
INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 6, 205–219, 2003

Efficient Noise Robust Feature Extraction Algorithms for Distributed Speech Recognition (DSR) Systems

BOJAN KOTNIK, DAMJAN VLAJ AND BOGOMIR HORVAT
University of Maribor, Faculty of Electrical Engineering and Computer Science, Maribor, Slovenia
bojan.kotnik@uni-mb.si

Abstract. The evolution of robust speech recognition systems that maintain a high level of recognition accuracy in difficult and dynamically-varying acoustical environments is becoming increasingly important as speech recognition technology becomes a more integral part of mobile applications. In distributed speech recognition (DSR) architecture the recogniser's front-end is located in the terminal and is connected over a data network to a remote back-end recognition server. The terminal performs the feature parameter extraction, or the front-end of the speech recognition system. These features are transmitted over a data channel to the remote back-end recogniser. DSR provides particular benefits for the applications of mobile devices such as improved recognition performance compared to using the voice channel and ubiquitous access from different networks with a guaranteed level of recognition performance. A feature extraction algorithm integrated into the DSR system is required to operate in real-time as well as with the lowest possible computational costs.

In this paper, two innovative front-end processing techniques for noise robust speech recognition are presented and compared, time-domain based frame-attenuation (TD-FrAtt) and frequency-domain based frame-attenuation (FD-FrAtt). These techniques include different forms of frame-attenuation, improvement of spectral subtraction based on minimum statistics, as well as a mel-cepstrum feature extraction procedure. Tests are performed using the Slovenian SpeechDat II fixed telephone database and the Aurora 2 database together with the HTK speech recognition toolkit. The results obtained are especially encouraging for mobile DSR systems with limited sizes of available memory and processing power.

Keywords: noise robustness, distributed speech recognition, frame-attenuation, spectral subtraction, feature extraction

1. Introduction

As automatic speech recognition (ASR) technology becomes more and more appealing to wireless applications, applications conducted in automobile environments or hands-free communication, noise robustness determines the usability of ASR systems in these applications. The performance of current ASR systems radically deteriorates when noise (in most cases, background noise or background speech) interferes with input speech. The automatic speech recognition accuracy achieved in laboratory environments is relatively high. As soon as recognition is placed in a natural environment, however, recognition accuracy almost always is significantly worse. These two facts, noise and environment dependencies, reduce the success of an ASR system in real-world applications. The degradation of the performance of the ASR system is due mainly to the mismatch between training and recognition conditions, and most of the methods for robust speech recognition are focused in the minimization of the mismatch. Thus far, many researchers' interests have been attracted to this problem and a large research effort has been conducted in the field of robustness. The studies conducted can be categorized in one of these main groups:

- Robust parameterizations: the speech signal is represented using parameters that are minimally affected by the noise.
- Compensation of the noise effect over the representation of the speech: these methods attempt to remove noise from parameters representing the speech.
- Adaptation of the models to noise conditions: the models in the recognizer are contaminated in order to properly model the noisy speech (Varga and Moore, 1990).

It is also possible to combine different approaches from these main groups. The final result of research and development activities in the field of robust speech technology will be multiconditionally operative algorithms.

In this paper, we focus on improving the robustness of Mel Frequency Cepstral Coefficients (MFCC) speech parameterization only, although the robustness in both front-end and back-end parts of an ASR system should be considered for practical applications. The feature extraction procedure is very important in the process of automatic speech recognition and has a great influence on the efficiency of automatic speech recognition systems. Two different front-end processing techniques developed for noise robust speech recognition are presented and compared, time-domain based frame-attenuation (TD-FrAtt) and frequency-domain based frame-attenuation (FD-FrAtt). Both of them use spectral subtraction based on minimum statistics as an additive noise reduction scheme. The main differences between the proposed techniques are two different voice activity detection and frame attenuation strategies. In a very noisy environment, recognition performance decreases, in part, due to imperfect speech detection; therefore, efficient speech/non-speech detection is crucial. In TD-FrAtt a special weighting function based on the zero crossing measure and frame energy is introduced. In FD-FrAtt a spectral energy-based voice activity detector with hangover criterion is incorporated. A novel frame attenuation algorithm, instead of the generally-used frame dropping, is presented; the non-speech frames are attenuated and not dropped from further processing. At the final stage of the TD-FrAtt and FD-FrAtt algorithms, a standard mel frequency cepstral coefficient feature extraction procedure is performed.

Generally speaking, the feature extraction algorithm, which is a part of the DSR system, where the size of memory available and the processing power are limited, is expected to meet the following requirements. First, it is expected to operate in real-time with the lowest possible computational costs, and second, it is required, as much as possible, to produce speech feature representation invariant to noisy environments and SNRs.

Section 2 presents a general overview of the distributed speech recognition system. The noise robust feature extraction algorithm is an important part of a DSR system. Two proposed noise reduction algorithms combined with the mel-cepstrum feature extraction procedure are described in Section 3. Two speech databases were used for the experimental work. The Aurora 2 database is designed to evaluate the performance of speech recognition algorithms under noisy conditions. The Slovenian SpeechDat II fixed-telephone database was recorded in a natural environment over fixed Public Switched Telephone Network (PSTN) using a digital ISDN interface. Section 4 presents both databases, the test procedures, and the results. The results are discussed in Section 5. Finally, conclusions are given in Section 6.

2. Description of the Distributed Speech Recognition System

The trend to ever-increasing use of data communication in modern telecommunication technology is spreading to the mobile wireless world. People now want the ability to access information while on the move, and the technology is now starting to be deployed to enable this to happen. Small portable multimodal devices (e.g., PDA, personal navigator) that will be used to access these data services require improved user interfaces using speech input (Kotnik et al., 2001a). At present, however, the complexities of medium and large vocabulary speech recognition systems are beyond the memory and computational resources of such devices (Oviatt, 2000).

Centralised servers can share the computational load between users and so enable upgrading of the provided technologies and services. Mobile voice networks, however, can degrade speech recognition performance obtained from centrally deployed recognisers. These degradations are the result of both low bit rate speech coding and channel transmission errors. The basic idea of a distributed speech recognition (DSR) system is to overcome these problems by eliminating the speech channel and using an error-protected data channel to send a parameterised speech representation that is suitable for recognition. Figure 1 displays a block diagram of the DSR system (Pearce, 2000).

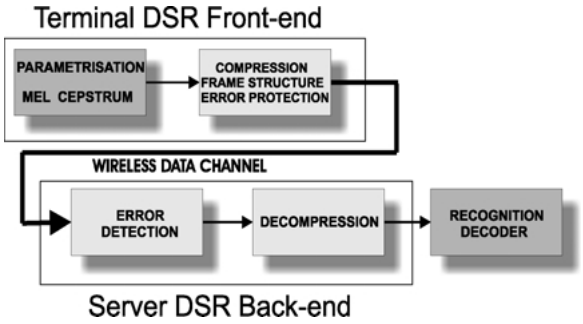

Figure 1. Block diagram of the DSR system.

