ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/263925154

Speech Enhancement of Movie Sound

Conference Paper · January 2008

CITATIONS
6

READS
29

3 authors, including:

Christian Uhle
Fraunhofer Institute for Integrated Circuits IIS
26 PUBLICATIONS **324** CITATIONS

SEE PROFILE

All content following this page was uploaded by Christian Uhle on 01 February 2019.

The user has requested enhancement of the downloaded file.

Audio Engineering Society

Convention Paper

Presented at the 125th Convention
2008 October 2–5 San Francisco, CA, USA

This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Speech enhancement of movie sound

Christian Uhle[1], Oliver Hellmuth[1], and Jan Weigel[1]

[1] *Fraunhofer Institute for Integrated Circuits, Erlangen, Germany*

Correspondence should be addressed to Christian Uhle (christian.uhle@iis.fraunhofer.de)

ABSTRACT

Today, many people have problems understanding the speech content of a movie, e.g. due to hearing impairments. This paper describes a method for improving the speech intelligibility of movie sound. Speech is detected by means of a pattern recognition method; the audio signal is then attenuated during periods where speech is absent. The speech signals are further processed by a spectral weighting method aiming at the suppression of the background noise. The spectral weights are computed by means of feature extraction and a neural network regression method. The output signal finally carries all relevant speech with reduced background noise allowing the listener to follow the plot of the movie more easily. Results of numerical evaluations and of listening tests are presented.

1. INTRODUCTION

Movies tell stories, and beside the visual and acoustic impressions it is the plot of the movie that is of major interest to the moviegoer. To follow the plot it is important to understand the relevant speech of the audio track, e.g. monologues, dialogues, announcements and narrations. People who are hard of hearing often experience that background sounds, e.g. environmental noise and music are presented at too high level with respect to the speech. In this case, it is desired to increase the level of the speech signals and to attenuate the background sounds.

This is why "Services for Hearing Impaired" are an essential part of the ongoing European Integrated Project "Enhanced Digital Cinema" (EDCine, IST-038454) [1] aiming at the optimization and enhancement of the digital cinema of the future. People with normal hearing capabilities may benefit from such processing as well. A movie experience in the home cinema late at night is often a compromise between understanding the actors and not disturbing the family or neighbors. Dynamic compression is a valid solution to this problem. However, the superior solution is to increase the level of the speech signals in the movie sound. As the separated speech signals are usually not available in today's content distribution formats, additional speech detection and enhancement techniques are required to allow for such an enhanced "late night (listening)

mode".

This paper describes a method to enhance the speech and attenuate all non-speech components by applying speech detection and speech enhancement. The relevant components are detected by means of a pattern recognition system. The detected speech signals are then processed using a spectral weighting method to attenuate the background noise. The spectral weights are computed using a neural network regression method. The output signal finally carries an enhanced version of all relevant speech signal components of the original audio material.

The contribution of this paper is twofold: it investigates speech detection methods for the processing of movie sound and presents a novel approach to noise estimation for speech enhancement in non-stationary conditions. It is organized as follow: Section 2 presents the state of the art in speech detection and speech enhancement. Section 3 describes the proposed method, which is divided into the two separate processes, namely speech detection and speech enhancement. The results of the evaluations are presented in Section 4 and conclusions are given in Section 5.

2. BACKGROUND

Movie sound is a composite of signals of different characteristic and origin, e.g. speech, music, environmental noise and effect sounds. A particular feature of the speech in movie sound compared to other speech signals (e.g. newscasts) is the variety of speech sounds, including expressive, whispered and shouted speech sounds as well disguised voices. Words may occur isolated or are spoken with speeds ranging from very slow to very fast. From the point of view of speech enhancement, the speech is the desired signal $s[k]$ and all other signals are considered as background noise $b[k]$. In the following the movie sound signal $x[k]$ is assumed to be an additive mixture $x[k] = s[k] + b[k]$.

The application described in this paper combines two digital speech processing methods, namely speech detection and speech enhancement. The review of the state of the art is focused on the processing of single-microphone observations due to the following reasons:

- A huge supply of legacy content produced in mono exists.
- The audio signals of a movie produced in stereo are often highly correlated.
- The center channel of multi-channel surround movie sound contains often less background noise compared to the other channels and is therefore an appropriate input signal for the processing described here.

2.1. Speech detection - State of the art

The task of speech detection is to analyze an audio signal and to decide whether speech is present or not. Various methods to speech detection are reported in the literature with applications to speech enhancement, information retrieval, audio coding and automated speech recognition. In general, these methods implement the pattern recognition approach to speech detection as illustrated in Figure 1. Pattern recognition is "the act of taking in raw data and taking an action based on the category of the pattern" [2], whereas the term *pattern* describes an underlying similarity which can be found between features of the objects from different categories. This approach is valid for a diversity of classification problems and can be adapted to the specific task by an appropriate selection of the features. These features are chosen to be very similar among objects of the same class (intra-class compactness) and very different among objects of different classes (inter-class separability). A third requirement inherited from the first and the second is that the features should be robust with respect to noise and to "transformations" of the input signal which do not affect the affiliation to a class, e.g. rotations in the context of optical character recognition or moderate band-pass filtering in the context of speaker identification.

A variety of features has been applied in previous work on speech detection, e.g. based on the formant shape [3], the skewness of the distribution of the zero-crossing rate [4], line-spectral frequencies [5], and amplitude modulation features [6, 7, 8, 9]. Specific combinations of features [10, 11] or features derived from specific signal representations [12, 13] were reported. The posterior probability of a speech recognizer is used in [14]. In general, the features

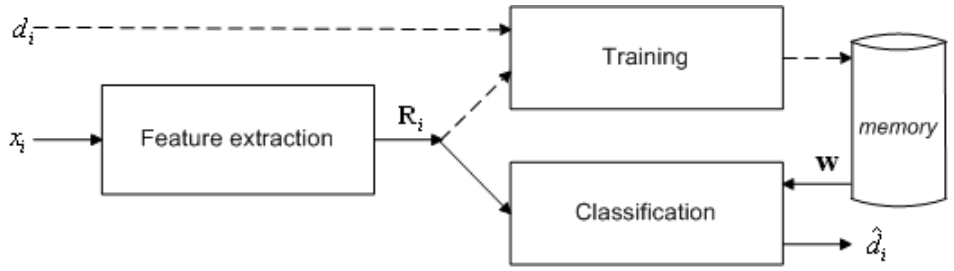


Fig. 1: Block diagram of the pattern recognition system. The objects x_i are classified into classes labelled $\widehat{d}_i$ using a set of features $\mathbf{R}_i$. The parameter set $\mathbf{w}$ of the classifier are obtained during a training process (shown by dotted lines) from examples with known labels d_i and are stored in memory. The parameters are then used to classify novel patterns (shown by solid lines).

used for other tasks of audio content classification have been applied successfully, e.g. in [15, 16].

