EURASIP Journal on Applied Signal Processing 2005:4, 487–497

A Computationally Efficient Mel-Filter Bank VAD Algorithm for Distributed Speech Recognition Systems

Damjan Vlaj
Institute of Electronics, Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova 17, 2000 Maribor, Slovenia
Email: damjan.vlaj@uni-mb.si

Bojan Kotnik
Institute of Electronics, Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova 17, 2000 Maribor, Slovenia
Email: bojan.kotnik@uni-mb.si

Bogomir Horvat
Institute of Electronics, Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova 17, 2000 Maribor, Slovenia
Email: bogo.horvat@uni-mb.si

Zdravko Kačič
Institute of Electronics, Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova 17, 2000 Maribor, Slovenia
Email: kacic@uni-mb.si

Received 18 March 2004; Revised 23 September 2004; Recommended for Publication by Douglas O'Shaughnessy

This paper presents a novel computationally efficient voice activity detection (VAD) algorithm and emphasizes the importance of such algorithms in distributed speech recognition (DSR) systems. When using VAD algorithms in telecommunication systems, the required capacity of the speech transmission channel can be reduced if only the speech parts of the signal are transmitted. A similar objective can be adopted in DSR systems, where the nonspeech parameters are not sent over the transmission channel. A novel approach is proposed for VAD decisions based on mel-filter bank (MFB) outputs with the so-called Hangover criterion. Comparative tests are presented between the presented MFB VAD algorithm and three VAD algorithms used in the G.729, G.723.1, and DSR (advanced front-end) Standards. These tests were made on the Aurora 2 database, with different signal-to-noise (SNRs) ratios. In the speech recognition tests, the proposed MFB VAD outperformed all the three VAD algorithms used in the standards by 14.19% relative (G.723.1 VAD), by 12.84% relative (G.729 VAD), and by 4.17% relative (DSR VAD) in all SNRs.

Keywords and phrases: voice activity detection, distributed speech recognition, telecommunication systems.

1. INTRODUCTION

Voice activity detection (VAD) is an algorithm that is able to distinguish speech from noise and is an integral part of a variety of speech communication systems, such as speech recognition, speech coding, hands-free telephony, and audio conferencing. An effective VAD algorithm plays an important role in telecommunication systems, especially in automatic speech recognition (ASR) systems. This paper presents the use and the importance of VAD algorithms in ASR systems. ASR systems can work under various noisy conditions (e.g., in a restaurant, a train station, etc.). An input signal often contains many nonspeech parts, which can reduce the speech recognition performance of ASR systems. This is especially true when the ASR system operates under adverse conditions. A major cause of errors in ASR systems is incorrect detection at the beginning and ending boundaries of the test, and the reference patterns [1]. Correct determination of the endpoints is fairly easy if the signal-to-noise ratio (SNR) is high (e.g., greater than 35 dB). Unfortunately, the majority of applicable recognizers have to work with a much smaller SNR (typically between 25 dB and 15 dB, and as low as 5 dB). Under such conditions, it becomes very difficult to detect weak fricatives, weak nasals, and low-amplitude voiced sounds occurring at the beginning or the end of the utterances [1].

The majority of existing algorithms that try to distinguish between speech and noise perform decisions based on frames of various frame and shift length. The input signal is divided into overlapping frames where the signal is supposed to be stationary within one frame. The VAD algorithm makes constantly a decision, if the certain frame of the windowed input signal contains noise only or there is also presence of speech. The output decision of the VAD can be 1 (the frame of the input signal contains speech or noisy speech) or 0 (the frame of the input signal contains noise only).

VAD is very important in speech communication and telecommunication systems. VAD and two other algorithms, discontinuous transmission (DTX) and comfort noise generator (CNG) [2, 3], are used in telecommunication systems to reduce transmission rate during the silent periods of the input signal. When DTX is in operation, the transmitter is switched off if no speech is present. This increases the system capacity by reducing the cochannel interface and also reduces the transmitter power consumption (an important consideration for mobile phones). During a typical conversation, each speaker talks for less than 40% of the time, and it has been estimated that DTX with VAD decision could approximately double the transmission channels capacity [4]. At the same time, VAD sorts out the frames without speech, so the noise suppressor can use this information to estimate noise statistics. A similar objective can be used in distributed speech recognition (DSR) system, where the nonspeech parameters are not sent over the transmission channel. This work proposes a novel procedure for a VAD algorithm based on mel-filter bank (MFB) outputs.

A lot of existing VAD algorithms use features that depend on energy computation. Some algorithms use a combination of zero-crossing rate and energy [5], and the others a distance measure of the cepstral features [6]. More complex algorithms use more than one feature to detect speech. One of these is described in [7] where the entropy for the low-, high-, and full-frequency bands is calculated. Entropy-based VAD is also presented in [8]. A statistical-model-based VAD [9, 10] is also one of the algorithms used for VAD. The new DSR Standard (advanced front end) [11] uses two VAD algorithms. Both are shortly presented in Section 3.3.

In the following section, we give short descriptions of the DSR system. Section 3 presents four VAD algorithms: three VAD algorithms used in the G.729, G.723.1, and DSR (advanced front-end) Standards and a novel approach to VAD decisions based on the MFB outputs with the so-called Hangover criterion. We conducted comparative tests between the proposed VAD (MFB VAD) algorithm and the three VAD algorithms used in the G.729, G.723.1, and DSR (advanced front-end) Standards. Section 4 presents the tests, which were made on the Aurora 2 database [12] using different SNRs to ascertain how robust the tested VAD algorithms are to the background noise, and to perform comparative analysis. In the same section, we present the speech recognition performance results using a frame dropping strategy, when different VAD algorithms were used. The results are presented in Section 5 and discussed in Section 6. Conclusions are given in Section 7.

2. DSR SYSTEM

The use of data processing in modern telecommunication technologies such as speech recognition has been increasing in recent years. The growth of speech and Internet technology, and speech recognition over IP networks [13] has made the field of speech recognition research more extensive. Data processing is now being deployed in the area of increased mobile telephony, where people want to be able to access information while they are on the move. This also applies in the mobile wireless world [14]. To access these data services, small portable multimodal devices (e.g., personal digital assistant (PDA), personal navigator) will be used, which require improved user interfaces using speech input [15]. Since 1995, several researches have been made in the field of DSR systems [14, 16]. The DSR system involves many diversified technologies including ASR, network data transmission bandwidth and protocol, data compression, distributed computing, and others [16].