The automatic speech recognition process is distributed between the terminal and the network. The terminal performs feature parameter extraction of the spoken speech, or the front-end of the speech recognition system. Features are compressed and transmitted together with error protection and correction data (CRC—Cyclic Redundancy Check) over a data channel to a remote back-end recogniser. The main advantage of this approach is that the transmission channel has minimal impact on the recognition system performance, and channel invariability is achieved. This performance is useful for those DSR services provided for a particular network by a network operator. Additionally, it is useful for 3rd party DSR applications that may be accessed over a variety of different networks (Pearce, 2000).

2.1. DSR Standardization in ETSI

To enable widespread applications using DSR in the marketplace, a standard for the front-end is needed to ensure compatibility between the terminal and the remote recogniser. The Aurora DSR Working Group within ETSI has been actively developing this standard over the last three years. A reference database, Aurora 2 (Hirsch and Pearce, 2000), and an experimental framework have been established to evaluate alternative proposals for the front-end feature extraction algorithm. This database is based on the original TI-Digits database with controlled filtering and simulated noise addition over a range of signal to noise ratios from 20 dB to −5 dB. A reference recogniser configuration using a HTK HMM speech recognition toolkit was applied to investigate those changes solely at the front-end. This database has been made publicly available via the European Language Resources Association (ELRA) (Hirsch and Pearce, 2000).

Feature extraction algorithms that operate in DSR should produce speech feature representation which is, as far as possible, invariant to noisy environments and SNRs if the equivalent utterance is spoken in each of these different conditions. Moreover, DSR algorithms should operate in real-time as well as with the lowest possible computational costs.

3. Proposed Noise-Reduction and Feature-Extraction Algorithms

To improve the performance of modern ASR systems, it is crucial to develop new efficient approaches for speech signal pre-processing and feature extraction, because all the succeeding processing steps in ASR systems are highly dependent on the quality of the extracted features. Ideally, noise robustness could be solved in the feature extraction unit of an ASR system, which would eliminate the need for additional application-specific data collection or parameter compensation. When primarily attempting to improve automatic speech recognition accuracy in adverse environments, special attention was focused on the grey blocks in the Fig. 1 block diagram. Two proposed front-end algorithms and the automatic speech recognition system will be described.

Figure 2 shows two proposed noise robust front-end algorithms denoted as time-domain frame-attenuation (TD-FrAtt) and frequency-domain frame-attenuation (FD-FrAtt), which share some equivalent processing blocks. Both proposals involve spectral subtraction based on minimum statistics as a noise reduction technique and a mel-cepstrum feature extraction procedure as a final stage in the front-end. All processing steps from a block diagram will be described in subsequent subsections. As will be seen, TD-FrAtt and FD-FrAtt achieve different performance in different training modes and with different speech recognition systems.

3.1. Framing, Windowing and FFT Computations

In all practical signal processing applications, it is necessary to work with short terms or frames of the noisy input signal $y[n]$. It is also necessary to select a portion of the signal that can be reasonably assumed to be stationary (Deller et al., 1993). Frames of $W = 200$ samples (frame length of 25 ms at a sampling frequency of 8 kHz) were used. The frame shift interval was 80

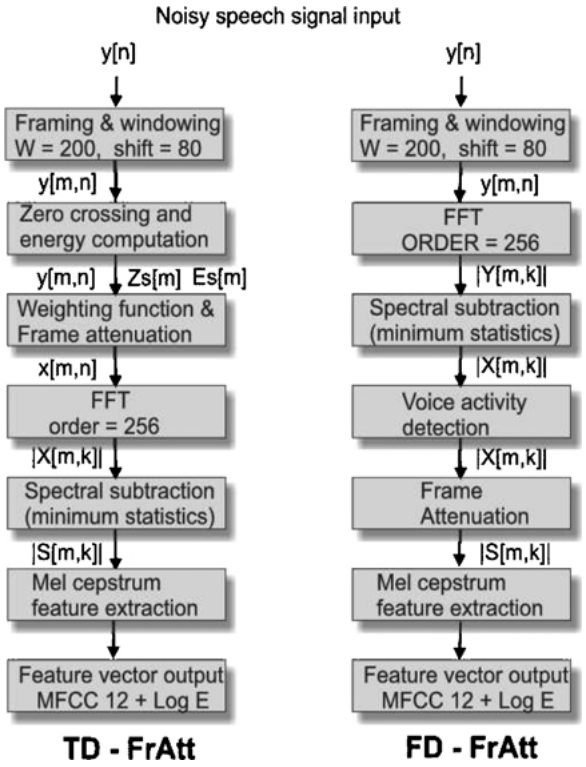

Figure 2. Two proposed noise robust front-end algorithms: time-domain frame-attenuation (TD-FrAtt) and frequency-domain frame-attenuation (FD-FrAtt).

samples (60% overlapped frames). The frame was multiplied by a Hamming window function, thus producing the windowed noisy input signal y[m,n], where m is a frame index and $n = 0, 1, 2, \ldots, W$ is a sample index. A zero padded 256-order (power of 2) FFT algorithm was applied, as a window length of 200 samples was used.

3.2. *Spectral Subtraction Based on Minimum Statistics*

Spectral subtraction is a method for restoring the magnitude spectrum or the power spectrum of a signal observed in additive noise, by subtracting an estimate of the noise spectrum from the noisy signal spectrum (Boll, 1979). TD-FrAtt and FD-FrAtt actually contain equivalent spectral subtraction algorithms. In Eqs. (1)–(6) the notation of FD-FrAtt (see Fig. 2) is used. When replacing variables Y with X and X with S, a corresponding notation for the TD-FrAtt is also achieved.

In real applications of speech recognition, the speech signal is usually affected by additive background noise, due to other audio sources in the environment where the speaker is. The noisy speech signal can be represented as:

$$y[m, n] = x[m, n] + \eta[m, n], \tag{1}$$

where x and η are speech and noise components, respectively. Furthermore, with $|Y[m, k]|^2$ (a noisy speech signal power spectrum) and $|N[m, k]|^2$ (an estimate of noise power spectrum) given, it is possible to estimate the power spectrum $|X[m, k]|^2$ of the original (uncorrupted) speech as:

$$|X[m, k]|^2 = |Y[m, k]|^2 - |N[m, k]|^2. \tag{2}$$

In order to reflect the uncertainty of the estimate $|N[m, k]|^2$, it was found that the estimate $|N[m, k]|^2$ should be weighted by an over-subtraction factor depending on the frequency k and the signal-to-noise ratio (SNR). Because equation (2) does not guarantee that $|X[m, k]|^2$ is always positive, it is necessary to involve a flooring factor. The two concepts—over-subtraction and flooring, and the estimator for $|N[m, k]|^2$ described in Martin (1994), lead to the following algorithm: Two smoothed power spectra are determined from $|Y[m, k]|^2$ according to the outputs of the first order digital low-pass filters:

$$\begin{aligned} |\bar{Y}_1[m, k]|^2 &= 0.40 \cdot |\bar{Y}_1[m-1, k]|^2 - 0.60 \cdot |Y[m, k]|^2 \\ |\bar{Y}_2[m, k]|^2 &= 0.75 \cdot |\bar{Y}_2[m-1, k]|^2 - 0.25 \cdot |Y[m, k]|^2. \end{aligned} \tag{3}$$

The power spectrum $|N[m, k]|^2$ is estimated using the minimum of the smoothed power spectrum within a moving interval I^m with fixed width D:

$$|Y_{\text{Min}}[m, k]|^2 = \underset{m \in I^m}{\text{Min}} |\bar{Y}_2[m, k]|^2 \quad I^m = [m - D, m]. \tag{4}$$