Many of the features mentioned above are computed on a frame-by-frame basis with a frame size between 10 and 32 milliseconds. On the other hand, the detection of speech largely depends on the temporal evolution of the signal characteristics over durations exceeding the length of this short-term spectral analysis (for both, human listeners and machines). Harb and Chen report that humans are able to classify speech easily for durations of about 200 milliseconds [17]. Consequently, the relation between neighboring frames is important for speech detection and is incorporated by evaluating features of neighboring frames using statistical moments of the features (e.g. in [4, 15]), histograms [11], delta features [16], or by subsuming the classification results of neighboring frames (e.g. in [3]). Some features capture information related to the temporal evolution of signal characteristics, e.g. spectro-temporal modulations (STM).

2.2. Speech enhancement - State of the Art

Speech enhancement is the improvement in the objective intelligibility and/or subjective quality of speech. A prominent approach to speech enhancement is spectral weighting, also referred to as short-term spectral attenuation, as illustrated in Figure 2. A frequency domain representation of the input signal is computed by means of a Short-term Fourier Transform (STFT), other time-frequency transforms or a filter bank. The input signal is than filtered in the frequency domain according to Equation 1,

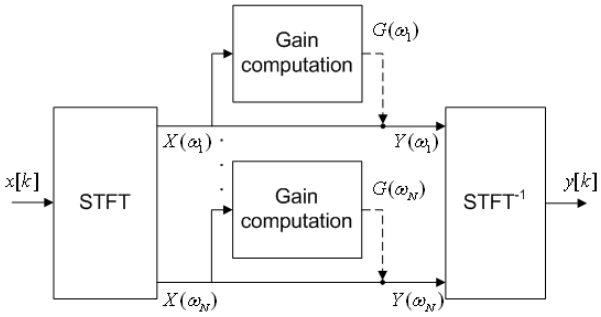


Fig. 2: Block diagram of the general spectral weighting approach. The output signal $y[k]$ is computed by attenuating the sub-band signals $X(\omega)$ of the input signal $x[k]$ depending on the noise energy within the sub-band signals.

whereas the frequency response $G(\omega)$ of the filter is computed such that the noise energy is reduced. The output signal is computed by means of the inverse processing of the time-frequency transforms or filter bank, respectively.

$$Y(\omega) = G(\omega)X(\omega) \tag{1}$$

Appropriate spectral weights $G(\omega)$ are computed using the input signal spectrum $X(\omega)$ and an estimate of the noise spectrum $\widehat{B}(\omega)$ or, equivalently, using an estimate of the linear sub-band SNR $\widehat{R}(\omega)$. Prominent examples of noise suppression rules are spectral subtraction [18] and Wiener filtering. Assuming that the input signal is an additive mixture of the speech

and the noise signals and that speech and noise are uncorrelated, the gain values for the spectral subtraction method are given in Equation 2.

$$G(\omega) = \sqrt{1 - \frac{\left|\widehat{B}(\omega)\right|^2}{\left|X(\omega)\right|^2}} \quad (2)$$

Similar weights are derived from estimates of the linear sub-band SNR $\widehat{R}(\omega)$ according to Equation 3.

$$G(\omega) = \sqrt{\frac{\widehat{R}(\omega)}{\widehat{R}(\omega) + 1}} \quad (3)$$

Various extensions to spectral subtraction have been proposed in the past, namely the use of an over-subtraction factor and spectral floor parameter [19], generalized forms [20], the use of perceptual criteria (e.g. [21]) and multi-band spectral subtraction (e.g. [22]). However, the crucial part of a spectral weighting method is the estimation of the instantaneous noise spectrum or of the sub-band SNR, which is prone to errors especially if the noise is non-stationary. Errors of the noise estimation lead residual noise, distortions of the speech components or musical noise (an artifact which has been described as "warbling with tonal quality" [23]).

A simple approach to noise estimation is to measure and averaging the noise spectrum during speech pauses. This approach does not yield satisfying results if the noise spectrum varies over time. Methods for estimating the noise spectrum even during speech activity can be classified according to [23] as

- Minimum tracking algorithms
- Time-recursive averaging algorithms
- Histogram based algorithms

The estimation of the noise spectrum using minimum statistics has been proposed in [24]. The method is based on the tracking of local minima of the signal energy in each sub-band. A non-linear update rule for the noise estimate and faster updating has been proposed in [25]. Time-recursive averaging algorithms estimate and update the noise spectrum whenever the estimated SNR at a particular frequency band is very low. This is done by computing recursively the weighted average of the past noise estimate and the present spectrum. The weights are determined as a function of the probability that speech is present or as a function of the estimated SNR in the particular frequency band, e.g. in [26, 27]. Histogram-based methods rely on the assumption that the histogram of the sub-band energy is often bimodal. A large low-energy mode accumulates energy values of segments without speech or with low-energy segments of speech. The high-energy mode accumulates energy values of segments with voiced speech and noise. The noise energy in a particular sub-band is determined from the low-energy mode [28]. For a comprehensive recent review it is referred to [23].

Methods for the estimation of the sub-band SNR based on supervised learning using amplitude modulation features are reported in [29, 30].

Other approaches to speech enhancement are pitch-synchronous filtering (e.g. in [31]), filtering of STM (e.g. in [32]), and filtering based on a sinusoidal model representation of the input signal (e.g. [33]).

3. SYSTEM DESCRIPTION

3.1. Overview

The signal processing presented in this work comprises a speech detection method and a speech enhancement method. Figure 3 illustrates an overview: The first processing detects the occurrences of relevant speech signals. The term "relevant" relates to speech signals which are necessary to follow the plot of the movie. Other speech sounds (for example a crowd of people speaking simultaneously in the background) are considered as background noise. If no relevant speech components are detected, the audio signal is attenuated (e.g. by 12 dB or more). The detected speech signals are processed by the speech enhancement processing to attenuate the corrupting background noise. Finally, the processed audio signal can be rendered at an increased output level since the loud non-speech signal components (e.g. gun shots, explosions or car accidents) are attenuated.

