A SUPERVISED LEARNING APPROACH TO AMBIENCE EXTRACTION FROM MONO RECORDINGS FOR BLIND UPMIXING

Christian Uhle

Fraunhofer Institute for Integrated Circuits,
Erlangen, Germany
Christian.Uhle@iis.fraunhofer.de

Christian Paul

Fraunhofer Institute for Integrated Circuits
Erlangen, Germany
Christian.Paul@iis.fraunhofer.de

ABSTRACT

A supervised learning approach to ambience extraction from one-channel audio signals is presented. The extracted ambient signals are applied for the blind upmixing of musical audio recordings to surround sound formats. The input signal is processed by means of short-term spectral attenuation. The spectral weights are computed using a low-level feature extraction process and a neural network regression method. The multi-channel audio signal is generated by feeding the computed ambient signal into the rear channels of a surround sound system.

1. INTRODUCTION

Multi-channel surround sound systems allow for an impressive reproduction of audio recordings in terms of conveying a natural and enveloping experience. The increasing availability of surround sound systems (e.g. home theatre and multimedia computer setups) evokes the consumers' desire to exploit their advantages for the reproduction of legacy content. The mismatch between the surround sound setup and the legacy content format (either mono or stereo) creates the need for content format conversion.

The term *upmixing* refers to the conversion of an audio signal to another format with more channels. Two concepts of upmixing are widely known: Upmixing with additional information guiding the upmix process (see e.g. [1] for a recent overview) and unguided ("blind") upmixing without the use of any side information, which is what this work relates to.

When processing stereo recordings, the difference between the left and right channel signals may be evaluated by the upmix process. Matrix-based techniques are one particular approach for upmixing in which linear combinations of the left and right input signal create the multi-channel output signals. Alternatively, some matrix-based upmix systems dynamically update the gain factors in the matrix based on a detection of the dominant element of the audio scene [2, 3]. Statistical techniques have been applied to the computation of the matrix elements in [4].

More advanced methods [5, 6] operate in the frequency domain, such as ambience-based techniques [3, 4, 5]. Their core component extracts an ambient signal which is fed into the rear channels of a multi-channel surround sound signal. The ambient sounds are those that form an impression of a (virtual) listening environment, including room reverberation, audience sounds (e.g. applause), environmental sounds (e.g. rain), artistically intended effect sounds (e.g. vinyl crackling) and background noise. This technique evokes an impression of envelopment ("immersed in sound") by the listener.

These approaches are applicable to audio recordings with more than one channel. A method for ambience extraction for mono recordings based on Non-negative Matrix Factorization has been described in [7]. The disadvantages of this previous method are high computational complexity and high latency.

This publication relates to the extraction of an ambient signal from audio recordings with one channel for the purpose of upmixing. The proposed method incorporates the extraction of low-level features and a supervised learning method to estimate the spectral weights, which are applied to the input signal in the frequency domain to compute the ambient signal. This approach is of low computational complexity and low latency compared to [7].

The processing is influenced by two techniques of audio signal processing, namely Adaptive Spectral Panoramization (ASP) [8] and noise suppression for speech enhancement.

1. ASP aims at the automated positioning of a sound source within a stereo panorama. The left and right channel signals of a stereo recording are time-varying filtered, whereas the filter characteristic is controlled by a feature extraction process applied to the input signal.
2. A prominent family of noise suppression methods for speech enhancement is based on a time-varying filtering process of the input signal and is known as *short-term spectral attenuation* (STSA) [9] or *spectral weighting* [10], whereas the filter characteristic is controlled by an estimate of the noise energy corrupting the speech signal[1].

This paper is organized as follows: Section 2 describes the underlying idea of the proposed method. Section 3 gives an overview of the processing, which is divided into two separate processes described in Section 4 (ambience estimation) and Section 5 (ambience extraction). The evaluation procedure and results are described in Section 6 and conclusions are given in Section 7.

2. PRELIMINARY CONSIDERATIONS

In general, a musical recording contains sound components emitted from one or more sound sources (e.g. instruments and singers) and reverberations of the room surrounding the sound sources. In the following, the sources are denoted as "direct sounds" (synonymous with the term "primary sound sources" used in e.g. [5]). The room reverberations add sound components, which evoke an impression of ambience when reproduced properly by the recording. There may be additional ambient sound sources as well, e.g. the audience in a live performance (applause), environmental sounds (like rain and wind) or other background noises.

[1] Another speech enhancement method is structurally very similar compared to a previous method for ambience extraction from stereo recordings[5].

A valid signal model for ambience extraction is to assume an additive mixture of the direct sounds $d[n]$ and the ambient sounds $a[n]$, such that the recorded sound $x[n]$ can be written as

$$x[n] = d[n] + a[n] \tag{1}$$

A similar signal model has been commonly applied in noise suppression methods for speech enhancement, e.g. in [11, 12]. For the following considerations related to speech enhancement, the observed signal x (e.g. the microphone signal) is assumed to be an additive mixture of a speech signal $s[n]$ (which is the desired signal) and a background noise $b[n]$ (which corrupts the desired signal).