Whereas centralized servers are able to share the computational burden between different users and, therefore, support an easy development of the provided technologies and services, mobile voice networks, however, can degrade the speech recognition performance obtained from central recognizers. This is due to low bit rate coding and speech transmission errors. The idea of a DSR system is to eliminate the impact of the transmission channel on the transmitted speech by using an error-protected data channel to send a parameterized speech representation which is suitable for speech recognition. The DSR system works in the following way (Figure 1): the recognizer's front end, which is located in the terminal, is connected over the wireless data channel to a remote back-end server. The feature parameter extraction of the spoken speech is performed on the terminal. This is called the "front end" of the speech recognition system. The features are then compressed and transmitted to a remote "back-end" recognizer over a data channel together with error protection and correction data (cyclic redundancy check (CRC)). In this way, both channel invariability and minimal impact of the transmission channel on recognition performance are achieved [14].

If these applications using DSR are to be marketed, a standard for the front end is required so as to ensure compatibility between the terminal and the remote recognizer. Over the last four years, the Aurora DSR Working Group within European Telecommunications Standards Institute (ETSI) has been developing this standard. In order to evaluate alternative proposals for the front-end feature extraction algorithm, a reference Aurora 2 database [12] and an experimental framework have been established. This database, which is based on the original TI-Digits database with controlled filtering and simulated noise addition over a range of SNRs from 20 dB to −5 dB, has been made publicly available via the European Language Resources Association (ELRA).

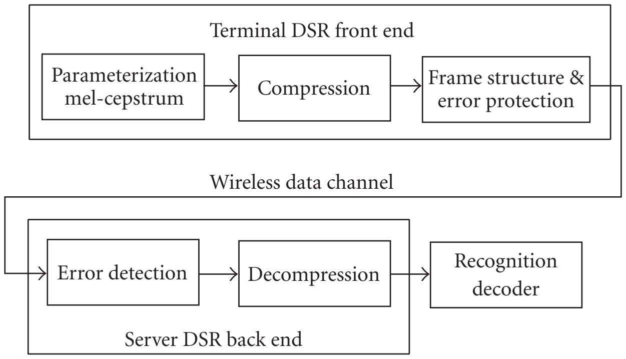


Figure 1: Block diagram of the DSR system.

The feature extraction algorithms that operate in DSR are expected to produce a speech feature representation invariant to noisy environments and SNRs if the equivalent utterance is spoken under each of these different conditions. Furthermore, they are expected to operate in real-time and with the lowest possible computational costs.

3. VAD ALGORITHMS

In the next four sections, the three VAD algorithms used in the Standards G.729, G.723.1, and DSR (advanced front end) are briefly described and the proposed MFB-output-based VAD algorithm is presented. All the described algorithms are frame-based. The VAD algorithm makes a voice activity decision every 10 milliseconds using the G.729, DSR (advanced front end), and proposed MFB VAD. When using the G.723.1 VAD algorithm, the decision is made every 30 milliseconds. The frame decision is divided into three decision frames for the G.723.1 VAD algorithm with the same information. In this way, the algorithms can be easily compared.

3.1. G.729 VAD algorithm

The basic parameters for the VAD algorithm used for the G.729 Standard are full- and low-band frame energy, a set of line spectral frequencies and the frame zero-crossing rate [2]. In the first stage, all four parametric features are extracted from the input signal. If the frame number is less than 32, an initialization stage for the long-term averages takes place, and the VAD decision is set to 1 if the frame energy from the linear predictive coding (LPC) analysis is above 15 dB. Otherwise, the VAD decision is forced to 0. At the next stage, a set of difference parameters are calculated. This set is generated as a difference measure between the current frame parameters and the running averages of the background noise characteristics. Four different measures are calculated: spectral distortion, energy difference, low-band energy difference, and zero-crossing difference. The initial VAD decision is made at the next stage, using multiboundary decision regions in the space of four difference measures [2]. The running averages have to be updated only in the presence of background noise. An adaptive threshold is tested, and the update takes place only if the threshold criterion is met.

3.2. G.723.1 VAD algorithm

Energy measure is the basic parameter for the VAD algorithm used for the G.723.1 Standard [3]. The energy of the inverse filtered signal is compared with the threshold. Speech is indicated whenever the threshold is exceeded. The threshold is computed using a two-step procedure. Firstly, the noise level is updated based on its previous value and the energy of the filtered signal. Secondly, the threshold is computed from the noise level via a logarithmic approximation [3]. A Hangover of six frames is added only in the case of speech bursts (VAD decision is 1) larger or equal to two frames [3].

3.3. DSR (advanced front-end) VAD algorithm

The DSR Standard (advanced front end) [11] uses two VAD algorithms. The first one is used for noise estimation and is based on energy computation. The second one is used for nonspeech frame dropping and has two stages: the first is a frame-based detection stage consisting of three measurements, and the second is the decision stage in which measurements are analyzed for speech likelihood. The final decision from the second stage is applied retrospectively to the earliest frame in the buffer [11]. The second VAD algorithm was used as a comparative test in this paper.

3.4. Proposed MFB VAD algorithm

The proposed MFB VAD algorithm is based on MFB outputs. The VAD algorithm is designed to be used in speech recognition systems. MFB outputs, which are part of the feature extraction module, were used to reduce unnecessary computational costs for VAD decision. The described VAD algorithm classifies frame m as speech $\sigma[m] = 1$ (speech mixed with noise) or nonspeech $\sigma[m] = 0$ (noise only) by comparing the SNR of the current frame to the threshold. The SNR of the current frame corresponds to the difference between the short-term and the long-term spectral energy estimates. The long-term estimate is updated when the VAD decides that the current frame corresponds to the noise only and the MFB output of the current frame is used as a short-term estimate. In the first step, an estimate of the MFB outputs' short-term energy $E_{\text{est}}[m]$ is calculated for the first ten frames as

$$E_{\text{est}}[m] = \begin{cases} \ln \sum_{i=1}^{N} f_{\text{bank}}[m,i], & m = 1, \\ \dfrac{E_{\text{est}}[m-1] + \ln \sum_{i=1}^{N} f_{\text{bank}}[m,i]}{2}, & 1 < m \leq 10, \end{cases} \tag{1}$$

where $N = 23$ represents the number of channels (MFB outputs) and $f_{\text{bank}}[m,i]$ is the ith MFB output of the mth frame. The estimates of short-term energy $E_{\text{est}}[m]$ are used to define the weighting factor q of the input signal:

$$q = \begin{cases} 32, & E_{\text{est}}[m] \leq \dfrac{6}{9} \cdot \text{MAX}[N], \\ 64, & \dfrac{6}{9} \cdot \text{MAX}[N] < E_{\text{est}}[m] < \dfrac{7}{9} \cdot \text{MAX}[N], \\ 128, & E_{\text{est}}[m] \geq \dfrac{7}{9} \cdot \text{MAX}[N], \end{cases} \tag{2}$$