The width $D = 25$ of interval I^m has to be chosen in such a way that the samples $Y[m, k]$, $m \in I^m$ cover approximately one syllable. Based on the following estimate:

$$|N[m, k]|^2 = |Y_{\text{Min}}[m, k]|^2, \tag{5}$$

the power spectrum $|X[m,k]|^2$ of uncorrupted speech is estimated by spectral subtraction (Martin, 1994):

$$|X_{\mathrm{SUB}}[m,k]|^2 = |Y[m,k]|^2 - \alpha \cdot \frac{|Y[m,k]|^2}{|\bar{Y}_1[m,k]|^2} \cdot |N[m,k]|^2$$

$$|X[m,k]|^2 = \begin{cases} \beta \cdot |Y[m,k]|^2, & \text{if } |X_{\mathrm{SUB}}[m,k]|^2 < \beta \cdot |Y[m,k]|^2 \\ |X_{\mathrm{SUB}}[m,k]|^2, & \text{if } |X_{\mathrm{SUB}}[m,k]|^2 \geq \beta \cdot |Y[m,k]|^2 \end{cases}, \tag{6}$$

where $\alpha = 1.5$ and $\beta = 0.1$ are over-subtraction and flooring factors, respectively. The parameter α in Eq. (6) controls the amount of noise subtracted from the noisy signal. For full noise subtraction, $\alpha = 1$, and for over-subtraction $\alpha > 1$. The described spectral subtraction algorithm based on minimum statistics is due to the symmetry of the FFT transform calculated for frequency bins $k = 0, 1, \ldots, N/2$ only, where N denotes a FFT order. The main advantage of this algorithm is that no explicit detection of non-speech segments in noise spectrum $|N[m,k]|^2$ estimation procedure is needed.

3.3. *Voice Activity Detection and Frame-Attenuation Algorithms*

Voice activity detection (VAD) is the ability 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, audio conferencing, and echo cancellation. A major cause of errors in automatic speech recognition systems is inaccurate detection of the beginning and ending boundaries of test and reference patterns (Junqua and Haton, 1996). It is essential for ASR algorithms that the speech segments are separated reliably from the non-speech parts. Similarly, VAD algorithms are used to attenuate noise-only frames of the noisy input speech signal and not to discard them completely. This process is called a frame-attenuation and stands in opposition to conventional frame dropping, where unwanted frames are dropped completely from further processing (Andrassy et al., 2001; Benitez et al., 2001). Accurate determination of endpoints is not very difficult if the SNR is high (greater than 35 dB). Unfortunately, the majority of practical recognizers must work with a much smaller SNR, typically 25 or 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 end of utterances. Two different VADs and frame attenuation algorithms are presented in Sections 3.3.1 and 3.3.2. The first one, which is called *frame weighting function*, is based on short time energy and the zero crossing measure and works in the time-domain. The second one is implemented in the frequency-domain and is based on the frame spectral energy and SNR measurements.

3.3.1. Time-Domain Based VAD and Frame-Attenuation Algorithm. A short time zero crossing value and a short time signal energy of the windowed input signal $y[m,n]$ are used for computing the frame weighting function. The zero crossing value is the sum of the number of times the sequence changes the sign. A short-term zero crossing measure for the window m of length N of the windowed input signal $y[m,n]$ is defined as (Deller et al., 1993):

$$Z_S[m] = \frac{1}{N} \sum_{n=1}^{N-1} \frac{|\mathrm{sgn}\{y[m,n]\} - \mathrm{sgn}\{y[m,n-1]\}|}{2}, \tag{7}$$

where $\mathrm{sgn}\{y[m,n]\}$ is a sign function, defined by the following equation:

$$\mathrm{sgn}\{y[m,n]\} = \begin{cases} +1, & y[m,n] \geq 0 \\ -1, & y[m,n] < 0 \end{cases}. \tag{8}$$

The short-term energy measure for the N-length frame m is defined by the equation:

$$E_S[m] = \frac{1}{N} \sum_{n=0}^{N-1} y^2[m,n]. \tag{9}$$

The short-term zero crossing and energy measures plotted for the word */forum/* are presented in Fig. 3.

The speech signal begins and ends with the noise regions (N). There the signal energy is low and, at the same time, the zero crossing value is high (Deller et al., 1993). The utterance has a strong fricative at the beginning (in the picture designated as (U)-unvoiced region); hence, the zero crossing level is very high and the energy is low initially, but higher than in the noise regions. The opposite is true when the signal enters the voiced portion (V) of the utterance. There the energy of the signal is very high and the zero crossing level very low (lower than in the noise regions). If it is necessary to find speech regions in the input signal (which consists of voiced and also unvoiced regions), the function

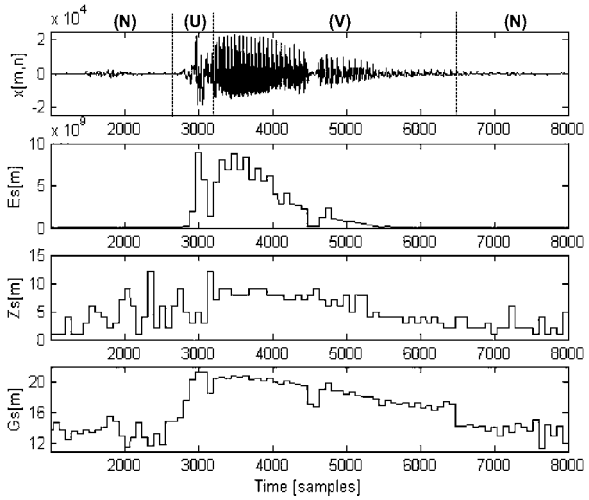

Figure 3. Interdependence between the zero crossing and the energy measures for the word */forum/*.

$G_S[m]$ can be defined as (Kotnik et al., 2001b):

$$G_S[m] = \log_e\left(\frac{1}{Z_S[m]} \cdot E_S[m]\right). \tag{10}$$

Function $G_S[m]$ is computed for each frame of the input speech signal. $G_S[m]$ helps to write the following weighting function:

$$x[m,n] = \begin{cases} a \cdot y[m,n], & G_S[m] < t_1 \\ b \cdot y[m,n], & t_1 \le G_S[m] < t_2 \\ c \cdot y[m,n], & t_2 \le G_S[m] < t_3 \\ d \cdot y[m,n], & t_3 \le G_S[m] \end{cases}. \tag{11}$$

The presented weighting function is used as frame-attenuation: noisy frames are not deleted, but appropriately weighted, and a suitable weighting coefficient is chosen on the basis of $G_S[m]$. The weighting coefficients a, b, c, and d were empirically estimated and are shown in Table 1.

Table 1. The values of weighting coefficients.