Noise suppression methods based on STSA may filter the input signal by computing a Short-term Fourier Transform (STFT) and weighting the spectral coefficients according to Equation 2.

$$S(m,k) = H(m,k)X(m,k) \tag{2}$$

Here, $X(m,k)$ and $S(m,k)$ denote the STFT coefficients of $x[n]$ and the estimate of the desired signal $s[k]$, respectively. The spectral weights $H(m,k)$ are positive and real-valued, k is the index of the time frame and m is the index of the frequency bin.

The spectral weights $H(m,k)$ can be computed using an estimate $R_s(m,k)$ of the time-frequency representation of the signal-to-noise ratio or an estimate $B(m,k)$ of the spectral coefficients of the background noise $b[n]$. A particular method is *spectral magnitude subtraction*, in which the spectral weights are computed according to Equation 3.

$$H(m,k) = \frac{R_s(m,k)}{R_s(m,k)+1} \tag{3}$$

Other gain values are derived by applying the Wiener filter rule (see e.g. [10]) or the spectral subtraction rule [11].

The STSA approach to noise suppression for speech enhancement can be summarized as two separate processing steps:

1. Noise estimation, i.e. the estimation of the power spectral density or the instantaneous spectra of the background noise or the estimation of the SNR in frequency bands.
2. Noise suppression, i.e. the attenuation of the noise in the observed signal.

The processing described above is equivalently applicable to the problem of ambience extraction, since the underlying signal model and the task are similar. The definition of the desired signal changes from "desired speech signal" to "ambient signal". The definition of the background signal changes from "corrupting background noise" to "direct signal components".

Consequently, the spectral weights may be computed according to Equation 3, whereas $R_s(m,k)$ is replaced by an estimate $R(m,k)$ of the ratio of ambient sound signals and direct sound signals.

$$R(m,k) = \frac{A(m,k)}{D(m,k)} \tag{4}$$

The preceding considerations lead to the problem of estimating the *ambient-to-direct ratio* (ADR) $R(m,k)$. In previous publications on ambience extraction from stereo recordings, the spectral weights for the ambience extraction are derived by evaluating the correlation between the left and right channel signals [4, 5, 6]. The correlation between the stereo channels in each frequency band is low in regions dominated by ambience and is therefore a valid cue for ambience extraction from stereo recordings. The sole use of information based on differences between the signals of a stereo or multi-channel recording for upmixing is clearly a restriction of such methods. It prevents them from processing mono signals or recordings with negligible inter-channel signal differences.

This publication investigates the application of a supervised learning method to the task of ambience extraction from mono signals. The underlying idea originates from the experience that ambient sounds are recognized even in mono recordings. There are signal characteristics guiding the discrimination between ambience and direct sound components.

The work begins with the question about the physical nature of ambience. Room reverberations result in an additive mixture of differently delayed and attenuated copies of the direct sound due to reflections of the sound by the walls, the ceiling and the floor. Consequently, ambient signals recorded by spaced or differently oriented microphones are less correlated compared to direct signals. Additionally, the following characteristics are observed:

- The ambient signal components in a stereo recording have comparable levels when averaged over time [13].
- The direct sounds have shorter attack times and decay times compared to ambient sounds.
- In general, the absorption of the sound energy by the reflecting surfaces of a room is greater at high frequencies. The reverberation time decreases with increasing frequency. Consequently, direct sounds are brighter sounding than ambient sounds, especially in rooms as used for musical recordings.

3. PROCESSING OVERVIEW

An overview of the presented method is shown in Figure 1. The processing is performed in the frequency domain. The spectral coefficients $X(m,k)$ are computed from the input signal $x[n]$ by means of the STFT, with time frame index m and frequency bin index k. The reported results are obtained with an STFT of overlapping data frames of 11.6 ms length each, a transform length of 23.2 ms and the Hann window function.

A set of low-level features Z is computed from the spectral coefficients $X(m,k)$ in frequency bands corresponding to the critical band scale [14], as described in Section 4.1. The features are fed into a neural network which is trained to estimate the positive and real-valued spectral weights $H_b(m,r)$ for each frequency band, with frequency band index r.

The spectral weights are interpolated to the frequency resolution of the input spectra, yielding $H(m,k)$. The STFT coefficients of the ambient signal $A(m,k)$ are computed by multiplying the input spectra $X(m,k)$ with the spectral weights $H(m,k)$. The ambient time signal $a[n]$ is derived by the inverse processing of the STFT computation.

4. AMBIENCE ESTIMATION

In this section, the estimation of the spectral weights $H_b(m,r)$ is described. A set of low-level features is computed for each time frame m from the spectral coefficients in each frequency band. In the following, the indices of the STFT coefficient corresponding to the lower and upper boundary of the frequency bands are denoted by l_r and u_r, respectively. The spectral weights are estimated from the low-level features by means of a neural network regression method.