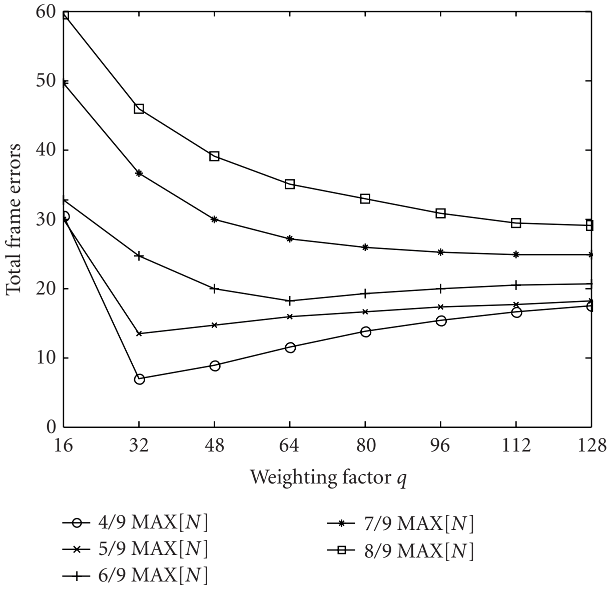


Figure 2: Optimal definition of the weighting factor q.

where MAX$[N]$ represents the logarithmic sum of the MFB outputs' maximum possible values in the frequency spectrum. The different values for the weighting factor q were defined by considering the fact that the lower q was more suitable for the higher SNRs and vice versa. The interdependence between the total frame errors and weighting factor q can be seen on Figure 2. When studying the weighting factor q, we obtained the best results for the clean signal when q was set to 32. For a very noisy signal (SNR $= -5$ dB), the best results were achieved when q was set to 128. If q was the same as for the clean signal, the frame detection for a very noisy signal would be incorrect, as the majority of frames which include speech would be recognized as noise. The values of the short-term energy estimates $E_{\text{est}}[m]$ are calculated for the first ten frames (1) and for every frame m when the VAD algorithm classifies it as nonspeech ($\sigma[m] = 0$). By changing the short-term energy estimates $E_{\text{est}}[m]$, the weighting factor q changes, respectively (see (2)). The weighting factor is therefore adjusted to various values of SNRs. The values of the short-term energy estimates $E_{\text{est}}[m]$ in the logarithmic domain were between four ninths of the maximum value for the silent part of the clean signal and eight ninths of the maximum value for the silent part of the noisy signal (SNR $= -5$ dB). We defined the limits for optimal definition of the weighting factor q in (2) based on this and the former facts. For the mth frame, the MAX$[N]$ is calculated as

$$\text{MAX}[N] = \ln\left(\sum_{k=1}^{N} \frac{\text{cbin}_{k+1} - \text{cbin}_{k-1} + 2}{2} \cdot Y_{\text{MAX}}[m,k]\right), \tag{3}$$

where N is the number of MFB outputs and $Y_{\text{MAX}}[m,k]$ is the maximum possible value of the kth bin of the frequency spectrum in the mth frame. The maximum possible value of the frequency spectrum is $2^Q/2$, where Q represents the resolution of the input signal quantization. In our case, the quantization resolution is 16. The calculation of the channels' centre frequencies (cbin_k for the kth channel) is described in more detail in [17]. The full-frequency band is divided into N channels equidistant in the mel-frequency domain. Each channel has a triangular-shaped frequency window. Consecutive channels are half overlapping.

We can now calculate the weighted short-term MFB output energy $E_f[m]$ for each frame, which is used in the VAD algorithm,

$$E_f[m] = q \cdot \ln\left(1 + \frac{\sum_{i=1}^{N} f_{\text{bank}}[m,i]}{w}\right), \tag{4}$$

where $N = 23$ represents a number of channels (MFB outputs) and $f_{\text{bank}}[m,i]$ the MFB output of the ith MBF output and mth frame. The weighting factor q is used to increase the slope of the $E_f[m]$ function in the logarithmic domain. The constant w is set to 1000 and is used to change the shape of the logarithmic function. The influence of the constant w in the logarithmic domain is presented in Figure 3. Figure 3a shows that we can reduce the total frame errors with the constant w, especially for the clean condition. The constant w is set to 1000, which gives us optimal attenuation for the sum of MFB outputs (4) against all other w constants, which are also presented in Figure 3b. The attenuation of the MFB outputs is too big for the bigger constant, w and the attenuation is too small for the lower constant w. There is no improvement in the reduction of the level of frame errors. If we choose too big or too small a constant w, the average of the total frame errors will increase. The analysis of the constant w shows that the chosen constant gave us the best results (Figure 3a). The short-term MFB bank output energy of the current frame $E_f[m]$ is used in the update of the long-term mean energy $E_m[m]$ as follows.

If

$$\left(E_f[m] - E_m[m-1]\right) < \textit{EenergyUpdate}, \tag{5}$$

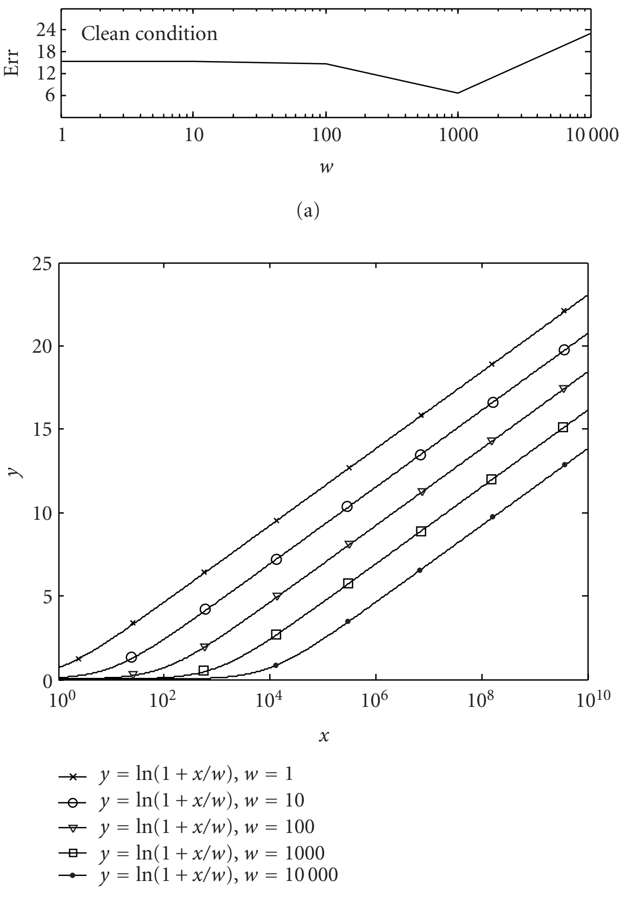
then

$$E_m[m] = E_m[m-1] + \frac{E_f[m] - E_m[m-1]}{\textit{EnergyReduction}}; \tag{6}$$