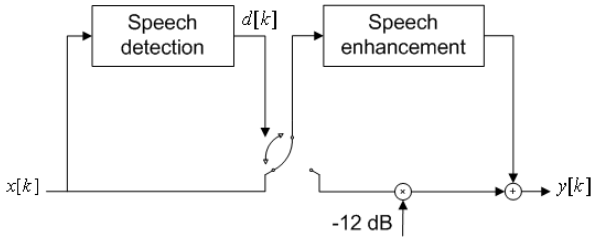


Fig. 3: Overview of the signal processing, with input signal $x[k]$, output signal $y[k]$, and speech detection result $d[k]$. The attenuation factor 12 dB is arbitrarily chosen and for illustrative reasons only.

The input signal is processed in the frequency domain. The STFT of overlapping data frames of 32 milliseconds length each is computed by means of the Discrete Fourier Transform with zero-padding, a block size of 64 milliseconds, an overlap of 8 ms and the Hann window function.

3.2. Speech detection

The speech components are detected by means of a pattern recognition system as shown in Figure 1. Various low-level features and the corresponding delta and delta-delta features have been investigated with respect to their abilities to discriminate speech and non-speech. Features derived from a small number of successive signal frames are subsumed to larger entities (in the following denoted as *group*). The groups are represented by the means and the variances of the feature values computed from the respective frames. The reference value for each group is computed as the median of the references of the corresponding frames.

The features are optionally post-processed by means of centering, variance normalization and Linear Discriminant Analysis.

Different classifiers are compared in this work:

- Linear classifier derived from the Fischer criterion (FLD)
- Gaussian mixture model with one component for each class derived using maximum likelihood estimation (GMM1)
- Gaussian mixture model with two components for each class derived using expectation maximization (GMM2)
- Feed-forward neural network (NN).

These classifiers represent discriminative functions of different complexity and allow extensive investigations of different parameter settings due to their fast training.

Experiments with different feature sets were carried out. The first feature set comprises various low-level features (LLF) computed from the spectral coefficients which correspond to 4 logarithmically spaced frequency bands in the range between 100 Hz and 8000 Hz. The features used in this work are the spectral flatness measure, the spectral flux, the spectral skewness, and the 4 Hz modulation energy. These features are frequently applied to audio content classification.

Another feature set is computed from the STM energy in critical bands. Features derived from STM have been successfully applied to speech detection, and they are robust with respect to stationary background noise [9].

Other feature sets comprise linear prediction coefficients (LPC) [34], mel-frequency cepstral coefficients (MFCC) [35], perceptual linear prediction (PLP) coefficients [36] and relative spectra perceptual prediction (RASTA-PLP) [37] coefficients. The features are described in Section 3.4.

3.3. Speech enhancement

The proposed method follows the well-known approach of spectral weighting but uses a novel method for the computation of the spectral weights. The noise estimation is based on a supervised learning method as described in [29, 30] but uses a different feature set. The features used in this work aim at the discimination of tonal versus noisy signal components (similarily to the sub-band amplitude modulation in the range 50-400 Hz as used in [29]). Additionally, the proposed features take the evolution of signal properties on a larger time scale into account (like the slow spectro-temporal modulations used in [30]).

The noise estimation method presented here is able to deal with a variety of nonstationary background sounds. A robust SNR estimation in non-stationary background noise is obtained by means of feature extraction and a neural network regression method as

illustrated in Figure 4. The real-valued weights are computed from estimates of the SNR in frequency bands whose spacing approximates the Bark scale. The spectral resolution of the SNR estimation is rather coarse in comparison to many other methods. However, a similar resolution has been applied successfully in previous works [29, 38] and is required for the proposed method with respect to computational complexity.

3.3.1. Feature extraction

A set of 21 different features has been investigated in order to identify the best feature set for the estimation of the sub-band SNR. These features were combined in various configurations and were evaluated by means of objective measurements as described in Section 4.2.1 and informal listening. The feature selection process results in a feature set comprising the spectral energy, the spectral flux, the spectral flatness, the spectral skewness, the LPC and the RASTA-PLP coefficients. The spectral energy, flux, flatness and skewness features are computed from the spectral coefficient corresponding to the critical band scale [39].

The features are detailed in Section 3.4. Additional features are the delta feature of the spectral energy and the delta-delta feature of the low-pass filtered spectral energy and of the spectral flux.

3.3.2. Structure of the neural network

The neural network is applied for the estimation of the sub-band SNR from the computed low-level features. The neural network has 220 input neurons (corresponding to the number of features) and one hidden layer with 50 neurons. The number of output neurons equals the number of frequency bands. The activation function of the hidden neurons is the hyperbolic tangent, the activation function of the output neurons is the identity.

3.3.3. Training of the neural network

The weights of the neural network are trained on mixtures of clean speech signals and background noises whose reference SNR are computed using the separated signals. The training process is illustrated on the left hand side of Figure 4. Speech and noise is mixed with an SNR of 3 dB per item and fed into the feature extraction. The data set comprises 2304 combinations of 48 speech signals and 48 noise signals of 2.5 seconds length each. The speech signals originated of different speakers with 7 languages. The noise signals are recordings of traffic noise, crowd noise, and various natural atmospheres.

For a given spectral weighting rule, two definitions of the output of the neural network are appropriate: The neural network can be trained using the reference values for the time-varying sub-band SNR $R(\omega)$ or with the spectral weights $G(\omega)$ (derived from the SNR values). Simulations with sub-band SNR as reference values yielded better objective results and better ratings in informal listening compared to nets which were trained with spectral weights. The neural network is trained using 100 iteration cycles. A training algorithm from the *NetLab* toolbox [40] is used in this work, which is based on scaled conjugate gradients.

3.3.4. Spectral weighting

The estimated sub-band SNR estimates are linearly interpolated to the frequency resolution of the input spectra and transformed to linear ratios $\widehat{R}$. The linear sub-band SNR are smoothed along time and along frequency using IIR low-pass filtering to reduce artifacts which may result from estimation errors. The low-pass filtering along frequency is further needed to reduce the effect of circular convolution which occurs if the impulse response of the spectral weighting exceeds the length of the DFT frames. It is performed twice, whereas the second filtering is done in reversed order (starting with the last sample) and the resulting filter has zero phase.