Coefficient	Value
a	0.3
b	0.7
c	1.2
d	0.8

Thresholds t_1, t_2 and t_3 have default values before the processing of the speech signal is started: $t_1 = 0.15\ G_S[0]$, $t_2 = 0.5\ G_S[0]$, and $t_3 = 0.85\ G_S[0]$. These values are then automatically updated using the following equations:

$$\begin{aligned} t_1 &= 0.15 \cdot \max(G_S[m]) + 0.85 \cdot \min(G_S[m]) \\ t_2 &= 0.50 \cdot \max(G_S[m]) + 0.50 \cdot \min(G_S[m]) \\ t_3 &= 0.85 \cdot \max(G_S[m]) + 0.15 \cdot \min(G_S[m]) \end{aligned}. \tag{12}$$

3.3.2. Frequency-Domain Based VAD and Frame-Attenuation Algorithm. A VAD module classifies frames as speech or non-speech (noise) by comparing the SNR to a threshold. The SNR corresponds to the difference between the short-term and the long-term signal log-energy estimates. The long-term estimate is updated when the VAD decides that the current frame corresponds to non-speech, and the energy of the current frame is used as a short-term estimate. In the first step, a short-term spectral energy $E_f[m]$ is calculated for each frame as:

$$E_f[m] = q \cdot \log_e\left(1 + \frac{1}{N}\sum_{k=0}^{N} |X[m,k]|^2\right), \tag{13}$$

where $N = 256$ represents an FFT order. q is an empirically estimated frame energy multiplication factor used to improve the speech-noise decision process; in this case it is set to 23. Then, the spectral energy of the current frame $E_f[m]$ is used in the update of the long-term mean spectral energy $E_{m\mathrm{NEW}}$ as:

$$\begin{aligned} &\text{IF } \left(E_f[m] - E_{m_{\mathrm{OLD}}}\right) < 20 \text{ THEN} \\ &\quad E_{m_{\mathrm{NEW}}} = E_{m_{\mathrm{OLD}}} + \frac{E_f[m] - E_{m_{\mathrm{OLD}}}}{100}. \\ &\text{ELSE} \\ &\quad E_{m_{\mathrm{NEW}}} = E_{m_{\mathrm{OLD}}} \end{aligned} \tag{14}$$

After determining the short-term spectral energy $E_f[m]$ and the long-term mean spectral energy $E_{m\mathrm{NEW}}$, the speech-noise decision procedure of the current frame can begin. Figure 4 shows a flowchart of the proposed frame spectral energy-based VAD algorithm. A similar scheme for VAD decision was used in ITU recommendation G.723.1 A (1996). It should be noted here that before the VAD processing of the input signal is started, the frame counters *SpeechToNoiseFrame* and *SpeechFrame*, as well as long-term mean spectral energy $E_{m\mathrm{OLD}}$, are initialized to 0.

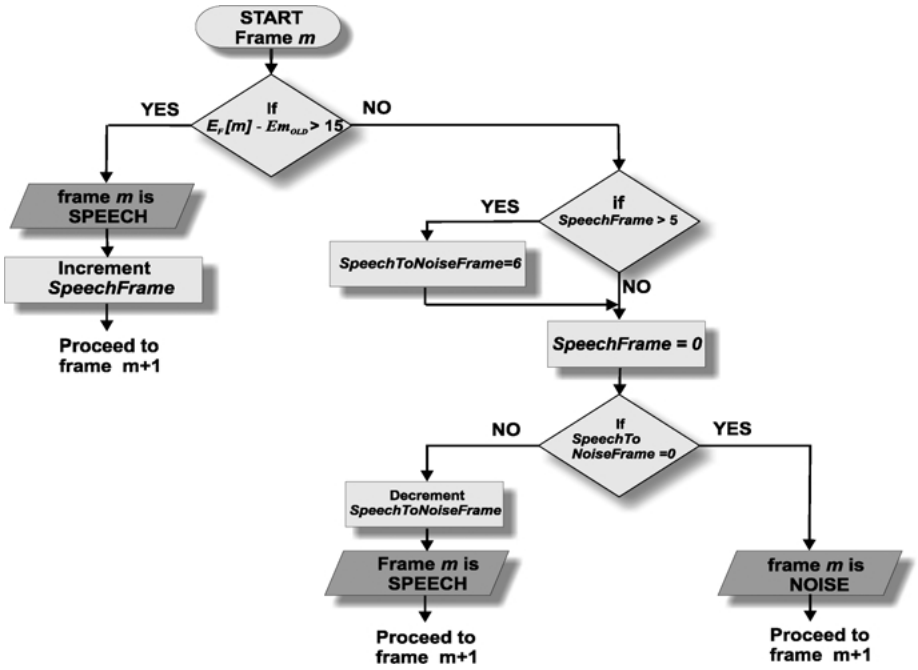

Figure 4. Flowchart of the frame spectral energy based voice activity detector.

If the current frame is declared as *noise*, then its spectral magnitude values $|S[m,k]|$ are attenuated:

$$|S[m,k]| = \frac{|X[m,k]|}{100}; \tag{15}$$

otherwise, the frame spectrum is left unchanged. The frames that are declared as "noise" are, in the conventional frame dropping algorithm, discarded from further processing. The frame attenuation used here does not drop frames that are declared as noise; it only attenuates them. In addition, a *SpeechToNoiseFrame* factor prevents a misclassification of weak fricatives to noise at the end of the speech segments, since seven frames at the end of a speech segment at least six frames long are also declared as speech.

3.4. Mel-Cepstrum Feature Vector Extraction Procedure

The low-frequency components of the magnitude spectrum are ignored. The useful frequency band lies between $f_{\text{start}} = 32$ Hz and half of the actual sampling frequency (in our case $f_{\text{sampl}} = 8$ kHz). This band is divided into $Ch = 36$ equidistant channels in the mel-frequency domain. Each channel has a triangle-shaped frequency window. Consecutive channels are half-overlapping. The output of the mel filter is a weighted sum of the FFT magnitude spectrum values $|S[m,k]|$ in each band. The output of mel filtering is subjected to a natural logarithm function. Twelve cepstral coefficients are calculated from the output of the non-linear transformation by discrete cosine transformation. The logarithmic frame energy measure $\log E$ for each frame is computed. This $\log E$ serves as the 13th parameter, i.e., element, in a final feature vector. The described procedure is repeated for each frame m of the input signal $|S[m,k]|$. The feature vector produced with proposed front-end consists of 13 parameters: 12 mel-cepstrum coefficients and the energy parameter. The dynamic features (delta and acceleration coefficients) are additionally calculated in the back-end, thus producing a final feature vector composed of 39 elements.

In both TD-FrAtt and FD-FrAtt noise robust feature extraction algorithms some empirically estimated constants are used. These parameters are database and recognition task independent. Therefore, TD-FrAtt and FD-FrAtt algorithms are portable to different speech databases. In our experimental framework the Aurora 2 and Slovenian SpeechDat II databases are considered.

4. The Description of the Experimental Framework

Experiments were made using the Aurora 2 database (Hirsch and Pearce, 2000), which is designed to evaluate the performance of speech recognition algorithms in noisy conditions, and the Slovenian SpeechDat II database (Kaiser and Kacic, 1997), which is recorded in a natural environment over a fixed PSTN using a digital ISDN interface. The tests on these two databases are described in the next two subsections.

4.1. Tests on Aurora 2 Database

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

The reference recognizer is based on the HTK software (Young, 1997). The training and recognition parameters were defined to compare the recognition results when applying different feature extraction schemes. The digits were modelled as whole word HMMs with 16 emitting states per word with simple left-to-right models without skips over states and three Gaussian mixtures per state. Two pause models were defined. The first one, called "sil", consists of three emitting states, and the second pause model, called "sp", was used to model pauses between words and consists of a single emitting state which is tied with the middle state of the first pause model. Six Gaussian mixtures for each state were used for the pause models.