else,

$$E_m[m] = E_m[m-1], \tag{7}$$

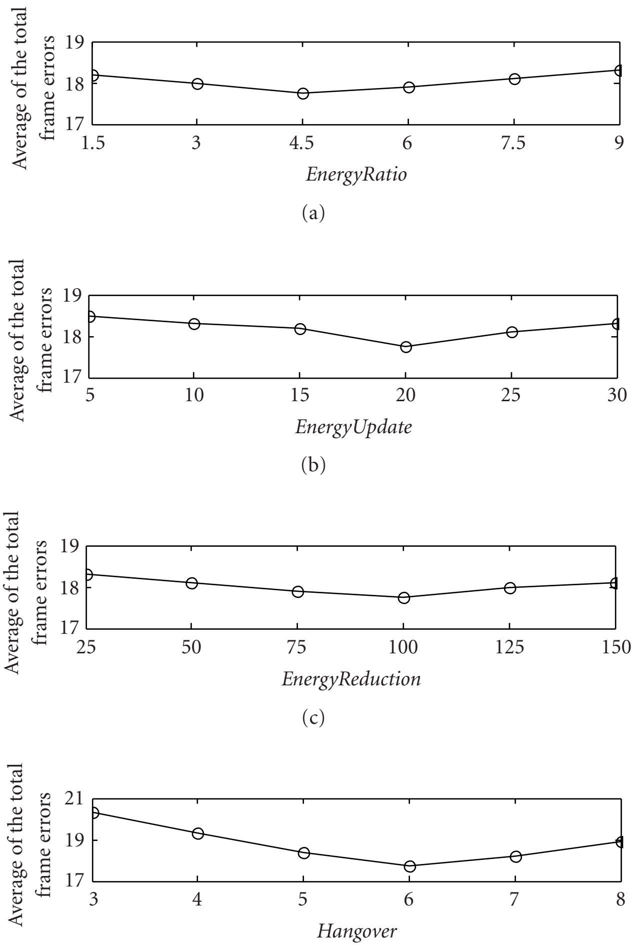
where the *EnergyUpdate* constant is set to 20 and the *EnergyReduction* constant to 100. The *EnergyReduction* constant is set to modify long-term mean energy $E_m[m]$. If the *EnergyReduction* is smaller than 100, the modification is bigger and vice versa. *EnergyReduction* was set to 100 after several analyses. The results of the analysis can be seen in Figure 4c. The VAD decision procedure of the current frame can begin



(a)

(b)

FIGURE 3: (a) The influence of the constant w on the total number of frame errors and (b) influence of the constant w on the sum of the MFB outputs in the logarithmic domain ($x = \Sigma_{i=1}^{N} f_{\text{bank}}[m, i]$).



(a)

(b)

(c)

(d)

FIGURE 4: Definition of the *EnergyRatio*, *EnergyUpdate*, *EnergyReduction*, and *Hangover* constants to achieve lower total frame errors.

after initial calculations on the first frame to determine the short-term spectral energy $E_f[m]$ and the long-term mean spectral energy $E_m[m]$. Figure 5 shows a flowchart of the proposed MFB output-based VAD algorithm. It should be noted here, that before the VAD processing of the input signal is started, the frame counters *Hangover* and *SpeechFrame*, as well as the long-term mean spectral energy $E_m[m - 1]$, are initialized to 0. The constant *EnergyRatio* is set to optimal value 4.5 (see Figure 4a). The constants *EnergyUpdate* and *EnergyRatio* were set together. Between both values, the long-term mean energy $E_m[m]$ is updated and the frame is declared as speech. When the difference is smaller than 4.5, only the long-term mean energy $E_m[m]$ is updated ($E_f[m] - E_m[m - 1] < 4.5$) and when the difference is bigger than 20, the frame is declared as speech without long-term mean energy $E_m[m]$ update ($E_f[m] - E_m[m - 1] > 20$). VAD also uses the so-called Hangover criterion with the *Hangover* factor, which prevents a misclassification of weak fricatives to noise at the end of the speech segments. Seven consecutive frames (experimentally defined—see Figure 4d) at the end of at least a four-frame-long speech segment are also declared as speech. If the speech segment is shorter than four frames then the Hangover criterion is unused. When the Hangover criterion is used, the seven consecutive frames at the end of the speech segment are also declared as speech when the conditions shown in the flowchart in Figure 5 are fulfilled (*Hangover* is set to 6). The Hangover criterion for the VAD decision was also used in [3, 11].

4. DESCRIPTION OF THE EXPERIMENTAL FRAMEWORK

Experiments were made using the Aurora 2 database [12], which is designed to evaluate the performance of speech recognition algorithms under noisy conditions. The Aurora 2 database is described in Section 4.1. The development of the VAD reference procedure on a frame-by-frame basis is presented in Section 4.2 on which basis we can compare all three tested algorithms to the same reference. The speech recognition experiments are presented in Section 4.3.

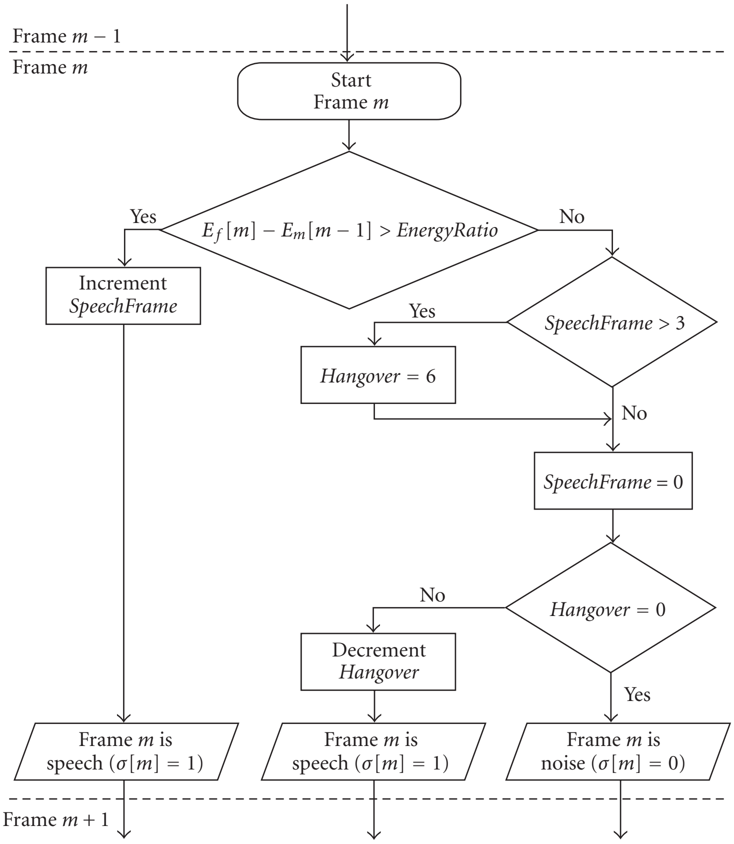


Figure 5: Flowchart of the MFB VAD.