The spectral weights are computed according to the modified spectral subtraction rule in Equation 4 and limited to -18 dB.

$$G(\omega) = \begin{cases} \frac{\widehat{R}(\omega)^{\alpha}}{\widehat{R}(\omega)^{\alpha}+1} & | \quad \widehat{R}(\omega) \leq 1 \\ \frac{\widehat{R}(\omega)^{\beta}}{\widehat{R}(\omega)^{\beta}+1} & | \quad \widehat{R}(\omega) > 1 \end{cases} \tag{4}$$

The parameters $\alpha = 3.5$ and $\beta = 1$ are determined experimentally. This particular attenuation above 0 dB SNR is chosen in order to avoid distortions of the speech signal at the expense of residual noise. The attenuation curve as a function of the SNR is illustrated in Figure 5.

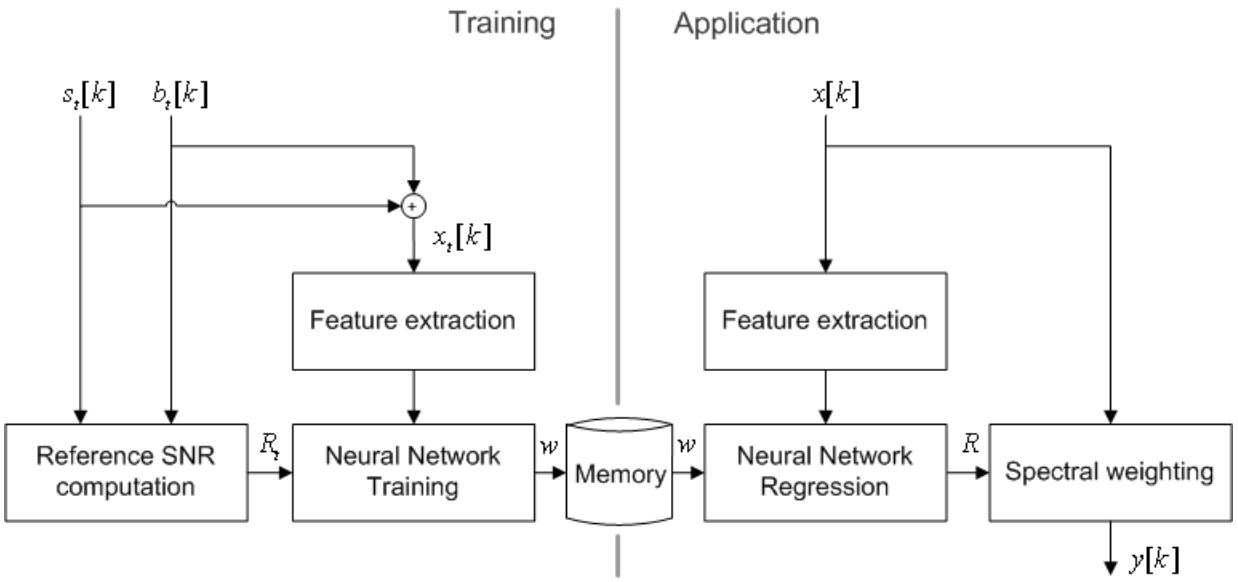


Fig. 4: Overview of spectral weighting processing using feature extraction and a neural network regression method. The parameters w of the neural network are computed using the reference sub-band SNR R_t and features from the training items $x_t[k]$ during the training phase (left hand side). The noise estimation and reduction is shown on the right hand side.

Figure 6 shows an example for the input and output signals, the estimated sub-band SNR and the spectral weights.

3.4. Description of the features

3.4.1. Spectral energy

The spectral energy is computed for each time frame and frequency band and normalized by the total energy of the frame. Additionally, the spectral energy is low-pass filtered over time using a second-order IIR filter.

3.4.2. Spectral flux

The spectral flux SF is defined as the dissimilarity between spectra of successive frames [41] and is frequently implemented by means of a distance function. In this work, the spectral flux is computed using the Euclidian distance according to Equation 5, with spectral coefficients $X(m,k)$, time frame index m, sub-band index r, lower and upper boundary of the frequency band l_r and u_r, respectively.

$$SF(m,r) = \sqrt{\sum_{q=l_r}^{u_r} (|X(m,q)| - |X(m-1,q)|)^2} \tag{5}$$

3.4.3. Spectral flatness measure

Various definitions for the computation of the flatness of a vector or the tonality of a spectrum (which is inversely related to the flatness of a spectrum) exist, e.g. in [42, 43]. The spectral flatness measure SFM used here is computed as the ratio of the geometric mean and the arithmetic mean of the L spectral coefficients of the sub-band signal as shown in Equation 6.

$$SFM(m,r) = \frac{e^{\left(\sum_{q=l_r}^{u_r} \log(|X(m,q)|)\right)/L}}{\frac{1}{L}\sum_{q=l_r}^{u_r} |X(m,q)|} \tag{6}$$

3.4.4. Spectral skewness

The skewness of a distribution measures its asymmetry around its centroid and is defined as the third

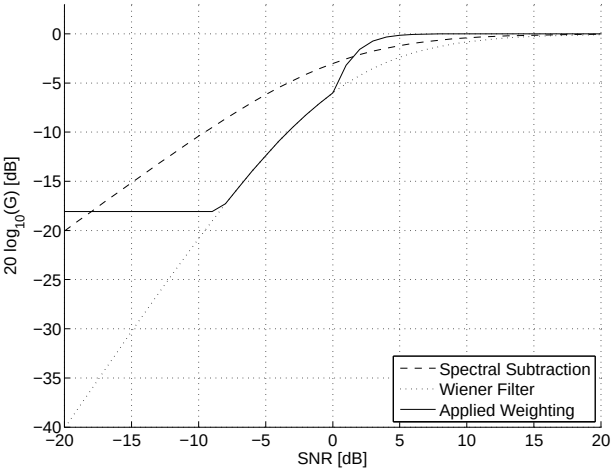


Fig. 5: Gain factor as a function of the SNR. The applied gains (solid line) are compared to spectral subtraction gains (dotted line) and the Wiener filter (dashed line).

central moment of a random variable divided by the cube of its standard deviation.

3.4.5. Linear Prediction Coefficients

The LPC are the coefficients of an all-pole filter which predicts the actual value $x(k)$ of a time series from the preceding values such that the squared error $E = \sum_k (\widehat{x}_k - x_k)^2$ is minimized.

$$\widehat{x}(k) = -\sum_{j=1}^{p} a_j x_{k-j} \tag{7}$$

The LPC are computed by means of the autocorrelation method described in [34].

3.4.6. Mel-frequency cepstral coefficients

The power spectra are warped according to the mel-scale using triangular weighting functions with unit weight for each frequency band. The MFCC are computed by taking the logarithm and computing the Discrete Cosine Transform.