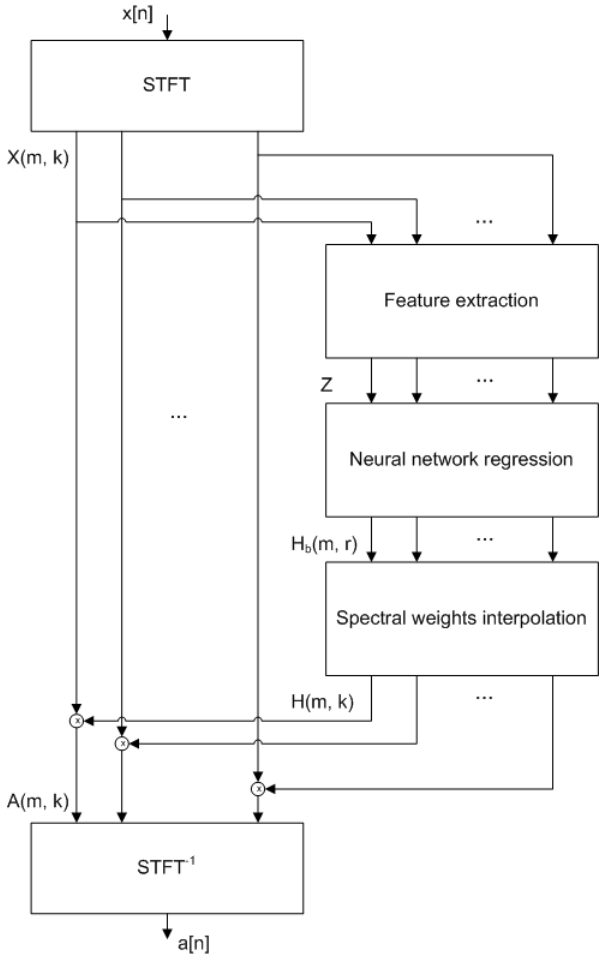

Figure 1: *Block diagram of the ambience extraction processing.*

4.1. Feature extraction

The choice of the extracted features is determined by the characteristics of ambience as described in Section 2. The initial feature set is comprised of the spectral energy, the spectral energy difference, the spectral flux, the spectral flatness, and the spectral centroid.

4.1.1. Spectral centroid

The spectral centroid $SC(m,r)$ corresponding to the r-th frequency band is computed from the STFT coefficients $X(m,k)$ with bin frequency $f(k)$ according to

$$SC(m,r) = \frac{\sum_{q=l_r}^{u_r} |X(m,q)|\, f(q)}{\sum_{q=l_r}^{u_r} |X(m,q)|} \tag{5}$$

The spectral centroid is a low-level feature that correlates (when computed over the whole frequency range of a spectrum) to the perceived brightness of a sound [15]. It is normalized to the frequency range of the sub-band such that $0 \leq SC(m,r) \leq 1$.

4.1.2. 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. [15, 16]. 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.

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

4.1.3. Spectral flux

The spectral flux SF is defined as the dissimilarity between spectra of successive frames [17] 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 7.

$$SF(m,r) = \sqrt{\sum_{q=l_r}^{u_r} (|X(m,q)| - |X(m-1,q)|)^2} \tag{7}$$

4.1.4. Spectral energy difference

The spectral energy difference SD is computed as the mean of the difference of the spectral energy of successive frames according to Equation 8.

$$SD(m,r) = \sum_{q=l_r}^{u_r} (|X(m,q)|^2 - |X(m-1,q)|^2) \tag{8}$$

Contrary to the spectral flux, the spectral energy difference distinguishes between the directions of the change in the temporal progression of the energy in frequency bands.

4.1.5. Spectral energy

The spectral energy SE is computed in each time frame and frequency band and normalized by the total energy of the time frame. Subsequently, the feature values are low-pass filtered over time by means of a second-order IIR filter.

4.2. Feature post-processing

The extracted features are accumulated into the feature set and further processed prior to the training and the application of the regression method.

4.2.1. Centering and variance normalization

The features $z \in Z$ are processed to have zero mean and unit variance according to

$$\widetilde{z} = \frac{z - E\{z\}}{\sqrt{E\left\{(z - E\{z\})^2\right\}}} \tag{9}$$

to eliminate side effects on the regression process due to different scaling of the features values, where $\widetilde{z}$ denotes the transformed feature and the expectation values are computed from the training data.

4.2.2. Feature grouping

The sub-band features computed from a small number of successive signal frames are subsumed to larger entities (in the following denoted as *group*) with a hop size of one frame. The rationale behind the grouping is to evaluate the progression of the features over time.

The groups are represented by the means and the variances of the feature values computed from the respective frames. The reference values for each group are computed as the mean of the references of the corresponding frames.

