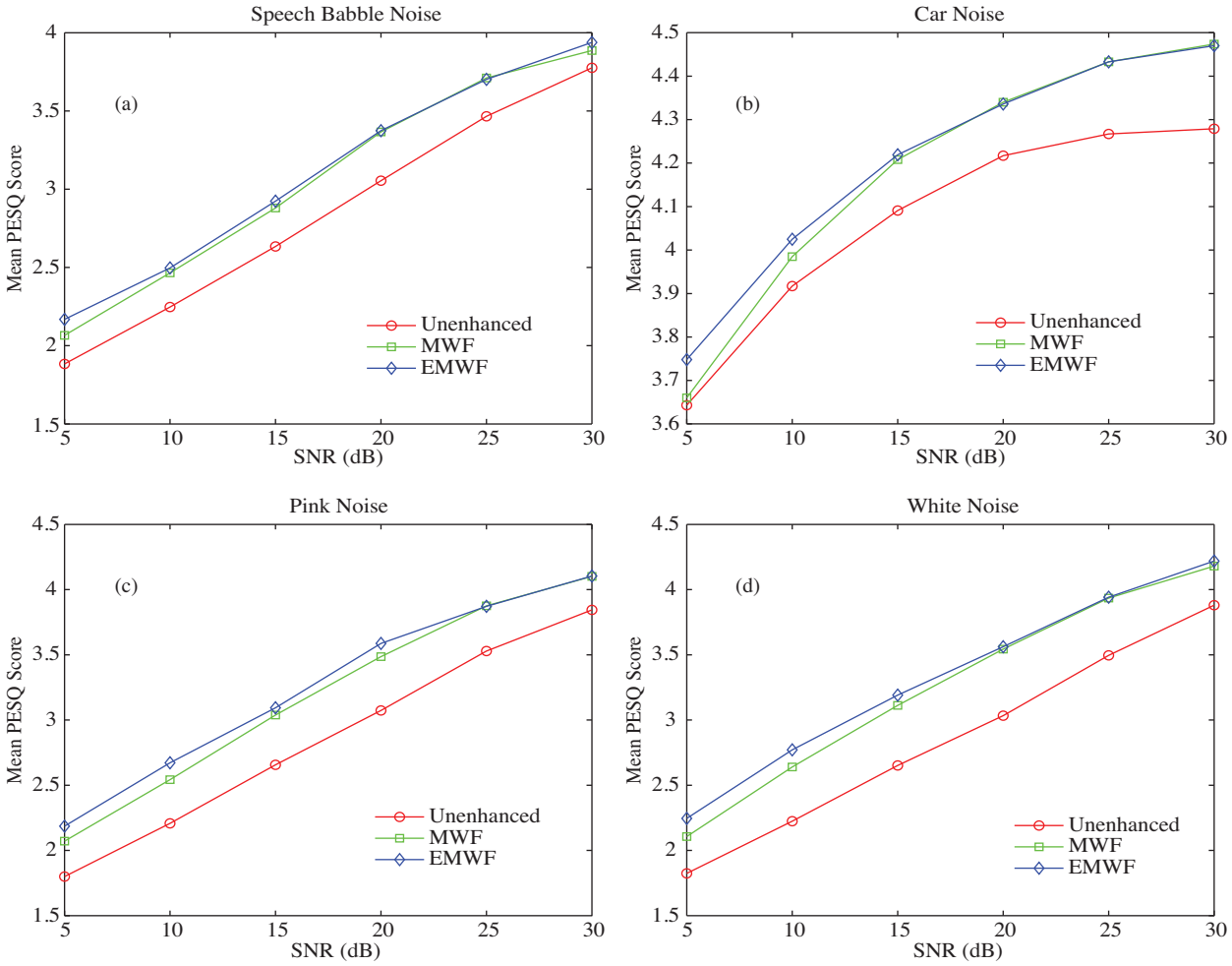

Figure 10. PESQ performances of MWF and EMWF outputs compared to unenhanced speech with (a) speech babble noise, (b) car noise, (c) pink noise, (d) white noise.

Tables 6 and 7 demonstrate the increased recognition performance of the unified system compared to the proposed VAD system. Recognition performance increased from 65.2% to 83.2% for actual noisy samples and from 87.5% to 91.3% for artificially noise added samples. As can be seen from the results, the unified system offers significant performance increase in speech recognition at low SNR levels due to the speech enhancement capability embedded in the system.

5. Discussion and conclusion

In this work, we follow a unified approach for VAD and speech enhancement problems. We demonstrate that the 2 problems are interrelated and implement a unified system for VAD and speech enhancement. The hybrid VAD and EMWF algorithms constitute the VAD and speech enhancement blocks of the proposed unified system, respectively.