3.4.7. Relative spectra perceptual prediction coefficients

The RASTA-PLP coefficients are computed from the power spectra in the following steps:

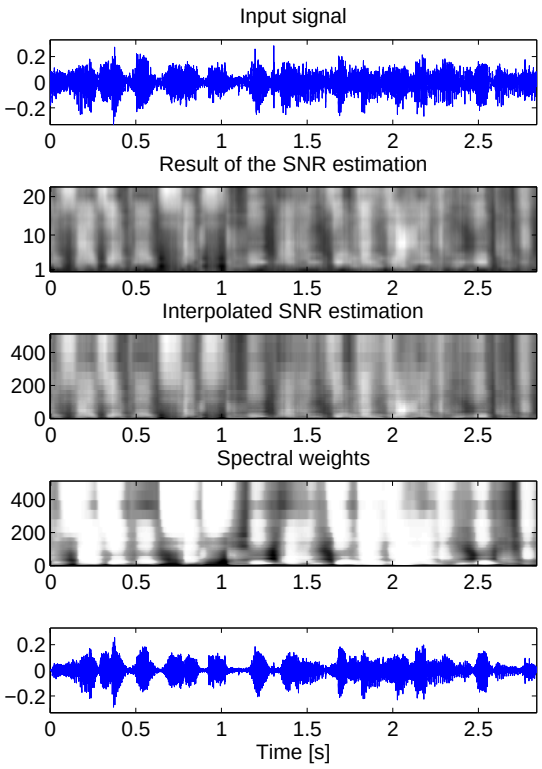


Fig. 6: Example of the spectral weighting: input time signal, estimated sub-band SNR, estimated SNR in frequency bins after interpolation, spectral weights and processed time signal.

1. Magnitude compression of the spectral coefficients
2. Band-pass filtering of the sub-band energy over time
3. Magnitude expansion which relates to the inverse processing of step 2
4. Multiplication with weights that correspond to an equal loudness curve
5. Simulation of loudness sensation by raising the the coefficients to the power of 0.33
6. Computation of an all-pole model of resulting spectrum by means of the autocorrelation method

The PLP are computed similar to the RASTA-PLP but without applying steps 1-3.

3.4.8. Spectro-temporal modulation

The STM feature is computed from the logarithmic power spectrum in critical bands. The sub-band envelope signals are derived by means low-pass filtering using a linear-phase FIR filter. The modulation energies are measured for three different modulation frequency ranges (0-3 Hz, 3-7 Hz, and 8-12 Hz) by means of band-pass filtering of the envelope signals using-second order IIR filters.

3.4.9. Delta features

Delta features have been successfully applied in automatic speech recognition and audio content classification in the past. Various ways for their computation exist. Here, they are computed by means of convolving the time sequence of a feature with a linear slope with a length of 9 samples (the sampling rate of the feature time series equals the frame rate of the STFT of 125 Hz). Delta-delta features are obtained by applying the delta operation to the delta features.

4. EVALUATION

4.1. Speech detection results

The speech detection is evaluated using a data set of 170 manually annotated movie sounds of 178 min length in total by means of 10-fold cross validation. The mean recognition rates over all runs are reported. Table 1 shows the recognition rates for different feature sets without delta features (d=0), with delta features (d=1), and with delta and delta-delta features (d=2). In these simulations, the features vectors were projected by means of an LDA onto a basis with 10 dimensions. However, very similar results were obtained without LDA for all features and classifiers. The results show that the use of delta features does not affect the recognition rate significantly. The linear classifier has shown comparable performance to more elaborate classifiers.

The influence of the grouping size on the recognition rate is shown in Figure 7. The results are obtained with RASTA-PLP coefficients and a projection using LDA.

Further experiments investigated feature sets comprising low-level features and RASTA-PLP coefficients and grouping using statistical moments of higher-order (skewness and kurtosis). However, no improvement in the recognition rates compared to the reported simulations was obtained by merging the feature sets.

	feature	FLD	GMM1	GMM2	NN
d=0	LLF	87,3	84,6	86,4	88,2
	LPC	84,0	80,2	81,2	88,4
	MFCC	89,0	86,3	88,1	89,4
	PLP	88,8	85,6	88,1	89,0
	RPLP	89,0	86,8	88,3	89,4
	STM	89,0	84,5	87,4	88,0
d=1	LLF	87,6	84,0	86,6	87,0
	LPC	83,7	80,5	81,1	84,6
	MFCC	89,1	87,1	88,4	89,6
	PLP	89,1	86,7	88,5	89,3
	RPLP	89,2	87,3	88,5	89,6
d=2	LLF	87,6	84,6	86,8	87,9
	LPC	83,2	80,3	80,7	84,6
	MFCC	88,8	86,7	88,4	89,4
	PLP	88,7	86,7	88,2	89,2
	RPLP	88,9	86,8	88,2	89,4

Table 1: Recognition rate of the speech detection in percent. The results show a comparison of different feature sets and classifiers. Compared are feature sets without delta features (d=0), with delta features (d=1), and with delta and delta-delta features (d=2).

Discussion The presented results show that an average recognition rate of nearly 90 % is obtained with various feature sets. These rates are smaller than the rates reported in previous works using other audio material, e.g. for speech/music discrimination. There are particular characteristics of movie sound which make the speech detection difficult. These are the variety of speech sounds and high background noise levels which occur frequently in movie sound.

Observations of the exemplarily chosen audio signals and detection results revealed that isolated words are difficult to detect with the presented method. Another cause of misclassification is that the segment boundaries of the speech are often not detected precisely. However, a valid solution is to apply the speech enhancement to the input signal for a small duration exceeding the detected speech region

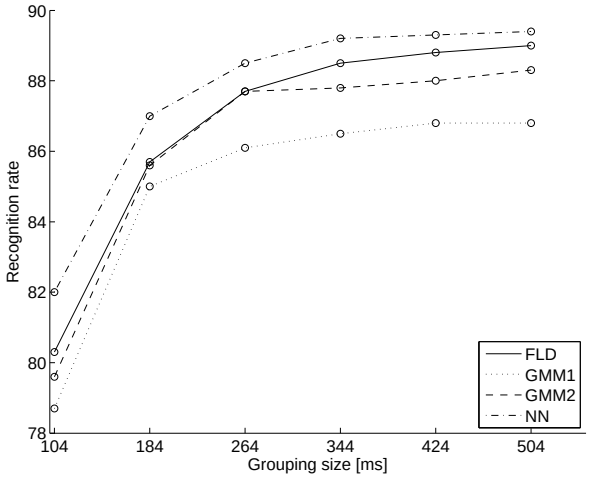


Fig. 7: Recognition rate for different grouping sizes.

at segment boundaries to account for false negative detections.