For training on clean data, 8440 utterances from the training part of the TI-Digits database (Leonard, 1991) were chosen, which contained the recordings of 55 male and 55 female adults. These signals were filtered by the G.712 characteristics (ITU recommendation G.712, 1996) with no noise added. The same 8440 utterances were used for the multicondition training. They were divided into 20 subsets, each of which included 422 utterances. There were a few utterances from all training speakers in each subset. The 20 subsets represented four different noise scenarios at five different SNRs. The noises were a suburban train, babble, a car and an exhibition hall. The SNRs were 20 dB, 15 dB, 10 dB, 5 dB and the clean condition. They were filtered by the G.712 characteristic before adding speech and noise to produce the noisy speech signal.

Three different test sets were defined (Hirsch and Pearce, 2000). Four subsets with 1001 utterances in each were obtained by splitting 4004 utterances from 52 male and 52 female speakers in the TI-Digits test part. The recordings of all speakers were present in each subset. To each subset one noise signal at SNRs of 20 dB, 15 dB, 10 dB, 5 dB, 0 dB and -5 dB was added, and the clean case without adding noise was taken as the seventh condition. Speech and noise were, once again, filtered by the G.712 characteristic before addition.

The first test set was called test set A. In this test set four noises—a suburban train, babble, a car and an exhibition hall—were added to the four subsets, so the set consisted of 28028 utterances. There was a high match of training and test data, owing to the fact that this test set contained the same noises 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—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 different noises other than those used for training were considered.

The third test set was called test set C, and it contained two out of four subsets with 1001 utterances each. Here speech and noise were filtered by a MIRS characteristic (ETSI-SMG technical specification, 1994), before being added to the SNRs of 20 dB, 15 dB, 10 dB, 5 dB, 0 dB and -5 dB. MIRS, which can be explained as a frequency characteristic that simulates the behaviour of a telecommunication terminal, meets the official requirements for the terminal input frequency response as specified e.g., in the technical specification (ETSI-SMG technical specification, 1994). The street and suburban train were used as additional noise. The purpose of this set was to show the influence on recognition performance when a different frequency characteristic was present at the input of the recogniser.

Tables 2 and 3 show the speech recognition results achieved by the Aurora 2 database and feature extraction algorithms TD-FrAtt and FD-FrAtt, respectively. Acoustical models were trained on clean and noisy speech data (Multiconditional training procedure).

Table 2. Aurora 2 database recognition results and performance relative to the current ETSI ES 201 108 standard achieved with feature extraction algorithm TD-FrAtt and multiconditional training.

	Aurora 2 multicondition training—results														
	A					B					C				
	Subway	Babble	Car	Exhibition	Average	Resturant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall	Percentage improvement
Clean	98.74	98.64	98.54	98.98	98.73	98.74	98.64	98.54	98.98	98.73	98.89	98.52	98.71	98.72	13.06
20 dB	98.04	97.61	98.09	98.15	97.97	97.91	97.43	97.85	98.67	97.97	97.73	97.85	97.79	97.93	19.65
15 dB	96.71	96.88	97.13	97.50	97.06	96.04	96.43	96.72	97.06	96.56	96.74	96.37	96.56	96.76	10.44
10 dB	94.13	93.80	95.88	94.90	94.68	91.71	94.28	93.76	95.12	93.72	93.72	94.52	94.12	94.18	4.66
5 dB	89.18	85.00	90.01	87.06	87.81	81.75	87.75	86.10	87.91	85.88	86.50	86.04	86.27	86.73	7.32
0 dB	73.48	56.66	70.10	69.25	67.37	54.58	68.65	63.92	63.99	62.79	63.98	61.85	62.92	64.65	12.53
−5 dB	38.92	21.41	27.79	34.72	30.71	19.78	32.10	26.92	25.27	26.02	28.20	25.38	26.79	28.05	4.55
Average	90.31	85.99	90.24	89.37	88.98	84.40	88.91	87.67	88.55	87.38	87.73	87.33	87.53	88.05	
	13.80%	−16.27%	27.59%	11.21%	9.54%	−6.80%	14.43%	0.23%	23.62%	8.09%	26.80%	19.22%	23.13%		12.20

Table 3. Aurora 2 database recognition results and performance relative to the current ETSI ES 201 108 standard achieved with feature extraction algorithm FD-FrAtt and multiconditional training.

	Aurora 2 multicondition training—results														
	A					B					C				
	Subway	Babble	Car	Exhibition	Average	Resturant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall	Percentage improvement
Clean	98.86	98.85	98.60	98.86	98.79	98.86	98.85	98.60	98.86	98.79	98.96	98.73	98.85	98.80	18.82
20 dB	98.28	98.46	98.21	98.24	98.30	98.50	97.70	98.39	98.33	98.23	98.00	97.85	97.93	98.20	30.33
15 dB	97.05	97.67	98.03	97.47	97.56	97.42	97.22	97.73	97.01	97.35	97.08	97.22	97.15	97.39	28.15
10 dB	95.09	95.44	96.63	95.28	95.61	93.06	95.56	95.74	95.50	94.97	94.44	95.47	94.96	95.22	22.04
5 dB	89.50	88.66	92.10	88.28	89.64	83.57	90.42	89.98	88.92	88.22	88.24	89.24	88.74	88.89	22.45
0 dB	74.95	64.30	75.84	72.17	71.82	59.38	73.16	72.74	70.84	69.03	68.01	68.68	68.35	70.01	25.73
−5 dB	39.70	27.36	32.90	38.44	34.60	24.78	37.24	37.61	33.88	33.38	30.89	33.01	31.95	33.58	11.93
Average	90.97	88.91	92.16	90.29	90.58	86.39	90.81	90.92	90.12	89.56	89.15	89.69	89.42	89.9	
	19.73	7.93%	41.84%	18.86%	22.71%	6.80%	29.12%	26.49%	34.09%	23.95%	35.27%	34.30%	34.80%		26.09

Absolute results (upper part of the table) at different SNRs and conditions as well as improvement relative to the current ETSI ES 201 108 standard for distributed speech recognition systems (ETSI ES 201 108, 2000) are shown (the row on the bottom and column on the right of the table).

Tables 4 and 5 present speech recognition results achieved by the Aurora 2 database and feature extraction algorithms TD-FrAtt and FD-FrAtt, respectively. Acoustic models were trained on clean only speech data (Clean training procedure). Absolute results (upper part of the table) at different SNRs and conditions are presented, as well as improvement relative to the current ETSI ES 201 108 front-end standard (the row on the bottom and column on the right).

4.2. Tests on Slovenian SpeechDat II Database

Within the SpeechDat II project (Van den Heuvel et al., 2001) a total of 28 databases have been collected, covering eleven European languages as well as some major dialectal variants and minority languages. Twenty databases were recorded over the fixed telephone network (FDB), five databases over the mobile network (MDB), and three databases were designed for speaker verification via telephone (SDB). The recordings of the FDB and MDB databases cover between 500 and 5000 calls by different speakers being recorded in a single session (except for two MDBs using multiple sessions). The duration of each recording session is 4–8 minutes.