4.1. Aurora 2 database

The concept of the Aurora 2 framework and experiments includes two training modes that are defined as training on clean data only, and as training on clean and noisy (multicondition) data. Owing to the fact that training on clean data only enables speech modeling without any noise distortions, such models are expected to be the best for representing all the available speech information. The weakness of these models is that they contain no information about possible distortions. This, however, is an advantage of multicondition training, where distorted speech signals are taken as training data.

For training on clean data, 8440 utterances were chosen, which contained the recordings of 55 male and 55 female adults. These signals were filtered using the G.712 characteristics [18] with no noise added. The same 8440 utterances were used for multicondition training. They were divided into 20 subsets, each of which included 422 utterances. There were several utterances from all the available speakers in each subset. The 20 subsets represented 4 different noise scenarios at 5 different SNRs. The noises were a suburban train, babble, a car, and an exhibition hall noise. The SNRs were 20 dB, 15 dB, 10 dB, 5 dB, and the clean condition. How the SNR was calculated is described in more detail in [12]. Speech and noise were filtered using the frequency response of the G.712 characteristic before mixing them with a noisy speech signal.

Three different test sets were defined [12]. Four subsets with 1001 utterances in each were obtained by splitting 4004 utterances from 52 male and 52 female speakers. The recordings of all speakers were present in each subset. Individual noise signals at SNRs of 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, and −5 dB were added and the clean case without adding noise was taken as the seventh condition. Speech and noise were, once again, filtered using the G.712 characteristic before addition.

The first test set was called test set A. In this test set, there are four noises, a suburban train, babble, a car, and an exhibition hall, were added to the 4 subsets. So, the set consisted of 28,028 utterances. There was a high match of training and test data, owing to the fact that this test set contained the same noises as used for the multicondition training.

The second test set was called test set B. This test was created in the same way, the only difference was that four different noises were used, which were a restaurant, a street, an airport, and a train station. In this case, a mismatch between training and test data also existed for multicondition training. This influenced the recognition accuracy when considering different noises other than those used for training.

The third test set was called test set C and it contained 2 out of 4 subsets with 1001 utterances each. Here, speech and noise were filtered using a Motorola Integrated Radio Systems (MIRS) characteristic [19], before being added to

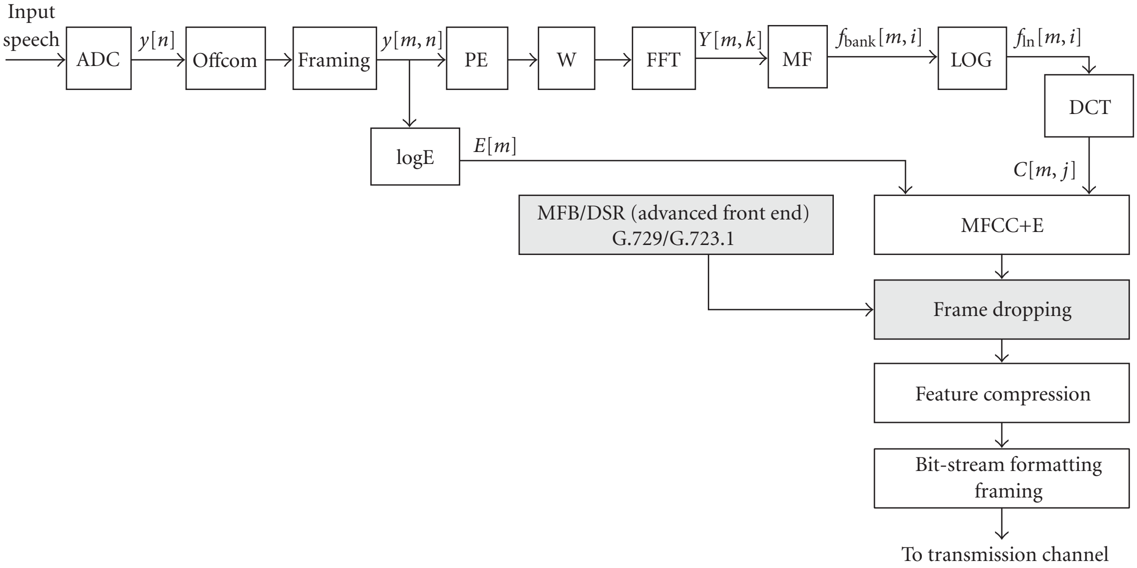


Figure 6: Block diagram of mel-cepstral DSR front-end Standard [17] with additionally integrated VAD decision and frame dropping strategy. (Sample index n, Frequency bin index k, and Cepstral coefficients index j). (ADC: analog-to-digital conversion; Offcom: offset compensation; PE: preemphasis; logE: energy measure computation; W: windowing; FFT: Fast Fourier transform (only magnitude components); LOG: nonlinear transformation; DCT: discrete cosine transform; and MFCC+E: Mel-frequency cepstral coefficients and energy.)

the SNRs of 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, and −5 dB. MIRS can be explained as a frequency characteristic that simulates the behavior of a telecommunication terminal, which meets the official requirements for the terminal input frequency response as specified, for example, in the ETSI-SMG (ETSI Special Mobile Group) Technical Specification [19]. The street and suburban train were used as added noise signals. The purpose of this set is to show the influence on recognition performance when a different frequency characteristic is present at the input of the recognizer.

4.2. Development of the frame-based reference for the VAD algorithm

The procedure for the frame-based reference for the VAD algorithm was carried out on 4004 clean speech data recordings, which were used in the test sets of the Aurora 2 database. The reference for the VAD algorithm was based on connected-digit speech recognition performed by the reference recognizer used in the Aurora 2 framework. The reference recognizer was based on Hidden Markov Models Toolkit (HTK) software [20]. The digits were modeled as whole word Hidden Markov Models (HMMs) with 16 emitting states per word with simple left-to-right models without skip-over states and 3 Gaussian mixtures per state. Two pause models were defined. The first one, called sil, consisted of 3 emitting states. The second pause model, called sp, was used to model pauses between words and consisted of a single emitting state, which was tied to the middle state of the first pause model. Six Gaussian mixtures for each state were used for the pause models.

Realignment of the training data was used to set the time boundaries of the speech utterances to create a frame-based reference for the VAD decision. The decision as to whether the frame contains a speech or silent part of the signal was set for every 10 milliseconds. The frame-based reference VAD algorithm was set in such a way that the result was best for the speech recognition process using the following procedure. If the input signal was recognized as speech within the time boundaries, the reference frames were set to 1. When the input signal was recognized as silence, the reference frames were set to 0. Comparative tests were made between the proposed MFB VAD algorithm and the three VAD algorithms used in the G.729, G.723.1, and DSR (advanced front-end) Standards. All tested VAD algorithms were compared to the frame-based reference for VAD algorithm. The results are shown in Section 5.