4.2. Speech enhancement results

4.2.1. Objective evaluation of the estimation of the sub-band SNR

The results of the objective evaluation are obtained from a 10-fold cross validation experiment using the synthetic mixtures of clean speech signals and background noises described in Section 5.3. The results are presented by means of

1. the mean of the absolute value of the estimation error between reference and estimates
2. the (normalized) correlation coefficient between references and estimates

The proposed method is compared to two previous noise estimation algorithms. The method C1 is a minimum tracking algorithm [25] and method C2 is a time-recursive averaging algorithm [44], whereas the implementations in [23] were used in this experiment. The methods have been chosen because they performed surprisingly well for the processing of movie sound compared to other previous methods in preliminary and informal experiments. The mean error and the correlation coefficient are shown separately for each frequency band in Figure 8.

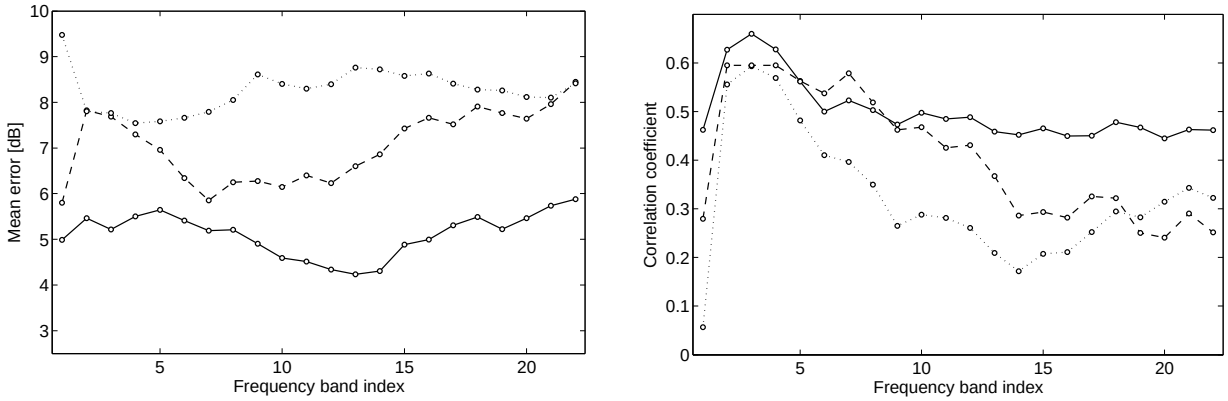
Discussion The mean error of the proposed method is about 5 dB and is smaller compared to the two previous methods under comparison. The correlation is moderate across all frequency bands for the proposed method, whereas the correlation decreases for higher frequencies for the compared methods.

4.2.2. Listening tests

The proposed method (PM) is evaluated by means of two listening tests and compared to the unprocessed audio signal and to speech enhancement using the noise estimation methods C1 and C2, as used in Section 4.2.1. The methods under comparison differ by the noise estimation only; the noise suppression rule is the standard Wiener filter rule. The noise suppression rule has been chosen to be equal for all conditions in order to make a comparison between the noise estimation methods. However, it should be noted that the results of the speech enhancement largely depends on the particular combination of both, the noise estimation and the noise suppression rule.

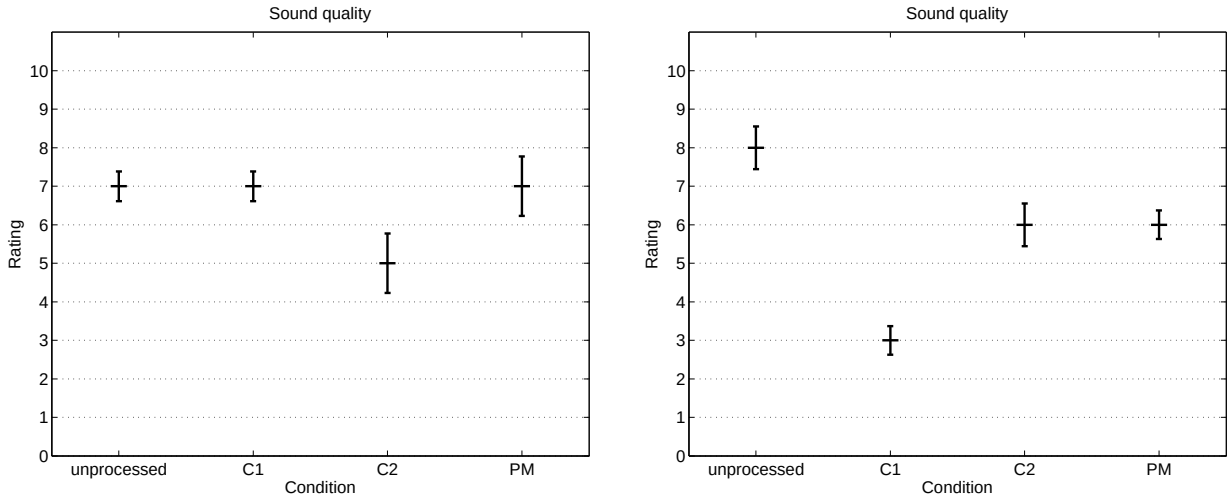
Six excerpts of movie sound with a length between 4 and 12 seconds each were presented. The unprocessed items were played first followed by the processed items in random order.

Two groups of listeners participated in the test: Group A comprises 11 hearing impaired children between 12 and 15 years old. Group B comprises 12 adult listeners with normal hearing and a professional background in audio signal processing, either as students or researcher at the institute where the listening test was carried out. The listeners of group A were not asked for information on the hearing impairments. The listeners were asked to rate the test conditions according to their personal preference with respect to *sound quality* and *speech quality*. Listeners of group B were additionally asked to rate the *noise reduction* achieved by the speech enhancement. The listeners were not asked to rate the noise reduction of the unprocessed items, since it is assumed to contain more background noise compared to the other conditions (although it should be noted that the level of the background noise can theoretically increase due to the processing, e.g. by artifacts, or it can at least be perceived by the listener as such). The ratings are expressed as a number in the range 1 to 10, where 1 relates to the lowest rating and 10 to the highest rating. The results are



(a) (b)

Fig. 8: Results of the objective evaluation of the noise estimation: (a) mean of the estimation error and (b) the correlation coefficient between the reference and the estimation. The proposed method (solid line) is compared to two previous methods C1 and C2 shown by dotted and dashed lines, respectively.