4.3. Neural network regression

The neural network is utilized to estimate the spectral weights from the low-level feature set Z. A training algorithm from the *NetLab* toolbox [18] is used in this work. For a given spectral weighting rule $g(\cdot)$ (as shown in Equation 3), two definitions of the output of the neural network are appropriate. The neural network can be trained using the reference values for the ADR $R_b(m,r)$ or with the spectral weights

$$H_b(m,r) = g\left(R_b(m,r)\right) \tag{10}$$

Figure 2 illustrates the distributions of the ADR and the spectral weights, for six selected sub-bands. The histogram plots reveal that the distribution of the ADR is not flat and concentrated at small values, whereas the spectral weights $H_b(m,r)$ show a rather flat distribution. In the following, the reference values for the spectral weights are used as the references for the training of the neural network.

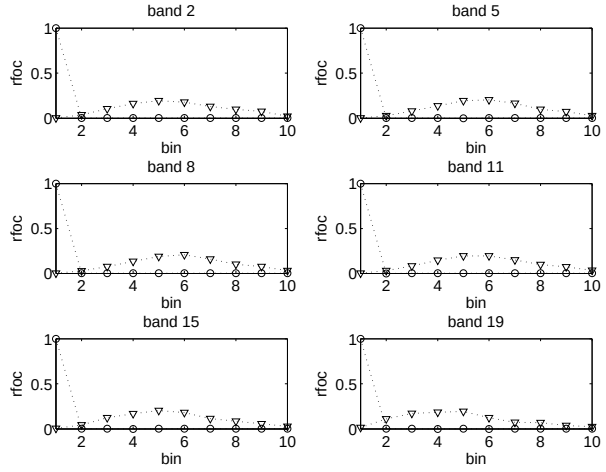

Figure 2: *Histograms of $R_b(m,r)$ (circle) and $H_b(m,r)$ (triangle) for six selected frequency band illustrating the relative frequency of occurrence (rfoc).*

4.3.1. Structure of the neural network

The neural network has N input neurons, one hidden layer with K neurons and M output neurons. With B being the number of frequency bands and W the number of features, the number of input neurons is $N = 2BW$ (if the mean values and the variances of the features of each group of successive frames are fed into the neural network). The number of output neurons equals the number of frequency bands, $M = B$.

The results presented here are obtained with $K = 40$ hidden neurons unless otherwise specified. The activation function of the hidden neurons is the hyperbolic tangent. The activation function of the output neurons is the identity, such that the required computations in the output layer are reduced to linear combinations of the features and the weights of the neural network. Each neural network is trained using 100 iteration cycles.

4.3.2. Reference data for training and test

A crucial aspect for the application of supervised learning methods is the proper choice of the reference values used for the training. The training of a neural network for the task of ambience estimation requires audio signals whose direct signal and ambient signal are separately available. Appropriate audio signals are ideally generated using anechoic recordings as direct signals and artificially reverberated copies of the recordings as ambient signals (whereas the direct path of the reverberation processor is muted).

Since a sufficient amount of anechoic recordings comprising different musical genres is not available, commercial recordings with a negligible amount of ambience (i.e. recordings which are in general considered as being very "dry") are considered as direct signals. The ambient signals are generated by convolving the audio signals with recordings of room impulse responses.

5. AMBIENCE EXTRACTION

The regression results $H_b(m,r)$ are interpolated to the frequency resolution of the input spectra. Prior to the modification of the input spectra $X(m,k)$, the estimated spectral weights are post-processed by a non-linear mapping function and a low-pass filter.

The non-linear mapping of spectral weights has been applied previously [5] and aims at increasing the ambient signal components in the output signal while reducing the direct signal components. The mapping function $g(H)$ applied here is given in Equation 11.

$$g\left(H\right) = \sin^2\left(H \cdot \pi/2\right) \tag{11}$$

Subsequently, the spectral weights are low-pass filtered along time (using a first-order IIR filter) to account for erroneously occurring fast fluctuations in the temporal progression of the estimation results.

The complex STFT coefficients of the ambient signal $A(m,k)$ are computed from the input spectra $X(m,k)$ and the spectral weights $H(m,k)$. The ambient time signal is resynthesized using the inverse processing of the STFT.

6. EVALUATION

6.1. Data sets

The estimation of the ambience weights is evaluated using a data set of 80 excerpts of musical recordings with a length of 10 seconds each. These items were recorded with a negligible amount of room reverberations. Different musical genres are considered in the choice of the audio items.

An ambient signal of each audio item is computed by convolution with one of a set of 25 room impulse responses. The impulse

responses were edited by attenuating the first impulse corresponding to the direct path. The direct signals and the ambient signals are additively mixed. Different mixing levels were chosen for each of the impulse responses to ensure that the mixture signals contain a reasonable amount of reverberation. The audio signals and the impulse responses were recorded in a two-channel stereo format. The mixture signals were downmixed to mono prior to the ambience estimation processing.

6.2. Regression results

The performance of the ambience estimation is evaluated by means of

1. the regression error computed as L1-norm of the differences between reference and estimated value
2. the (normalized) correlation coefficient between references and estimates

A ten-fold cross validation is applied by dividing the data set into sets of training data (90%) and test data (10%). If not noted otherwise, the experimental results are obtained with feature grouping of six frames using the mean values and the variances of the feature.