The databases are intended to be used for developing a number of applications such as information services (e.g., timetable information), transaction services (e.g., home shopping, home banking) and other call processing services. All SpeechDat databases are orthographically transcribed. The four categories of non-speech acoustic events are transcribed as:

- filled pauses ([fil]) and hesitations (uh, um, er, ah, mm, etc.),
- speaker noise ([spk]). All kinds of sounds and noises made by the calling speaker that have no relation to the prompted text, e.g., lip smack, cough, grunt, throat clear, tongue click, laugh, etc.,
- intermittent noise ([int]). This category contains noises of an intermittent nature not being generated by the speaker. Examples are music, door slam, background speech, phone ringing, and
- stationary noise ([sta]) not being intermittent (and not speaker generated).

As can be seen above, [fil] and [spk] noise types originate from the speaker and do not usually overlap with the target speech, whereas the last two may occur simultaneously with the target speech.

The COST 249 SpeechDat reference recognizer RefRec (Lindberg et al., 2000; COST 249 SpeechDat SIG, 2000) was used in tests on the Slovenian SpeechDat II database. The reference recogniser is a fully automatic training procedure for building a phonetic recogniser. The reference recogniser relies on the HTK software (Young, 1997) and SpeechDat II compatible databases and is designed to serve as a reference system in speech recognition research. Version 0.96 of the reference recogniser, which takes into account labelled non-speech acoustic events during training and provides robustness against them during testing, is used for the experiment. So, [spk] and [fil] labels are used during training and context independent models for speaker noise and filled pauses are then generated. The noise markers for stationary noise ([sta]) and intermittent noise ([int]) are ignored both during training and testing. The phonetic HMMs are trained from orthographic (word-level) transcriptions using a pronunciation lexicon and a "flat start" boot-strapping procedure. Training starts from context-independent, single Gaussian monophones. The procedure of training the HMMs is described in more detail in Lindberg et al. (2000). In the final training stage, the monophones and the tied state triphone models are improved by Gaussian mixture splitting and reestimation up to 32 components. The models trained by the reference recogniser can be used to provide the benchmark results for a number of different applications.

The baseline test was performed using the ETSI ES 201 108 front-end in model training and recognition procedures. In the next steps, both procedures were repeated with proposed feature extraction modules TD-FrAtt and FD-FrAtt, so that the speech recognition performance could be compared. Three common speech recognition tasks I, A, and BC—have been designed for some of the sub-corpora, as shown in Table 6.

Common test procedures are used for all the tests, ensuring identical rules of test design across databases. Currently, there are three such procedures, denoted SVIP (Small Vocabulary Isolated Phrase), SVWL (Small Vocabulary Word Loop) and MVIP (Medium Vocabulary Isolated Phrase). Only SVIP and SVWL were used in our tests. Utterances with OOV (out-of-vocabulary words), mispronunciation, unintelligible speech or truncations were excluded in all procedures,

Table 4. Aurora 2 database recognition results and performance relative to the current ETSI ES 201 108 standard achieved with feature extraction algorithm TD-FrAtt and clean training.

	Aurora 2 clean training—results														
	A					B					C				
	Subway	Babble	Car	Exhibition	Average	Resturant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall	Percentage improvement
Clean	99.32	99.09	99.16	99.44	99.25	99.32	99.09	99.16	99.44	99.25	99.23	99.12	99.18	99.24	21.44
20 dB	96.13	96.28	96.78	96.70	96.47	96.56	96.67	97.05	97.28	96.89	96.87	96.85	96.86	96.72	30.72
15 dB	94.04	94.46	94.63	93.88	94.25	94.07	94.41	95.32	94.81	94.65	94.10	94.61	94.36	94.43	53.79
10 dB	87.50	87.91	90.38	87.86	88.41	88.55	90.04	90.24	90.27	89.28	86.40	89.43	87.92	88.66	63.92
5 dB	75.02	73.11	80.93	72.69	75.44	67.92	77.08	77.20	78.48	75.17	72.97	75.26	74.12	75.07	58.21
0 dB	53.35	40.64	54.75	48.30	49.26	38.67	52.90	49.19	52.03	48.20	45.84	47.62	46.73	48.33	37.51
−5 dB	26.79	13.45	18.64	20.72	19.90	12.25	23.86	18.12	18.59	18.21	19.81	19.98	19.90	19.22	11.71
Average	81.21	78.48	83.49	79.89	80.77	76.75	82.22	81.80	82.57	80.84	79.24	80.75	80.00	80.64	
	38.42%	57.08%	58.10%	41.88%	50.25	50.96%	53.80%	61.07%	60.73%	56.70%	38.63%	43.20%	40.92%		51.52

Table 5. Aurora 2 database recognition results and performance relative to the current ETSI ES 201 108 standard achieved with feature extraction algorithm FD-FrAtt and clean training.

	Aurora 2 clean training—results														
	A					B					C				
	Subway	Babble	Car	Exhibition	Average	Resturant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall	Percentage improvement
Clean	98.93	99.24	98.90	99.23	99.08	98.93	99.24	98.90	99.23	99.08	99.02	99.18	99.10	99.08	5.04
20 dB	95.95	96.77	96.27	95.50	96.12	96.68	95.89	96.72	97.19	96.62	96.16	96.55	96.36	96.37	21.38
15 dB	91.99	93.77	93.59	90.99	92.59	93.28	93.65	94.54	94.20	93.92	93.21	94.14	93.68	93.34	42.90
10 dB	84.46	86.25	86.82	82.94	85.12	85.26	86.76	89.50	87.44	87.24	83.14	86.55	84.85	85.91	54.53
5 dB	70.56	69.71	71.88	66.28	69.61	68.10	71.95	75.16	72.85	72.02	66.66	70.80	68.73	70.40	50.23
0 dB	48.27	41.96	43.66	41.65	43.89	41.17	47.64	52.01	45.02	46.46	40.19	44.20	42.20	44.58	32.84
−5 dB	21.61	16.63	18.76	17.12	18.53	17.68	21.74	23.17	19.93	20.63	17.81	19.26	18.54	19.37	11.78
Average	78.25	77.69	78.44	75.47	77.46	76.90	79.18	81.59	79.34	79.25	75.87	78.45	77.16	78.12	
	28.71%	55.49%	45.29%	29.12%	41.70	51.27%	45.89%	60.61%	53.44%	53.11%	28.69%	36.39%	32.54		45.21

Table 6. Common tests for some sub-corpora. The test procedures are SVIP (Small Vocabulary Isolated Phrase) and SVWL (Small Vocabulary Word Loop).

Test	Recognition task	Procedure
I	Isolated digits	SVIP
A	30 isolated application words	SVIP
BC	Connected digit strings, unknown length	SVWL

because these were difficult to score without a particular application dialogue in mind. Noise markers were ignored, but utterances with non-speech acoustic events were kept in the test.

In Tables 7–9 the speech recognition test results (WER-word error rate) are shown with triphone models for sub-corpuses I, A and BC. The number of Gaussian mixture PDFs (probability density functions) increased from 1 to 32 in steps to the power of 2. The first columns in all the tables show the baseline results achieved with the ETSI ES 201 108 front-end. The middle column represents the WER achieved by the proposed feature extraction procedure TD-FrAtt (see Fig. 2 and description in Section 3), and the third column shows the achieved WER with the proposed front-end FD-FrAtt.