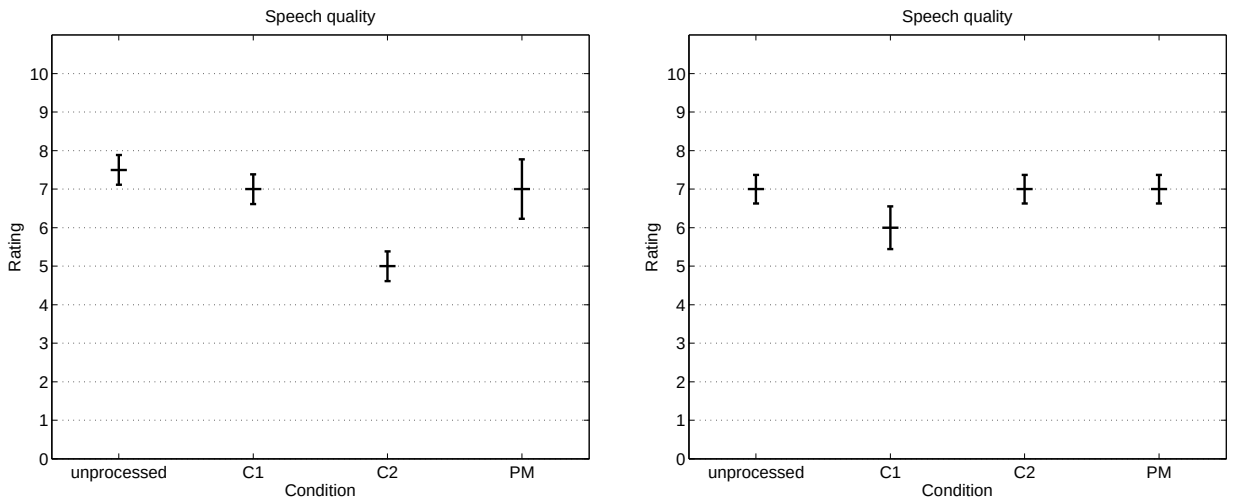
(a) Group A: Hearing impaired listeners (b) Group B: Listeners with normal hearing

Fig. 9: Results of the listening test with respect to *sound quality*.

presented in Figures 9 to 11 by means of the median of the ratings and the 95% confidence interval about the median.

Discussion The hearing impaired listeners rated the *sound quality* of the unprocessed audio, C1 and PM equally good and higher than C2. The listeners with normal hearing preferred the sound quality of the unprocessed items and rated the audio items processed by C1 distinctly lower compared to the other processed items. The results indicate that the sound quality of the processed items is perceived differently by the listeners of group A and group B.

The ratings with respect to the *speech quality* are



(a) Group A: Hearing impaired listeners

(b) Group B: Listeners with normal hearing

Fig. 10: Results of the listening test with respect to *speech quality*.

comparable to the ratings on sound quality in various ways. The hearing impaired listeners rated the C1 and PM equally good and higher than C2. The unprocessed audio items were rated slightly better compared to C1 and PM without statistical significance. The listeners with normal hearing rated the unprocessed audio, C2 and PM equally good. The results indicate that the sound quality of the processed items is perceived differently by the listeners of group A and group B.

The noise reduction achieved by PM is rated superior compared to the other methods, whereas the difference between C1 and C2 is not statistically significant. The results of the proposed method are more consistent among the two groups as for the C1 and C2.

4.3. Example signal

Figure 12 illustrates exemplarily an audio signal from a movie of 11 seconds length before and after the processing with the proposed method. Additionally, the reference and the result of the speech detection are shown. It should be noted that the speech enhancement is applied to the input signal for a small duration exceeding the detected speech region to account for false negative detections.

5. CONCLUSION

This paper presented a new approach to speech enhancement with the application to movie sound. The occurrence of speech is detected by means of a pattern recognition method. The detected speech signals are then enhanced using spectral weighting.

Results of experiments on speech detection were reported and particular characteristics of movie sound were discussed. A novel method for noise estimation was applied for the spectral weighting, whereas the time-varying sub-band SNR values are estimated by means of feature extraction and a neural network regression method. This improves the results of the objective and subjective evaluations for speech signals corrupted by nonstationary background noise in comparison to the previous methods under investigation.

The presented work is in particular directed towards the processing of movie sound to suit the needs of hearing impaired listeners. Another application is the "late night (listening) mode" for the home cinema which renders the movie sound with background noise at lower level.

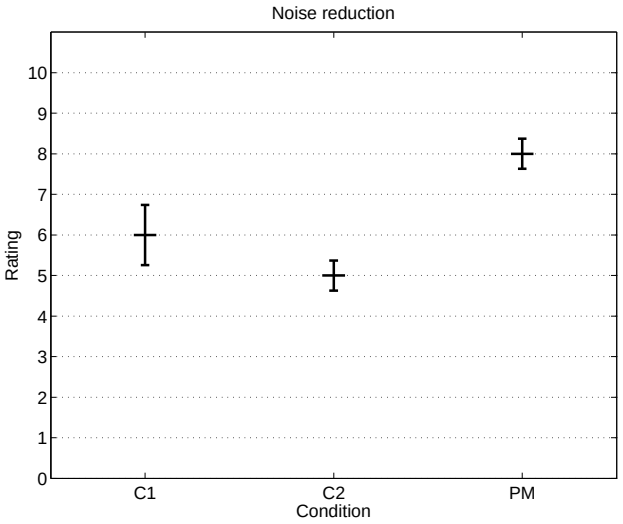


Fig. 11: Results of the listening test with respect to *noise reduction* of group B (listeners with normal hearing). No ratings for the noise reduction of the unprocessed audio signal are reported.

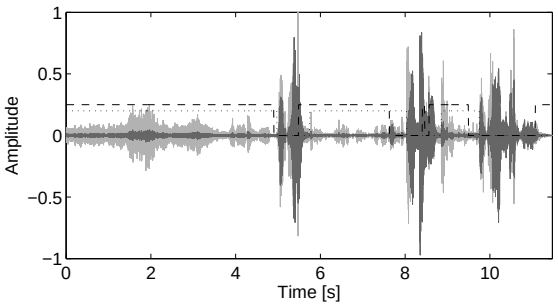


Fig. 12: Time-domain signal of an unprocessed audio signal (light) and processed signal (dark). Speech detection reference (dotted line) and speech detection result (dashed line) are shown, whereas the lower values indicate that speech is present.