Figure 3 details the mean regression error of the validation runs separately for each frequency band. The results for two selected parameter settings are shown. The first simulation is computed by using the complete feature set without grouping, the second simulation uses a reduced feature set $\{SFM, SF, SD\}$ with grouping.

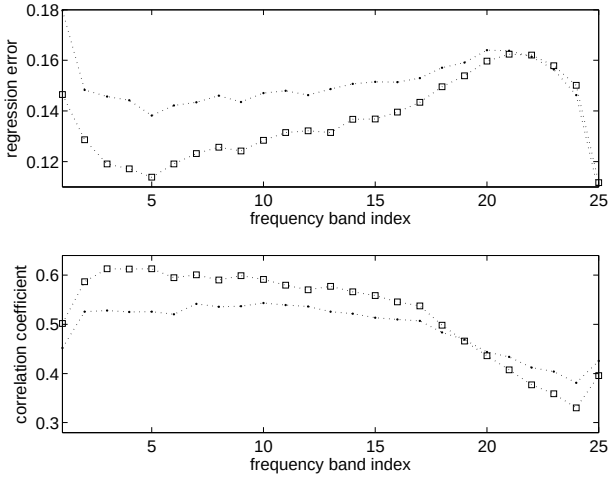

Figure 3: *Regression results per frequency band: regression error (L1-norm) (upper figure) and correlation coefficient between references and estimates (lower figure) for two different parameter settings. The initial feature set without grouping (point) is compared to a reduced feature set with grouping (square).*

The results indicate a moderate correlation between the estimated spectral weights and the reference values for the first eighteen frequency bands, with a mean correlation of 0.49 and 0.52 for the first and second condition, respectively. The regression results decrease at the upper frequency bands, probably due to the decrease in the ambient signal energy in that bands.

Examples of the estimated spectral weights and the reference values are shown in Figure 4 for three frequency bands of an audio signal with a length of 2.3 seconds from the training data set. The upper plot shows the input time signal, the other plots show the estimation results and the references for the frequency bands centered at 150 Hz, 845 Hz, and 2510 Hz.

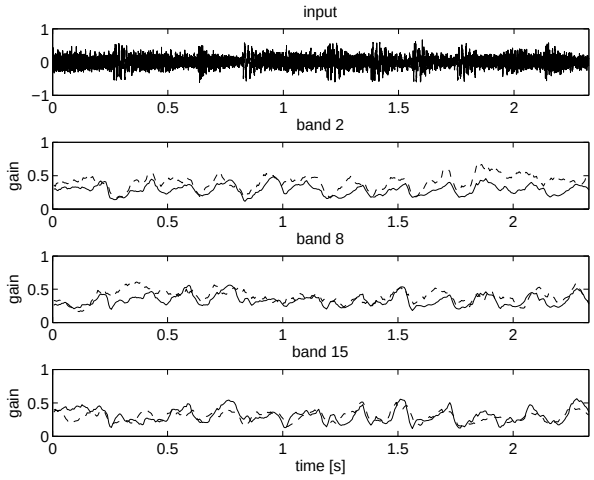

Figure 4: *Input time signal and examples of spectral weights of frequency band number 2, 8, and 15 for a signal with a length of 2.3 seconds. References are shown by dashed lines and the estimated weights by solid lines.*

6.3. Experiments on the influence of selected parameters

This section investigates the impact of selected parameters on the regression result. The following parameters are modified in the experiments:

- the feature set
- the grouping parameters
- the number of hidden neurons

6.3.1. Variation of the feature set

The impact of the feature set on the regression result is evaluated in experiments with backwards feature selection. The results obtained with the initial feature set $\{SC, SFM, SF, SD, SE\}$ are compared to conditions where one feature has been removed from the feature set. The regression results are shown in Figure 5.

Discarding either SD, SF or SFM leads to higher regression errors and lower correlation between the estimations and the references. The results are only slightly affected by the removal of the features SE or SC from the initial feature set. Experiments without feature grouping and with different STFT parameters (larger frame size) yield similar ranking of the features. This leads to the next experiment where both features (SE and SC) are discarded.

Figure 6 illustrates the recognition results for the initial feature set, and for feature sets where either SE or SC or both, SE and SC, were discarded. It is shown that the regression performance is improved when using the feature set $\{SFM, SF, SD\}$.