4.3. Speech recognition experiments

Speech recognition experiments were made on the Aurora 2 database with the ETSI standardized feature extraction module [17]. The standardized feature extraction algorithm with the added VAD algorithm and frame dropping block is shown in Figure 6. The VAD block makes the decision as to which frame should be sent over the wireless data channel to the back-end recognizer. The back-end recognizer is based on the HTK software [20]. Its structure is presented more in detail in [12]. The frame dropping block allows transmission of the SpeechFrames over the wireless data channel. At the same time, the frame dropping block prevents the transmission of noisy frames being sent over the wireless data channel. The frame dropping strategy was also used in [21, 22] to improve the speech recognition accuracy. The frame dropping strategy is important because, in this case, the speech recognizer does not need to deal with noisy frames and the result is faster and more accurate ASR.

Table 1: Percentage of frame errors obtained by the G.729 VAD algorithm.

	FEC	MSC	NDS	OVER	Total
Clean	4.52	7.47	0.42	0.43	12.84
SNR 20	0.44	4.73	13.83	5.53	24.53
SNR 15	0.30	5.28	14.42	6.13	26.13
SNR 10	0.31	6.27	14.43	6.37	27.38
SNR 5	0.35	7.75	14.43	6.60	29.13
SNR 0	0.43	10.72	14.39	6.69	32.23
SNR −5	0.60	13.53	14.40	6.68	35.21

Table 2: Percentage of frame errors obtained by the G.723.1 VAD algorithm.

	FEC	MSC	NDS	OVER	Total
Clean	4.67	0.35	5.25	9.18	19.45
SNR 20	3.88	0.51	7.73	9.19	21.31
SNR 15	2.43	0.51	10.18	10.17	23.29
SNR 10	0.78	0.30	12.57	10.79	24.44
SNR 5	0.10	0.08	13.27	12.85	26.30
SNR 0	0.00	0.02	13.38	13.16	26.56
SNR −5	0.00	0.00	13.38	13.20	28.58

Table 3: Percentage of frame errors obtained by the DSR (advanced front-end) VAD algorithm.

	FEC	MSC	NDS	OVER	Total
Clean	0.00	0.00	7.57	11.00	18.57
SNR 20	0.36	0.03	4.54	10.40	15.33
SNR 15	0.55	0.06	4.41	10.23	15.25
SNR 10	0.81	0.18	4.06	9.84	14.89
SNR 5	1.14	0.38	3.97	9.44	14.93
SNR 0	1.59	1.64	3.98	8.82	16.03
SNR −5	3.03	7.23	3.98	7.35	21.59

Table 4: Percentage of frame errors obtained by the MFB VAD algorithm.

	FEC	MSC	NDS	OVER	Total
Clean	3.28	1.83	0.58	1.23	6.92
SNR 20	1.88	3.28	5.96	4.27	15.39
SNR 15	1.86	4.45	6.78	4.61	17.70
SNR 10	1.83	5.94	7.47	4.88	20.12
SNR 5	1.86	7.23	8.13	5.53	22.75
SNR 0	1.94	9.94	8.32	5.96	26.16
SNR −5	2.36	14.26	8.52	5.95	31.09

5. RESULTS

5.1. Results on tested VAD algorithms

The tests based on frame errors were made on an MFB VAD algorithm and three VAD algorithms that are used in G.729, G.723.1, and DSR (advanced front-end) Standards. The statistics were obtained on [4] and are presented in Tables 1, 2, 3, and 4 as follows:

(i) clipping rate at the front of the speech segment (FEC),
(ii) clipping in the middle of the speech segment (MSC),
(iii) noise detected as speech in the silent region (NDS),
(iv) quantity of time during which the output of the tested VAD is on, after the reference VAD has switched off (OVER),
(v) total number of all the frame errors (Total).

We have also analyzed if word clipping (WC) occurred, but none of the tested VAD algorithms performed this kind of error.

5.2. Speech recognition results

The speech recognition experiments were carried out on the Aurora 2 database with clean and multicondition training on all three subsets. Tables 5, 6, 7, and 8 present the absolute performances of the algorithms indicated in Figure 6 (G.729, G.723.1, DSR (advanced front-end) and MFB VAD algorithms) and relative performance to the mel-cepstral DSR front-end Standard [17].

6. DISCUSSION

The frame dropping strategy used for the ASR shows that speech recognition performance can be improved, without any other noise reduction technique, if an effective VAD algorithm with frame dropping strategy is used (Tables 5, 6, 7, and 8). Speech recognition accuracy is defined as the ratio between the difference between the number of correctly recognized words and the number of inserted words compared to the number of all in advance correctly defined words [20]. From this definition, speech recognition accuracy can be increased by reducing the number of inserted words, which can be achieved by the correctly dropped noisy frames. We made comparative tests using ASR on different VAD algorithms. In the speech recognition tests, the proposed MFB VAD outperformed all the three VAD algorithms used in the standards by 14.19% relative (G.723.1 VAD), by 12.84% relative (G.729 VAD), and by 4.17% relative (DSR VAD—advanced front end) in all SNRs. The percentages are calculated from Tables 5, 6, 7, and 8: Performance relative to mel-cepstrum. The percentage results are obtained from the difference between overall average values of the proposed MFB VAD algorithm and overall average values of the compared VAD algorithms. The speech recognition improvement is especially noticeable when clean data training is used for speech recognition (see Tables 5, 6, 7, and 8: "Clean only" training mode). The use of a frame dropping strategy is important because it prevents the speech recognizer from having to deal with noisy frames and the result is faster ASR at the back end of the DSR system. Another advantage of this strategy is that the noisy frames are not sent over the wireless data channel. For clean training, the relative improvement with the use of an MFB VAD algorithm and frame dropping strategy is 28.06%. The idea of multicondition training is to train HMMs under different noise conditions. HMMs are not trained well enough when using the frame dropping strategy, because the number of dropped frames is greater in the middle of the

Table 5: Speech recognition performance by the G.729 VAD algorithm.

Absolute performance				
Training mode	Set A	Set B	Set C	Overall
Multicondition	85.95	84.25	80.31	84.14
Clean only	67.68	66.88	67.34	67.29
Average	76.81	75.57	73.82	75.72
Performance relative to mel-cepstrum				
Training mode	Set A	Set B	Set C	Overall
Multicondition	−15.31%	−14.69%	−21.38%	−16.51%
Clean only	16.39%	25.15%	3.35%	18.09%
Average	0.54%	5.23%	−8.93%	0.79%

Table 6: Speech recognition performance by the G.723.1 VAD algorithm.