A new and robust VAD algorithm is implemented to be used for the proposed unified system. The newly proposed VAD algorithm uses a periodicity measure and an energy measure that is computed according to the spectral properties of human speech. The new algorithm associates a soft decision value, rather than a strict speech/nonspeech decision, with each frame to indicate the speech likeliness of that frame. The final

Table 5. Average MOS scores on a scale from 1 to 5 over 10 utterances recorded in a car driven in traffic for 3 different conditions: (i) Original noisy speech, (ii) enhanced speech using MWF, and (iii) enhanced speech using EMWF.

Listener	Noisy speech	Enhanced speech with MWF	Enhanced speech with EMWF
Subject 1	2.56	2.97	3.05
Subject 2	2.00	2.34	2.39
Subject 3	2.91	3.39	3.43
Subject 4	2.63	3.07	3.09
Subject 5	2.46	2.91	2.91
Subject 6	2.91	3.36	3.43
Subject 7	2.74	3.17	3.22
Subject 8	2.43	2.87	2.87
Subject 9	2.17	2.59	2.61
Subject 10	3.07	3.54	3.55
Subject 11	2.51	2.93	2.97
Subject 12	2.13	2.46	2.49
Subject 13	2.65	3.08	3.11
Subject 14	2.85	3.30	3.35
Subject 15	1.87	2.35	2.40
Mean	2.53	2.96	2.99
St. Dev.	0.36	0.38	0.38

Table 6. Unified system evaluations for samples in Turkish (actual noise).

Sample	Recognition rate in Recognition rate	Recognition rate Recognition rate
Sample 1	40.1%	71.4%
Sample 2	91.3%	100.0%
Sample 3	93.6%	95.7%
Sample 4	76.1%	100.0%
Sample 5	94.1%	98.0%
Sample 6	91.3%	100.0%
Sample 7	23.9%	54.5%
Sample 8	65.9%	79.5%
Sample 9	90.9%	95.5%
Sample 10	93.6%	95.7%
Total	65.2%	83.2%

speech/nonspeech decision is based on a history of soft decision values. Employing speech enhancement in the form of MWF is shown to improve the performance of the proposed VAD algorithm. Utilization of speech enhancement in the hybrid VAD algorithm not only decreased the average error but also increased the utterance detection rate while decreasing the false alarm rate. Benefits of utilizing VAD for speech enhancement are demonstrated by implementing the EMWF algorithm. The EMWF algorithm results in less spectral distortion of noise power spectrum estimate compared to the standard MWF algorithm. Increased noise power spectrum estimation accuracy of EMWF relative to MWF enables the EMWF algorithm to employ a more aggressive enhancement at nonspeech intervals, and a more mild filtering at the speech segments compared to MWF. This provides better speech enhancement. For comparison, the EMWF algorithm is tested under simulated and actual car noise conditions and is shown to outperform the standard MWF in both subjective and objective speech quality evaluations.

Table 7. Unified system evaluations for samples in English (artificially added noise).

Sample set	Noise characteristics	Recognition rate in standard VAD (%)	Recognition rate in unified system (%)
Set 1	speech babble, 5 dB SNR	73.3%	90.0%
Set 2	speech babble, 10 dB SNR	100.0%	93.3%
Set 3	speech babble, 15 dB SNR	100.0%	90.0%
Set 4	speech babble, 20 dB SNR	100.0%	86.7%
Set 5	speech babble, 25 dB SNR	100.0%	90.0%
Set 6	speech babble, 30 dB SNR	96.7%	90.0%
Set 7	car noise, 5 dB SNR	100.0%	100.0%
Set 8	car noise, 10 dB SNR	100.0%	96.7%
Set 9	car noise, 15 dB SNR	100.0%	96.7%
Set 10	car noise, 20 dB SNR	96.7%	96.7%
Set 11	car noise, 25 dB SNR	96.7%	96.7%
Set 12	car noise, 30 dB SNR	96.7%	96.7%
Set 13	pink noise, 5 dB SNR	6.7%	70.0%
Set 14	pink noise, 10 dB SNR	73.3%	83.3%
Set 15	pink noise, 15 dB SNR	96.7%	90.0%
Set 16	pink noise, 20 dB SNR	100.0%	96.7%
Set 17	pink noise, 25 dB SNR	100.0%	96.7%
Set 18	pink noise, 30 dB SNR	100.0%	100.0%
Set 19	white noise, 5 dB SNR	3.3%	56.7%
Set 20	white noise, 10 dB SNR	63.3%	90.0%
Set 21	white noise, 15 dB SNR	96.7%	90.0%
Set 22	white noise, 20 dB SNR	100.0%	93.3%
Set 23	white noise, 25 dB SNR	100.0%	100.0%
Set 24	white noise, 30 dB SNR	100.0%	100.0%
Total		87.5%	91.3%

Finally, the unified system is evaluated as a preprocessor to a speech recognition engine where actual noisy and artificially noise added signals are used in the experiment. Compared to the single VAD system, the usage of the unified system enables performance increase in speech recognition rates, especially at low SNR levels.