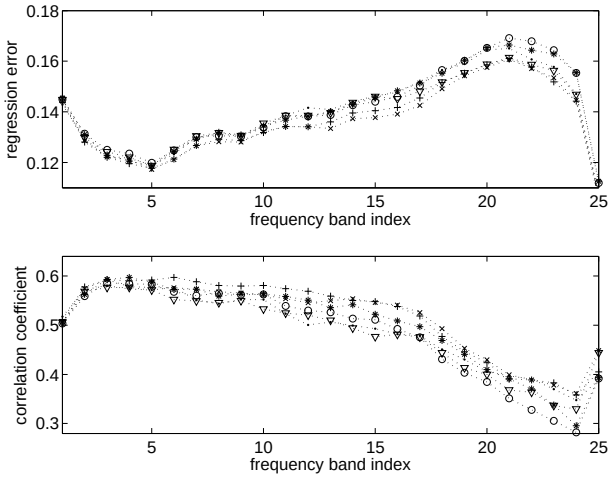

Figure 5: *Regression results for different feature sets: initial feature set (plus), without SD (triangle), without SE (x-mark), without SF (circle), without SC (star), without SFM (point)*

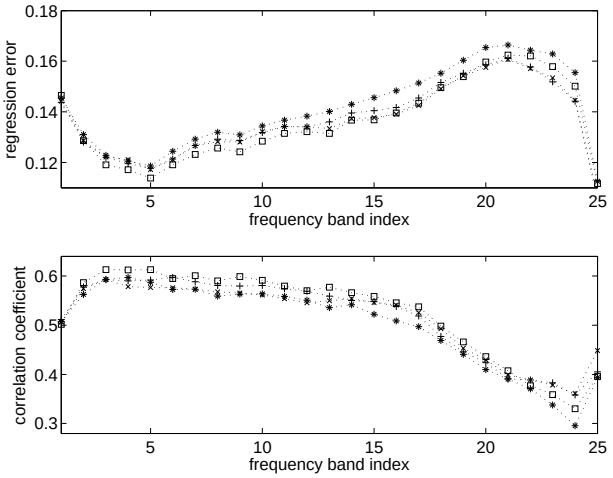

Figure 6: *Regression results for different feature sets obtained by feature selection: initial feature set (plus), without SE (x-mark), without SC (star), without SE and SC (square)*

6.3.2. Grouping of frames

The influence of the grouping parameters on the regression results is investigated in the following experiment. Figure 7 illustrates the result of regressions

- without grouping
- with grouping using the means of the feature values
- with grouping using the means and the variances
- with grouping with varied grouping size.

The experiments indicate an improvement of the regression results when grouping is applied for the lowest 17 frequency bands, corresponding to a frequency range up to 3150 Hz. For this frequency range, the best result in terms of regression error and correlation is obtained with grouping using the means and variances of the feature values and grouping size of six frames, although the differences are small compared to a smaller grouping size.

The impact of the grouping parameters on the regression results is different for frequencies larger than 3700 Hz. This observation leads to the assumption that further improvements can be obtained by using different grouping parameters in different frequency bands, which is not investigated further.

6.3.3. Number of hidden neurons

The number of hidden neurons influences the computational load of the ambience estimation process, particularly owing to the nonlinear activation function. Therefore, it is desired to determine the number of hidden neurons such that the use of additional neurons does not improve the recognition rate significantly.

Figure 8 illustrates the regression result obtained from neural networks in which the number of input neurons is varied in the range between 1 and 100. The regression error and the correlation coefficients are averaged over the first 17 frequency bands. It is shown that only minor improvements are obtained by using a neural net with more than 40 neurons.

6.4. Comparison to ambience extraction from stereo recording

For the purpose of comparison, spectral weights are computed from stereo recordings by means of the normalized cross-correlation coefficient between the left and the right channel as used in [5] from the data base described in Section 6.1 (prior to the downmixing to mono). The normalized cross-correlation coefficient is computed recursively and processed using a non-linear mapping function as described in [5]. The spectral weights are subsequently averaged over the frequency bins corresponding to the critical bands. It should be noted that no additional criterion for the detection of signal components panned completely to one side is used for the computation of the spectral weights for stereo signals.

Figure 9 illustrates the regression error and the correlation coefficients between the reference values and spectral weights obtained from the inter-channel cross-correlation coefficient. Surprisingly, the evaluation metrics used in this work indicate comparable regression results for the processing of mono recordings using the method presented here and stereo recordings using a method based on the inter-channel correlation.

6.5. Informal listening

In order to gain an impression of the ambient signal in the context of upmixing, various real-world items were processed. Multichannel signals were assembled by feeding the unprocessed two-channel signal into the left and right front channels and by feeding the ambient signal into the rear channels with a delay of 11 ms. The rear channel signals were not subjected to additional post-processing (as e.g. described in [19]).

The resulting multi-channel signals evoke the impression of envelopment when played back on a 5.0 surround sound system with moderate ambience volume level. However, the localization of the direct sound sources remains in the front as desired.