Absolute performance				
Training mode	Set A	Set B	Set C	Overall
Multicondition	87.82	86.36	84.06	86.48
Clean only	59.46	57.49	62.82	59.34
Average	73.64	71.92	73.44	72.91
Performance relative to mel-cepstrum				
Training mode	Set A	Set B	Set C	Overall
Multicondition	0.02%	0.67%	1.74%	0.69%
Clean only	−4.87%	3.93%	−9.82%	−1.81%
Average	−2.43%	2.30%	−4.04%	−0.56%

Table 7: Speech recognition performance by the DSR (advanced front-end) VAD algorithm.

Absolute performance				
Training mode	Set A	Set B	Set C	Overall
Multicondition	87.79	87.16	84.32	86.84
Clean only	66.87	65.90	65.88	66.29
Average	77.33	76.53	75.10	76.57
Performance relative to mel-cepstrum				
Training mode	Set A	Set B	Set C	Overall
Multicondition	−0.17%	6.45%	3.37%	3.34%
Clean only	14.31%	22.95%	−0.77%	15.58%
Average	7.07%	14.70%	1.30%	9.46%

Table 8: Speech recognition performance by the MFB VAD algorithm.

Absolute performance				
Training mode	Set A	Set B	Set C	Overall
Multicondition	87.64	87.26	81.62	86.28
Clean only	71.89	72.96	66.65	71.27
Average	79.76	80.11	74.13	78.77
Performance relative to mel-cepstrum				
Training mode	Set A	Set B	Set C	Overall
Multicondition	−1.48%	7.17%	−13.30%	−0.80%
Clean only	27.28%	38.90%	1.50%	28.06%
Average	12.90%	23.03%	−5.90%	13.63%

speech segment (MSC in Table 4) as under the clean training condition. In this case, it is better to use a frame attenuation strategy [23, 24]. However, this strategy does not have the two advantages mentioned in the case of frame dropping strategy [21, 22]. Speech recognition performance by multicondition training can be improved if the errors in clipping at the front of the speech (FEC) and clipping in the middle of the speech (MSC) are as few as possible. In this case, the HMMs can be trained well enough.

When compared to the VAD G.729 and G.723.1 Standards, the total number of frame errors is always smaller in the case of MFB VAD (Tables 1, 2, and 4). The only exception is the G.723.1 Standard, where the total number of frame errors is smaller when the SNR is −5 dB, because FEC and MSC frame errors are zero. When we compare MFB VAD algorithm (Table 4) to DSR (advanced front-end) VAD algorithm (Table 3), the percentage of Total errors is almost always smaller by the DSR VAD algorithm. But, as we can see from Tables 7 and 8, the speech recognition accuracy is better by MFB VAD algorithm if we compare overall average and overall speech recognition by clean condition training mode. The reason for this is the smaller number of inserted words when computing speech recognition accuracy. From this, we can conclude that the higher NDS and OVER frame errors increase the number of inserted words and reduce the speech recognition accuracy. The same conclusion can be set for the G.729 and G.723.1 VAD algorithms (Tables 1, 2, 5, and 6).

We can conclude from the results obtained for the G.723.1 VAD, that the VAD used for this standard is very effective for the SNRs 5 dB, 0 dB, and −5 dB. This is due to the fact that this VAD does not properly distinguish between noise and speech. On the contrary, it considers the whole information as speech, therefore the FEC and MSC frame error rates are decreasing when the SNR is being reduced. If we use a VAD algorithm for noise estimation when the decision is set at 0, it is very important that VAD also works well at low SNRs. Additional proof that almost all frames are declared as speech is given in Table 6. We can see that the performance of the ASR with VAD used in the G.723.1 Standard relative to the mel-cepstral DSR front-end Standard (Figure 6) stays almost the same whether the G.723.1 VAD and frame dropping strategy are used or not (Table 6).

When we made the analysis of the FEC and MSC errors obtained by the MFB VAD algorithm, we came to a conclusion that the errors occurred especially in the region of the stops and fricatives (Table 4). For the DSR (advanced front-end) VAD algorithm, the obtained FEC and MSC errors were smaller (Table 3) than the errors obtained by MFB VAD algorithm. But, when the training of the whole word acoustical models and speech recognition procedure use the same feature extraction procedure with included VAD algorithm, then FEC and MSC errors have smaller influence on the speech recognition accuracy than NDS and OVER errors.

Table 9: The amount of time needed for VAD computation in one frame.

VAD	Computational costs	Relative to the MFB VAD
G.729	78 μs	0.527
G.723.1	935 μs	6.317
DSR	96 μs	0.648
MFB	148 μs	1
MFB*	0.59 μs	0.004

*In this case, we assume that the MFB outputs are known, as they are used for the ASR process.

A lower computational cost is another advantage of MFB VAD in respect to VAD algorithms used in the standards. The computational costs were calculated with the amount of time needed for VAD computation in one frame. The amount of time needed for one frame is presented in Table 9. The tests were made on an HP-UX B2000 workstation. The amount of time was defined as the average computation over the million frames for all tested VAD algorithms. Table 9 shows that the amount of time needed for algorithm computation of the proposed MFB VAD is significantly lower than the ones for G.729, G.723.1, and DSR (advanced front-end) if it is assumed that we already have information about MFB outputs. It should also be noticed that the amount of time needed for VAD used in the G.723.1 Standard should be divided by three, because VAD decisions are made every 30 milliseconds, the MFB, DSR (advanced front-end), and G.729 VAD decisions are made every 10 milliseconds.

The use of a VAD algorithm together with frame dropping strategy is important at the back end of the DSR system, because the speech recognizer does not need to deal with noisy frames and the result is faster and more accurate ASR (Table 8). At the font end of the DSR system, where the processing power and memory available are limited, a reduction in computational costs using the MFB VAD is also welcome.

7. CONCLUSIONS

A novel approach for a VAD algorithm was presented based on MFB outputs with the so-called Hangover criterion. Comparative tests were made between the proposed MFB VAD algorithm and the three VAD algorithms used in the G.729, G.723.1, and DSR (advanced front-end) Standards. The results show that the total number of frame errors is higher for G.729 and G.723.1 VAD algorithms compared to MFB VAD algorithm. The total number of frame errors obtained by the DSR (advanced front-end) is lower than the one obtained by MFB VAD algorithm, however, the results of speech recognition performance with frame dropping strategy show better results with MFB VAD algorithm especially when clean only training mode is used. The computational costs of the MFB VAD are significantly lower than the costs when the standardized VAD algorithms are used (Table 9), especially when MFB feature extraction is used in speech recognition. This advantage is important at the front-end of the DSR system, where the processing power and memory available are limited.

REFERENCES

[1] J. C. Junqua and J. P. Haton, “Dealing with noisy speech and channel distortions,” in *Robustness in Automatic Speech Recognition: Fundamentals and Applications*, chapter 5, pp. 155–189, Kluwer Academic Publishers, Norwell, Mass, USA, 1996.