Table 7. Comparison of WER achieved by different frontends with sub-corpus I.

	Speech recognition results (WER) (%)		
Numb. of Gauss. PDFs	Baseline	TD-FrAtt	FD-FrAtt
1	6.68	4.66	5.18
2	6.33	5.70	4.66
4	5.72	4.15	5.18
8	5.63	3.11	4.15
16	5.70	4.15	4.15
32	4.15	3.11	3.63

Table 8. Comparison of WER achieved by different frontends with sub-corpus A.

	Speech recognition results (WER) (%)		
Numb. of Gauss. PDFs	Baseline	TD-FrAtt	FD-FrAtt
1	7.03	6.01	3.65
2	6.37	4.88	3.93
4	5.69	4.51	3.46
8	5.64	3.66	3.00
16	3.93	3.38	2.90
32	4.11	3.00	2.72

Table 9. Comparison of WER achieved by different frontends with sub-corpus BC.

	Speech recognition results (WER) (%)		
Numb. of Gauss. PDFs	Baseline	TD-FrAtt	FD-FrAtt
1	6.77	7.73	5.06
2	6.33	6.05	4.25
4	5.81	4.57	3.27
8	5.60	4.25	2.57
16	4.95	3.73	2.32
32	4.28	3.48	1.86

5. Discussion

The speech recognition results achieved on Aurora 2 and Slovenian SpeechDat II databases with TD-FrAtt and FD-FrAtt are discussed and compared in subsections 5.1 and 5.2.

5.1. Aurora 2 Database Recognition Results

Aurora 2 experiments and tests have shown that, compared to the baseline system—the current Aurora front-end standard for distributed speech recognition, both described feature extraction algorithms TD-FrAtt and FD-FrAtt increase word recognition accuracy. The biggest recognition improvement was achieved when the system was trained on clean data irrespective of TD-FrAtt or FD-FrAtt. The achieved improvement with the TD-FrAtt was 51.52%, and the achieved improvement with the FD-FrAtt was 45.21%, relative to the baseline system. In the clean training the mismatch between train and test conditions is very high, so the improvement with the clean training procedure is higher, compared to multiconditional training. When the multiconditional training was performed, the improvement was 12.20% and 26.09% for TD-FrAtt and FD-FrAtt, respectively. When compared for particular conditions, the improvement was smaller in the cases of highly colored non-stationary noises (babble, restaurant) and multiconditional training. These colored noises have a non-white predominantly low frequency spectrum. Spectral energy distribution of colored noises is very similar to the spectral energy distribution of vowels. This feature is the main reason why speech recognition

efficiency decreases more in environments with colored noise characteristics, as compared to those with white noise characteristics (car, subway), even though the SNR is equal in both cases. It is well known that recognition improvement deteriorates with decreasing SNR. Both proposed methods proved to be efficient at SNRs between 0 and 20 dB.

As compared to FD-FrAtt, TD-FrAtt achieves a greater recognition improvement when clean training is performed. It can be seen that TD-FrAtt is more appropriate in situations, where the train and test conditions do not match. Namely, the zero crossing rate of human speech, and particularly colored noise, are very similar at lower SNRs. Due to zero crossing computation in the TD-FrAtt, more accurate acoustical modelling with clean-only speech data is achieved. The opposite is true when multiconditional training results are compared. In this case the mismatch is lower and FD-FrAtt achieves better performance. This statement is confirmed when the Slovenian SpeechDat II database recognition results are analysed.

Tables 10 and 11 show the automatic speech recognition performance summary for the time-domain frame-dropping (TD-FrDrop) and frequency-domain frame-dropping algorithms (FD-FrDrop), respectively. The results show that the frame-attenuation algorithm is more efficient than the frame-dropping algorithm, irrespective of the training mode.

In very low SNR conditions a VAD may result in an incorrect speech/non-speech decision. If this incorrect VAD decision is used in the frame-dropping algorithm, a corrupted speech parameterization is performed. The frame-attenuation principle does not emphasize this VAD error. The cepstral coefficients C_1–C_{12} are generally left unchanged after frame-attenuation. Actually, only the 13th feature vector element is affected in the frame-attenuation procedure.

Table 10. Aurora 2 database recognition results for the feature extraction algorithm TD-FrDrop and multiconditional training, and performance relative to the current ETSI ES 201 108 standard.

Absolute performance				
Training mode	Set A	Set B	Set C	Overall
Multicondition (%)	88.88	86.73	85.29	87.30
Clean only (%)	80.01	79.31	74.17	78.56
Average (%)	84.45	83.02	79.73	82.93
Performance relative to Mel-cepstrum				
Multicondition (%)	8.77	3.31	9.33	6.70
Clean only (%)	48.28	53.24	23.72	46.31
Average (%)	28.53	28.28	16.53	26.51

Table 11. Aurora 2 database recognition results for the feature extraction algorithm FD-FrDrop and multiconditional training, and performance relative to the current ETSI ES 201 108 standard.

Absolute performance				
Training mode	Set A	Set B	Set C	Overall
Multicondition (%)	89.58	89.03	87.42	88.93
Clean only (%)	76.47	77.98	75.01	76.78
Average (%)	83.02	83.51	81.22	82.86
Performance relative to Mel-cepstrum				
Multicondition (%)	14.49	20.09	22.44	18.64
Clean only (%)	39.13	50.25	26.21	41.86
Average (%)	26.81	35.17	24.32	30.25

The experiments on the Aurora 2 database are also appropriate for comparison of achievements of different researchers in the field of automatic speech recognition (Yapanel et al., 2001). At the Eurospeech 2001 conference many different techniques for noise robustness were compared in an Aurora 2—Special session.

5.2. *Slovenian SpeechDat II Database Recognition Results*

The experiments and tests on the Slovenian SpeechDat II database have shown that both proposed algorithms significantly decrease the word recognition error rates, when compared to the baseline tests. The best achieved word recognition error rates for all three sub-corpuses I, A and BC, when triphone models with 32 Gaussian mixture PDFs were used, are compared in Table 12.

Table 12 shows that the TD-FrAtt algorithm provides better performance with sub-corpus I than does FD-FrAtt (With TD-FrAtt, a 0.52% lower WER is achieved absolute to the FD-FrAtt). TD-FrAtt underperformed FD-FrAtt with the remaining sub-corpuses, A and BC. The obtained results also show that the

Table 12. Comparison of the best achieved recognition results (word error rates).

Sub-corpus	Baseline (%)	TD-FrAtt (%)	FD-FrAtt (%)
I	4.15	3.11	3.63
A	4.11	3.00	2.72
BC	4.28	3.48	1.86

proposed methods give high recognition accuracy at lower speech recognition system complexity. When using FD-FrAtt with sub-corpus BC, for example, a word recognition error rate of 4.25% at two Gaussian mixture PDFs was achieved (see Table 9). A similar result (WER of 4.28%) was achieved by the baseline system configuration when using the 32 Gaussian mixture PDFs. Here we can see that better word recognition accuracy was achieved at 16 times lower computational complexity (processing time, memory consumption, etc.). With the FD-FrAtt algorithm at higher model complexity (32 Gaussians) with BC sub-corpus, the lowest WER of 1.86% was achieved. Due to overtraining of acoustical models, sometimes fluctuations in the WER of speech recognition results can occur. Even when model complexity has been increased, the WER does not decrease proportionally. This WER anomaly can be seen especially in Table 7, where the test results with sub-corpus I are presented.