6. ACKNOWLEDGEMENT

The presented work is part of the project "Enhanced Digital Cinema" (EDCine, IST-038454) which is supported by the IST sixth framework program of the European Commission.

7. REFERENCES

[1] EDCine consortium, "Enhanced Digital Cinema (EDCine)", 2008, URL `http://www.edcine.org`

[2] R. Duda, P. Hart, D. Stork, *Pattern Classification*, Wiley, 2nd ed., 2000

[3] J. Hoyt, H. Wechsler, "Detection of human speech in structured noise", *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP, Adelaide, Australia*, 1994

[4] J. Saunders, "Real-time discrimination of broadcast speech/music", *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP, Atlanta, USA*, 1996

[5] K. El-Maleh, M. Klein, G. Petrucci, P. Kabal, "Speech/music discrimination for multimedia applications", *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP, Istanbul, Turkey*, 2000

[6] J. Tchorz, B. Kollmeier, "Speech detection and SNR prediction basing on amplitude modulation pattern recognition", *Proc. of Eurospeech*, 1999

[7] S. Karnebäck, "Discrimination between speech and music based on a low frequency modulation feature", *Proc. of Eurospeech, Aalborg, Denmark*, 2001

[8] J. Pinquier, J.-L. Rouas, R. André-Obrecht, "A fusion study in speech/music classification", *Proc. of the Int. Conf. on Multimedia and Expo, ICME*, 2003

[9] N. Mesgarani, M. Slaney, S. Shamma, "Discrimination of speech from non-speech based on multiscale spectro-temporal modulations", *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 920–930, 2006

[10] R. Aarts, R. Dekkers, "A real-time speech-music discriminator", *J. Audio Eng. Soc.*, vol. 47, no. 9, pp. 720–725, 1999

[11] J. Barbedo, A. Lopes, "A robust and computationally efficient speech/music discriminator", *J. Audio Eng. Soc.*, vol. 54, no. 7, pp. 571–588, 2006

[12] R. Jarina, N. O'Connor, S. Marlow, N. Murphy, "Rhythm detection for speech-music discrimination in MPEG compressed domain", *Proc. of the 14th Int. Conf. on Digital Signal Processing, Santorini, Greece*, 2002

[13] B. Thoshkahna, V. Sudha, K. Ramakrishnan, "A speech-music discriminator using HILN-features", *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP, Toulouse, France*, 2006

[14] G. Williams, D. Ellis, "Speech/music discrimination based on posterior probability features", *Proc. of Eurospeech, Budapest, Hungary*, 1999

[15] E. Scheirer, M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator", *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP, Munich, Germany*, 1997

[16] M. Carey, E. Parris, H. Lloyd-Thomas, "A comparison of features for speech, music discrimination", *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP, Phoenix, USA*, 1999

[17] H. Harb, L. Chen, "Robust speech music discrimination using spectrum's first order statistics and neural networks", *Proc. of Int. Symposium on Signal Processing and It's Applications*, 2003

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

[19] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of speech corrupted by acoustic noise", *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP*, 1979

[20] J. Lim, A. Oppenheim, "Enhancement and bandwidth compression of noisy speech", *Proc. of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979

[21] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system", *IEEE Trans. Speech and Audio Proc.*, vol. 7, no. 2, pp. 126–137, 1999

[22] S. Kamath, P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise", *Proc. of the IEEE Int. Conf. Acoust. Speech Signal Processing*, 2002

[23] P. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2007

[24] R. Martin, "Spectral subtraction based on minimum statistics", *Proc. of EUSIPCO, Edinburgh, UK*, 1994

[25] G. Doblinger, "Computationally Efficient Speech Enhancement By Spectral Minima Tracking In Subbands", *Proc. of Eurospeech, Madrid, Spain*, 1995

[26] I. Cohen, "Noise estimation by minima controlled recursive averaging for robust speech enhancement", *IEEE Signal Proc. Letters*, vol. 9, no. 1, pp. 12–15, 2002

[27] L. Lin, W. Holmes, E. Ambikairajah, "Adaptive noise estimation algorithm for speech enhancement", *Electronic Letters*, vol. 39, no. 9, pp. 754–755, 2003

[28] H. Hirsch, C. Ehrlicher, "Noise estimation techniques for robust speech recognition", *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP, Detroit, USA*, 1995

[29] J. Tchorz, B. Kollmeier, "SNR Estimation based on amplitude modulation analysis with applications to noise suppression", *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 3, pp. 184–192, 2003

[30] M. Kleinschmidt, V. Hohmann, "Sub-band SNR estimation using auditory feature processing", *Speech Communication: Special Issue on Speech Processing for Hearing Aids*, vol. 39, pp. 47–64, 2003

[31] R. Frazier, S. Samsam, L. Braida, A. Oppenheim, "Enhancement of speech by adaptive filtering", *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP, Philadelphia, USA*, 1976

[32] N. Mesgarani, S. Shamma, "Speech enhancement based on filtering the spectrotemporal modulations", *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP, Philadelphia, USA*, 2005

[33] J. Jensen, J. Hansen, "Speech enhancement using a constrained iterative sinusoidal model", *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 7, pp. 731–740, 2001

[34] J. Makhoul, "Linear Prediction: A tutorial review", *Proc. of the IEEE*, vol. 63, 1975

[35] S. Davis, P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. om Acoustics, Speech, and Signal Proc.*, vol. 28, no. 4, pp. 357–366, 1980

[36] H. Hermansky, "Perceptual linear predictive (PLP) analysis for speech", *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, 1990

[37] H. Hermansky, N. Morgan, "RASTA Processing of Speech", *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994

[38] A. Favrot, C. Faller, "Perceptually motivated gain filter smoothing for noise suppression", *Proc. of the AES 123rd Conv., New York, USA*, 2007

[39] E. Zwicker, H. Fastl, *Psychoacoustics*, Springer, 2nd ed., 1999

[40] I. T. Nabney, *NetLab - Algorithms for pattern recognition*, Springer, 2002

[41] P. Masri, "Computer modelling of sound for transformation and synthesis of musical signals", Ph.D. thesis, University of Bristol, 1996

[42] ISO/MPEG, "ISO/IEC 15938-4 MPEG-7", Int. Standard, 2002

[43] ISO/MPEG, "ISO/IEC 11172-3 MPEG-1", Int. Standard, 1993

[44] K. Sorensen, S. Andersen, "Speech enhancement with natural sounding residual noise based on connected time-frequency speech presence regions", *EURASIP J. on Appl. Signal Proc.*, vol. 18, pp. 2954–2964, 2005

View publication stats