Listening to the extracted ambient signals over headphones reveals that their sound characteristics are very similar to that of the room reverberations. Prominent direct signal components are sig-

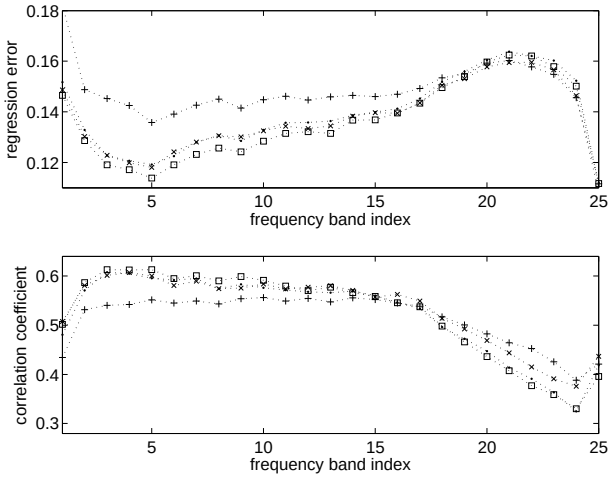

Figure 7: *Regression results for different grouping configurations: without grouping (plus), grouping using mean values with grouping size 4 (x-mark), grouping using mean values and variances with grouping size 4 (point), grouping using mean values and variances with grouping size 6 (square)*

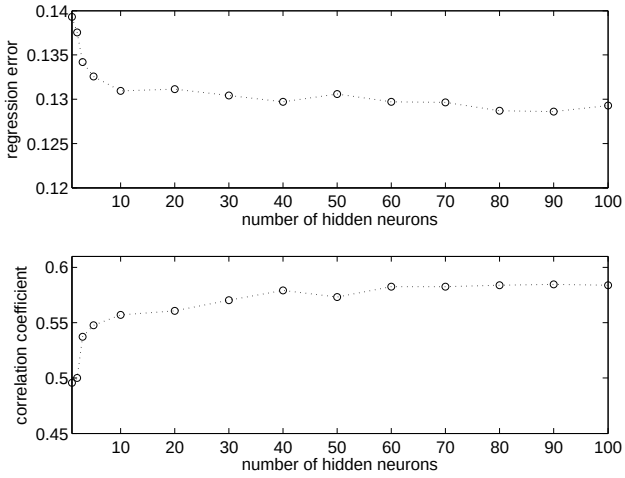

Figure 8: *Influence of the number of hidden neurons on the regression error and the correlation between estimated values and references.*

nificantly although not completely attenuated. Minor pumping artifacts may result from the processing but are masked by the front channel signals when played back on a surround sound system.

6.6. Listening test

The novel method was compared to unprocessed audio signals and to a previous approach to ambience-based upmixing of one-channel audio signals by means of a listening test. The previous method computes an ambient signal using a Non-negative Matrix Factorization of a spectrograms of the input signal of overlapping segments of a 4 seconds length each [7].

Six excerpts from commercial recordings from a variety of musical genres with a length between 8 and 16 seconds each were presented. The signal levels of the audio items were adjusted such that the audio items under comparison were perceived to be equally loud. The audio signals were presented using a surround sound setup with 5 loudspeakers Genelec 8250A arranged according to ITU-R BS.775. A group of 11 subjects participated in the test, 8 of them participated in listening tests with surround sound before. The listeners were asked to rate the test conditions according to their personal preference.

The results of the listening test are illustrated in the box plot in Figure 10. The two conditions with surround sound were rated higher compared to the original. The listening test indicates no significant difference between the two upmix methods.

6.7. Computational complexity

The computational load of the presented method is mainly caused by the STFT, the feature extraction and post-processing, and the regression function. The regression takes C (see Equation 12) real-valued multiplications and additions plus K computations of the hyperbolic tangent for each frame.

$$C = K \cdot (2BW + B) \tag{12}$$

A typical configuration of the regression function as used in the presented experiments may result in 7000 operations (without the computations of the hyperbolic tangent) per frame. The computational complexity of the feature extraction and post-processing is moderate. However, the computational load of the proposed ambience extraction process is much lower compared to the method for ambience extraction of mono signals as described in [7], which takes about 120000 operations per frame for each iteration of the numerical optimization method.

7. CONCLUSIONS

A novel approach to ambience extraction from audio recordings which is applicable to one-channel audio signals has been presented. The core component of the proposed method is the estimation of spectral weights which relate to the energy ratio of the ambience signal components and the direct signal components for each time frame and frequency band.

The results of the ambience estimation are evaluated by means of the averaged magnitude differences and the correlation coefficient between the references and the regression results of the spectral weights. The evaluation of the regression results indicates that the correlation between the references and the estimates is comparable to the correlation between the same references and spectral weights computed using the inter-channel cross-correlation coefficient between the left and right channel of a stereo recording.

Listening to the extracted ambience signals over headphones and to multi-channel surround sound produced by the presented method reveal that the sound characteristics of the recorded ambience are successfully captured. The result of a listening test confirms that surround sound produced by the presented method is preferred compared to the unprocessed audio signal. There was no significant difference in the listeners' preference compared to a previous method with distinctly higher computational complexity and higher latency. The results are promising and indicate that ambience-based upmixing is applicable to mono recordings.