References

[1] J.H. Chang, N.S. Kim, S.K. Mitra, "Voice Activity Detection Based on Multiple Statistical Models", IEEE Transactions on Signal Processing, vol. 54, no. 6, pp. 1965–1976, 2006.

[2] L. Rabiner, M. Sambur, "Voiced-unvoiced-silence detection using the Itakura LPC distance measure", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 323–326, 1977.

[3] J.D. Hoyt, H. Wechsler, "Detection of human speech in structured noise", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 19–22, 1994.

[4] J.A. Haigh, J.S. Mason, "Robust voice activity detection using cepstral features", Proceedings of the IEEE Conference on Computer, Communication, Control and Power Engineering, vol. 3, pp. 321–324, 1993.

[5] R. Tucker, "Voice activity detection using a periodicity measure", Proceedings of the IEE Conference on Communications, Speech and Vision, vol. 139, no. 4, pp. 377–380, 1992.

[6] J. Sohn, W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 365–368, 1998.

[7] L.M. Arslan, "Modified Wiener Filtering", Signal Processing, vol. 86, no. 2, pp. 267–272, 2006.

[8] C.P. Loizou, Speech Enhancement: Theory and Practice, Boca Raton, FL, CRC Press Inc., 2007.

[9] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of speech corrupted by acoustic noise", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 4, pp. 208–211, 1979.

[10] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 27, no. 2, pp. 113–120, 1979.

[11] R. McAulay, M. Malpass, "Speech enhancement using a soft-decision noise suppression filter", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28, no. 2, pp. 137–145, 1980.

[12] L.M. Arslan, J.H.L. Hansen, "Minimum cost based phoneme class detection for improved iterative speech enhancement", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 45–48, 1994.

[13] Y. Ephraim, D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 32, no. 6, pp. 1109–1121, 1984.

[14] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 33, no. 2, pp. 443–445, 1985.

[15] J.H.L. Hansen, M.A. Clements, "Constrained iterative speech enhancement with application to speech recognition", IEEE Transactions on Signal Processing, vol. 39, no. 4, pp. 795–805, 1991.

[16] J.H.L. Hansen, L.M. Arslan, "Markov model based phoneme class partitioning for improved constrained iterative speech enhancement", IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, pp. 98–104, 1995.

[17] P. Scalart, J.V. Filho, "Speech enhancement based on a priori signal to noise estimation", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, no. 2, pp. 629–632, 1996.

[18] M. Dendrinos, S. Bakamidis, G. Carayannis, "Speech enhancement from noise: A regenerative approach", Speech Communication, vol. 10, no. 1, pp. 45–57, 1991.

[19] Y. Ephraim, H.L.V. Trees, "A signal subspace approach for speech enhancement", IEEE Transactions on Speech and Audio Processing, vol. 3, no. 4, pp. 251–266, 1995.

[20] L.M. Arslan, Ph.D. Thesis, Duke University, 1996.

[21] G.E. Peterson, H.L. Barney, "Control Methods Used in a Study of the Vowels", The Journal of the Acoustical Society of America, vol. 24, no. 2, pp. 175–184, 1952.

[22] D.A. Schwartz, C.Q. Howe, D. Purves, "The Statistical Structure of Human Speech Sounds Predicts Musical Universals", The Journal of Neuroscience, vol. 23, no. 18, pp. 7160–7168, 2003.

[23] L. Rabiner, R. Schafer, Digital Processing of Speech Signals, New Jersey, Prentice-Hall Inc., 1978.

[24] J. Makhoul, "Linear Prediction: A Tutorial Review", Proceedings of IEEE, vol. 63, no. 4, pp. 561–580, 1975.

[25] ITU, "A silence compression scheme for G.729 optimized for terminals conforming to ITU-T V.70. ITU-T Rec. G. 729, Annex B", 1996.

[26] ETSI, "Voice activity detector (VAD) for adaptive multi-rate (AMR) speech traffic channels, ETSI EN 301 708 Recommendation", 1999.

[27] H. Sak, T. Güngör, Y. Safkan, "A Corpus-Based Concatenative Speech Synthesis System for Turkish", Turkish Journal of Electrical Engineering and Computer Sciences, vol. 14, no. 2, pp. 209–223, 2006.

[28] M.H. Asyalı, M. Yılmaz, M. Tokmakçı, K. Sedef, B.H. Aksebzeci, R. Mittal, "Design and implementation of a voice-controlled prosthetic hand", Turkish Journal of Electrical Engineering and Computer Sciences, vol. 19, no. 1, pp. 33–46, 2011.