It can be summarized, from the automatic speech recognition results, that the FD-FrAtt outperformed the TD-FrAtt when the train and test conditions matched. The matched conditions were at multiconditional training with the Aurora II database and also with the Slovenian SpeechDat II database.

6. Conclusion

This paper presents two robust front-end algorithms TD-FrAtt and FD-FrAtt (see Fig. 2). Both algorithms include a noise reduction procedure based on minimum statistic spectral subtraction and two different forms of frame-attenuation. In TD-FrAtt, a time-domain frame-attenuation was used, based on a special frame weighting function. This front-end algorithm resulted in the best performance on the Aurora 2 database when training on clean data was applied (improvement of 51.52% relative to the baseline system) as well as on the Slovenian SpeechDat II fixed database with sub-corpus I (WER of 3.11%). The FD-FrAtt algorithm exploits frequency-domain voice activity detection based on short and long term log-spectral energy estimation combined with frame-attenuation. The front-end algorithm FD-FrAtt produced its best performance on the Aurora 2 database when training on multiconditional data was performed (improvement of 26.09% relative to the baseline system), as well as on the Slovenian SpeechDat II database with A and BC sub-corpuses (WER of 2.72% and 1.86% for sub-corpuses A and BC, respectively).

It can be seen that TD-FrAtt and FD-FrAtt achieve different recognition rates in different training modes. This implies that TD-FrAtt and FD-FrAtt are suitable for different distributed speech recognition system configurations. For those DSR applications where a noisy speech database exists from which to produce acoustical models, it is a better choice to use the FD-FrAtt front-end algorithm, which has good performance for matched conditions. The use of TD-FrAtt in a distributed speech recognition system is preferred in those cases where noisy speech databases are unavailable and only an unmatched speech database (e.g., speech database recorded in a studio) exists for acoustical model generation.

The results obtained using the Aurora 2 and Slovenian SpeechDat II databases show that both proposed methods give higher recognition accuracy at lower SNRs, as well as at lower speech recognition system complexity, when compared to baseline systems, and are appropriate for usage in distributed speech recognition systems.

Acknowledgments

The authors would like to thank Siemens AG (Germany) for providing the Slovenian SpeechDat II fixed telephone speech database for evaluation of the proposed algorithms.

This work was funded by the Ministry of Education, Science and Sport of the Republic of Slovenia under Grant PP-0796/99.

References

Andrassy, B., Vlaj, D., and Beaugeant, C. (2001). Recognition performance of the siemens front-end with and without frame dropping on the Aurora 2 database. *EUROSPEECH 2001 Proceedings*. Aalborg, Denmark, pp. 193–196.

Benitez, C., Burget, L., Chen, B., Dupont, S., Garudadri, H., Hermansky, H., Jain, P., Kajarekar, S., Morgan, N., and Sivadas, S. (2001). Robust ASR front-end using spectral-based and discriminant features: Experiments on the Aurora tasks. *EUROSPEECH 2001 Proceedings*. Aalborg, Denmark, pp. 429–432.

Boll, S.F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *27*(2):113–120.

COST 249 SpeechDat SIG (2000). The RefRec Homepage. http://www.telenor.no/fou/prosjekter/taletek/refrec/

Deller, J.R., Proakis, J.G., and Hansen, J.H.L. (1993). *Discrete-Time Processing of Speech Signals*. New York, USA: Macmillan Publishing Company.

ETSI standard document (2000). Speech processing, transmission and quality aspects (STQ), distributed speech recognition,

front-end feature extraction algorithm, compression algorithm. ETSI ES 201 108 v1.1.1 (2000-02). Sophia Antipolis, France.

ETSI-SMG technical specification (1994). European digital cellular telecommunication system (Phase 1)—Transmission planning aspects for the speech service in GSM PLMN system—GSM03.50, version3.4.0. Sophia Antipolis, France.

Hirsch, H.G. and Pearce, D. (2000). The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. *ISCA ITRW ASR 2000 Proceedings*. Paris, France.

ITU recommendation G.712 (1996). Transmission performance characteristics of pulse code modulation channels. Geneva, Switzerland.

ITU recommendation G.723.1 A (1996). Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s. Annex A: Silence compression scheme. Geneva, Switzerland.

Junqua, J.-C. and Haton, J.-P. (1996). *Robustness in Automatic Speech Recognition*. Kluwer Academic Publishers. Norwell, Massachusetts, USA.

Kaiser, J. and Kacic, Z. (1997). *SpeechDat II Slovenian Database for the Fixed Telephone Network*. Maribor, Slovenia: University of Maribor.

Kotnik, B., Rotovnik, T., Kacic, Z., and Horvat, B. (2001a). The design of mobile multimodal communication device—personal navigator. *EUROCON 2001 Proceedings*, Bratislava, Slovakia, pp. 337–340.

Kotnik, B., Kacic, Z., and Horvat, B. (2001b). A Multiconditional Robust Front-End Feature Extraction with a Noise Reduction Procedure Based on Improved Spectral Subtraction Algorithm. *EUROSPEECH 2001 Proceedings*. Aalborg, Denmark, pp. 197–200.

Leonard, R.G. (1991). *A Speaker-Independent Connected-Digit Database*. Texas Instruments Inc., Dallas, Texas, USA,

Lindberg, B., Johansen, F.T., Warakagoda, N., Lehtinen, G., Kacic, Z., Zgank, A., Elenius, K., and Salvi, G. (2000). A noise robust multilingual reference recogniser based on SpeechDat II. *ICSLP 2000 Proceedings*. Beijing, China. Paper No. 01775.

Martin, R. (1994). Spectral subtraction based on minimum statistics. *EUSIPCO 1994 Proceedings*. Edinburgh, Scotland, UK. pp. 1182–1185.

Oviatt, S. (2000). Multimodal signal processing in naturalistic noisy environments. *ICSLP 2000 Proceedings*. Beijing, China, pp. 696–699.

Pearce, D. (2000). An overview of the ETSI standards activities for distributed speech recognition front-ends. *AVIOS 2000 Proceedings*. San Jose, CA, USA.

Van den Heuvel, H., Boves, L., Moreno, A., Omologo, M., Richard, G., and Sanders, E. (2001). Annotation in the SpeechDat projects. *International Journal of Speech Technology*, 4(2):127–143.

Varga, A.P. and Moore, R.K. (1990). Hidden Markov model decomposition of speech and noise. *ICASSP 1990 Proceedings*. Albuquerque, New Mexico, USA, pp. 845–848.

Yapanel, U., Hansen, J.H.L., Sarikaya, R., and Pellom, B. (2001). Robust digit recognition in noise: An evaluation using the AURORA Corpus. *EUROSPEECH 2001 Proceedings*. Aalborg, Denmark, pp. 209–212.

Young, S. (1997). *HTK Book—Version 2.1*, Cambridge, UK: Entropic Cambridge Research Laboratory.