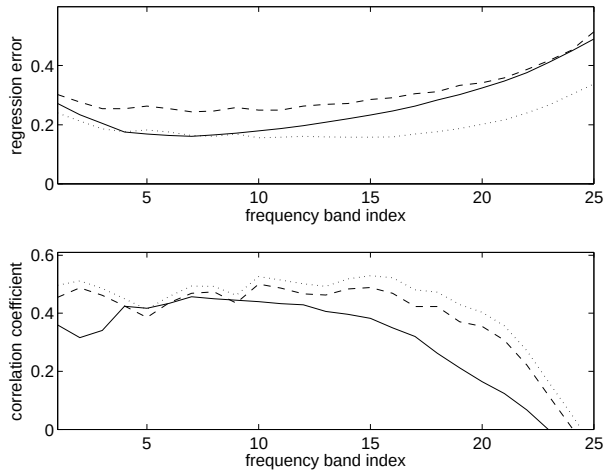

Figure 9: *Regression error (L1-norm) (upper figure) and correlation coefficient (lower figure) between references and cross-correlation-based weights in stereo recordings. Weights are computed without mapping (dotted line), with mapping (dashed line) and with mapping and additional low-pass filtering (solid line)*

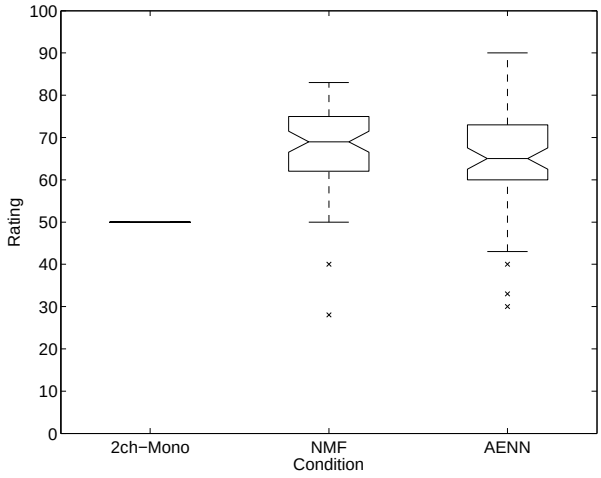

Figure 10: *Results of the listening test. The notches indicate the 95 % confidence interval about the median. The maximum whisker length is 1.5 times the interquartile range of the sample. Outliers are shown by x-marks.*

8. REFERENCES

[1] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Röden, W. Oomen, K. Linzmeier, and K. S. Chong, "MPEG Surround - the ISO/MPEG standard for efficient and compatible multichannel audio coding," in *Proc. of the 122nd AES Convention, Vienna, Austria*, 2007.

[2] R. Dressler, "Dolby Surround Pro Logic 2 Decoder: Principles of operation," *Dolby Laboratories Information*, 2000.

[3] D. Griesinger, "Multichannel matrix decoders for two-eared listeners," in *Proc. of the AES 101st Convention, Los Angeles, USA*, 1996.

[4] R. Irwan and R. Aarts, "Two-to-five channel sound processing," *J. Audio Eng. Soc.*, vol. 50, pp. 914–926, 2002.

[5] C. Avendano and J. M. Jot, "Ambience extraction and synthesis from stereo signals for multi-channel audio upmix," in *Proc. of the ICASSP, Orlando, Florida*, 2002.

[6] C. Faller, "Multiple-loudspeaker playback of stereo signals," *J. Audio Eng. Soc.*, vol. 54, pp. 1051–1064, 2006.

[7] C. Uhle, A. Walther, O. Hellmuth, and J. Herre, "Ambience separation from mono recordings using Non-negative Matrix Factorization," in *Proc. of the AES 30th Conference, Saariselkä, Finland*, 2007.

[8] V. Verfaille, U. Zölzer, and D. Arfib, "Adaptive digital audio effects (A-DAFx): A new class of sound transformations," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, pp. 1817–1831, 2006.

[9] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim-Malah noise suppressor," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 345–349, 1994.

[10] G. Schmidt, "Single-channel noise suppression based on spectral weighting," *Eurasip Newsletter*, 2004.

[11] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, 1979.

[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *Trans. on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 1109–1121, 1984.

[13] C. Avendano and J.-M. Jot, "A frequency-domain approach to multi-channel upmix," *J. Audio Eng. Soc.*, vol. 52, pp. 740–749, 2004.

[14] E. Zwicker and H. Fastl, *Psychoacoustics*, Springer Verlag, 2nd edition, 1999.

[15] ISO/MPEG, "ISO/IEC 15938-4 MPEG-7," International Standard, 2002.

[16] ISO/MPEG, "ISO/IEC 11172-3 MPEG-1," International Standard, 1993.

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

[18] Ian T. Nabney, *NetLab - Algorithms for pattern recognition*, Springer, 2002.

[19] C. Uhle, A. Walther, and M. Ivertowski, "Blind One-to-N upmixing," in *Proc. of the Audio Mostly Conf., Ilmenau, Germany*, 2007.