[2] ITU, “Coding of speech at 8 kbit/s using conjugate structure algebraiccode-excited linear-prediction (CS-ACELP) Annex B: A silence compression scheme,” ITU Recommendation G.729, 1996.

[3] ITU, “Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s. Annex A: Silence compression scheme,” ITU Recommendation G.723.1, 1996.

[4] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd, “The voice activity detector for the Pan-European digital cellular mobile telephone service,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP ’89)*, vol. 1, pp. 369–372, Glasgow, UK, May 1989.

[5] B. Kotnik, Z. Kacic, and B. Horvat, “A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction algorithm,” in *Proc. 7th European Conference on Speech Communication and Technology (ISCA EUROSPEECH ’01)*, pp. 197–200, Aalborg, Denmark, September 2001.

[6] J. Haigh and J. S. Mason, “A voice activity detector based on cepstral analysis,” in *Proc. 3rd European Conference on Speech Communication and Technology (ISCA EUROSPEECH ’93)*, pp. 1103–1106, Berlin, Germany, September 1993.

[7] S. McClellan and J. D. Gibson, “Variable-rate CELP based on subband flatness,” *IEEE Trans. Speech Audio Processing*, vol. 5, no. 2, pp. 120–130, 1997.

[8] P. Renevey and A. Drygajlo, “Entropy based voice activity detection in very noisy conditions,” in *Proc. European Conference on Speech Communication and Technology (ISCA EUROSPEECH ’01)*, pp. 1887–1890, Aalborg, Denmark, September 2001.

[9] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Processing Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[10] J. Stadermann, V. Stahl, and G. Rose, “Voice activity detection in noisy environments,” in *Proc. European Conference on Speech Communication and Technology (ISCA EUROSPEECH ’01)*, pp. 1851–1854, Aalborg, Denmark, September 2001.

[11] ETSI, “Speech processing, transmission and quality aspects (STQ), distributed speech recognition, advanced front-end feature extraction algorithm, compression algorithm,” ES 202 050 v1.1.1, 2002.

[12] H.-G. Hirsch and D. Pearce, “The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions,” in *Proc. Automatic Speech Recognition: Challenges for the Next Millennium (ISCA ITRW ASR ’00)*, pp. 181–188, Paris, France, September 2000.

[13] B. Milner and S. Semnani, “Robust speech recognition over IP networks,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP ’00)*, vol. 3, pp. 1791–1794, Istanbul, Turkey, June 2000.

[14] D. Pearce, “Enabling new speech driven services for mobile devices: an overview of the ETSI standards activities for distributed speech recognition front-ends,” in *Proc. Applied Voice Input/Output Society Conference (AVIOS ’00)*, San Jose, Calif, USA, May 2000.

[15] B. Kotnik, T. Rotovnik, Z. Kačič, B. Horvat, and I. Kramberger, “The design of mobile multimodal communication device-personal navigator,” in *Proc. International Conference on Trends in Communications (EUROCON ’01)*, vol. 2, pp. 337–340, Bratislava, Slovakia, July 2001.

[16] W. Zhang, L. He, Y. Chow, R. Yang, and Y. Su, "The study on distributed speech recognition system," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '00)*, vol. 3, pp. 1431–1434, Istanbul, Turkey, June 2000.

[17] ETSI, "Speech processing, transmission and quality aspects (STQ), distributed speech recognition, front-end feature extraction algorithm, compression algorithm," ES 201 108 v1.1.1, 2000.

[18] ITU, "Transmission performance characteristics of pulse code modulation channels," ITU Recommendation G.712, 1996.

[19] ETSI-SMG, "European digital cellular telecommunication system (phase 1)—transmission planning aspects for the speech service in GSM PLMN system," TSGSM03.50, 3.4.0, 1994.

[20] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (Version 3.0)*, Microsoft Corporation, Redmond, Wash, USA, 2000.

[21] B. Andrassy, D. Vlaj, and C. Beaugeant, "Recognition performance of the siemens front-end with and without frame dropping on the aurora 2 database," in *Proc. 7th European Conference on Speech Communication and Technology (ISCA EUROSPEECH '01)*, pp. 193–196, Aalborg, Denmark, September 2001.

[22] C. Benitez, L. Burget, B. Chen, et al., "Robust ASR front-end using spectral-based and discriminant features: experiments on the Aurora tasks," in *Proc. 7th European Conference on Speech Communication and Technology (ISCA EUROSPEECH '01)*, pp. 429–432, Aalborg, Denmark, September 2001.

[23] B. Kotnik, D. Vlaj, and B. Horvat, "Efficient noise robust feature extraction algorithms for distributed speech recognition (DSR) systems," *International Journal of Speech Technology*, vol. 6, no. 3, pp. 205–219, 2003.

[24] B. Kotnik, D. Vlaj, Z. Kačič, and B. Horvat, "Robust MFCC feature extraction algorithm using efficient additive and convolutional noise reduction procedures," in *Proc. International Conf. on Spoken Language Processing (ICSLP '02)*, pp. 445–448, Denver, Colo, USA, September 2002.

Damjan Vlaj graduated from the Faculty of Electrical Engineering and Computer Science, University of Maribor, in 1998. From April 1998 to October 2000, he worked at the Research and Studies Center, University of Maribor. Since October 2000, he has been employed at the Faculty of Electrical Engineering and Computer Science, University of Maribor. He has been a postgraduate student in the same faculty since October 1999. His research interests lie in the areas of robust speech recognition in the environment of fix and mobile telephony.

Bojan Kotnik received the Diploma degree from the University of Maribor in 2000. Since November 2000, he has been a postgraduate student and a young researcher at the Faculty of Electrical Engineering and Computer Science, University of Maribor. His research interests are in the areas of feature extraction, feature postprocessing, speech enhancement, and robust speech recognition.

Bogomir Horvat graduated from the Faculty of Electrical Engineering, University of Ljubljana, in 1963. He received the M.S. degree in 1975 and the Ph.D. degree in 1983 from the same faculty. He is currently employed in the Faculty of Electrical Engineering and Computer Science, University of Maribor, as a Full Professor, where he is the Head of the Institute of Electronics and Telecommunications and the Head of the Laboratory for Digital and Information Systems. His research interests include computer communications, man-machine communication, and system verification and validation.

Zdravko Kačič graduated from the Faculty of Electrical Engineering and Computer Science, University of Maribor, in 1986. He received the M.S. degree in 1989 and the Ph.D. degree in 1992 from the same faculty, where he is currently employed as a Full Professor. He is the Head of the Laboratory for Digital Signal Processing. His research interests are in the analysis of the complex sound scenes, systems for automatic speech recognition, and the creation of language resources.