available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/brainres

BRAIN RESEARCH

Research Report

Sparse gammatone signal model optimized for English speech does not match the human auditory filters☆

Stefan Strahl[a,*], *Alfred Mertins*[b]

[a]*International Graduate School for Neurosensory Science and Systems, Carl von Ossietzky University, D-26111 Oldenburg, Germany*
[b]*Institute for Signal Processing, University of Lübeck, Ratzeburger Allee 160, D-23538 Lübeck, Germany*

ARTICLE INFO

Article history:
Accepted 28 November 2007
Available online 16 January 2008

Keywords:
Sparse coding
Matching pursuit
Gammatone
Efficient coding hypothesis

ABSTRACT

Evidence that neurosensory systems use sparse signal representations as well as improved performance of signal processing algorithms using sparse signal models raised interest in sparse signal coding in the last years. For natural audio signals like speech and environmental sounds, gammatone atoms have been derived as expansion functions that generate a nearly optimal sparse signal model (Smith, E., Lewicki, M., 2006. Efficient auditory coding. Nature 439, 978–982). Furthermore, gammatone functions are established models for the human auditory filters. Thus far, a practical application of a sparse gammatone signal model has been prevented by the fact that deriving the sparsest representation is, in general, computationally intractable. In this paper, we applied an accelerated version of the matching pursuit algorithm for gammatone dictionaries allowing real-time and large data set applications. We show that a sparse signal model in general has advantages in audio coding and that a sparse gammatone signal model encodes speech more efficiently in terms of sparseness than a sparse modified discrete cosine transform (MDCT) signal model. We also show that the optimal gammatone parameters derived for English speech do not match the human auditory filters, suggesting for signal processing applications to derive the parameters individually for each applied signal class instead of using psychometrically derived parameters. For brain research, it means that care should be taken with directly transferring findings of optimality for technical to biological systems.

1. Introduction

There is evidence (Olshausen et al., 1996; Attwell and Laughlin, 2001; Olshausen and Field, 2004) that neurosensory systems encode stimuli by activating only a small number of neurons out of a large population at the same time. This concept of a 'sparse' signal representation has gained interest in the signal processing community in the last years (Mallat and Zhang, 1993; Davis, 1994; Chen, 1995; Gorodnitsky and Rao, 1997; Gribonval, 1999; Hoyer, 2002; Donoho and Tsaig, 2005; Aharon et al., 2006) as it shows improved performance in signal compression, analysis and denoising tasks (Neff and Zakhor, 1997; Chen et al., 1998; Gribonval, 2001; Donoho et al., 2006). A sparse signal model indicates the fundamental features of the signal as it necessarily involves expansion functions that are highly correlated with the signal. For natural audio signals like speech and environmental sounds, gammatone atoms have been derived as expansion functions that generate a nearly optimal

☆ This work was partly funded by the German Science Foundation (DFG) through the International Graduate School for Neurosensory Science and Systems and the SFB/TRR 31: 'The Active Auditory System'.

* *Corresponding author*. Present address: UCL Ear Institute, 332 Gray's Inn Rd, London WC1X 8EE, UK. Fax: +44 20 7679 8990.
E-mail addresses: stefan.strahl@uni-oldenburg.de (S. Strahl), mertins@isip.uni-luebeck.de (A. Mertins).

0006-8993/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.brainres.2007.11.059

sparse signal model (Lewicki, 2002; Smith and Lewicki, 2006). Gammatone functions are also known as filters modeling the human cochlea (Patterson and Moore, 1986; Patterson et al., 1988) and gammatone filterbanks are applied successfully in simulating the human auditory processing (Dau et al., 1996a,b; Patterson, 2000; Chi et al., 2005). Deriving the sparsest representation of a signal has been proven to be NP-hard (Davis, 1994, ch. 2) and is therefore, in general, computationally intractable. In this paper, we apply an accelerated sparse signal model for gammatone functions which is a specialization of Matching Pursuit (Mallat and Zhang, 1993). This time-frequency algorithm computes a sparse signal model from a given dictionary of atoms. At every iteration, the dictionary atom that best matches the signal is chosen and removed from the signal. This is repeated until the signal residuum is small enough or a maximal number of iterations have been reached. It has been shown (Goodwin, 1998; Gribonval, 1999; Krstulovic and Gribonval, 2006) that the complexity of such an algorithm can be reduced for dictionaries that exhibit a special structure and we apply these results to gammatone dictionaries resulting in a computational complexity of $\mathcal{O}(N \log N)$ per iteration. The achieved acceleration makes it possible to apply this physiologically motivated signal model in real-time applications like speech coding and analyze its performance in state-of-the-art audio compression schemes like MPEG-4 AAC (Brandenburg et al., 2000). The possibility to evaluate the sparse gammatone signal model on a large data set enables the statistical analysis of the selected gammatone parameters for a given sound corpus. According to Barlow's efficient-coding hypothesis (Barlow, 1961), the human auditory filters have been optimized under a strong evolutionary pressure to optimally encode the relevant acoustic stimuli. We analyze the TIMIT speech corpus (Garofolo et al., 1990) and compare the derived gammatone parameters with the known parameters from psychoacoustic experiments.

2. Results

2.1. *Audio coding*

In audio coding schemes such as MPEG-2/4 AAC, the modified discrete cosine transform (MDCT) is used to convert overlapping blocks of the time signal into a frequency-domain representation. With a time shift of N samples, this transform maps $2N$ real numbers onto N real coefficients by using modulated versions of a symmetric window like shown in Fig. 1a. In the older MPEG-1-layer-3 (MP3) standard, a bank of bandpass filters is used in combination with the MDCT. The symmetry property of the used window results from the Princen–Bradley condition the MDCT has to satisfy in order to yield a perfect reconstruction transform (Malvar, 1999). In the MPEG-4 AAC coder, a sine-shaped and a Kaiser–Bessel derived (KBD) window (Oppenheim and Schafer, 1989) can be chosen. In both MP3 and AAC, the window length can be switched between a short and a long window, which allows the encoder to find the best compromise between a high coding gain in stationary sections (long window) and reduced pre-echoes when the signal contains strong transient components (short window).

As argued by Smith and Lewicki (2005), transforms using a block-wise analysis are very sensitive to small time shifts of the incoming signal and do not encode well transients and periodic components that are located in the middle of the overlap region of two adjacent blocks or window positions.

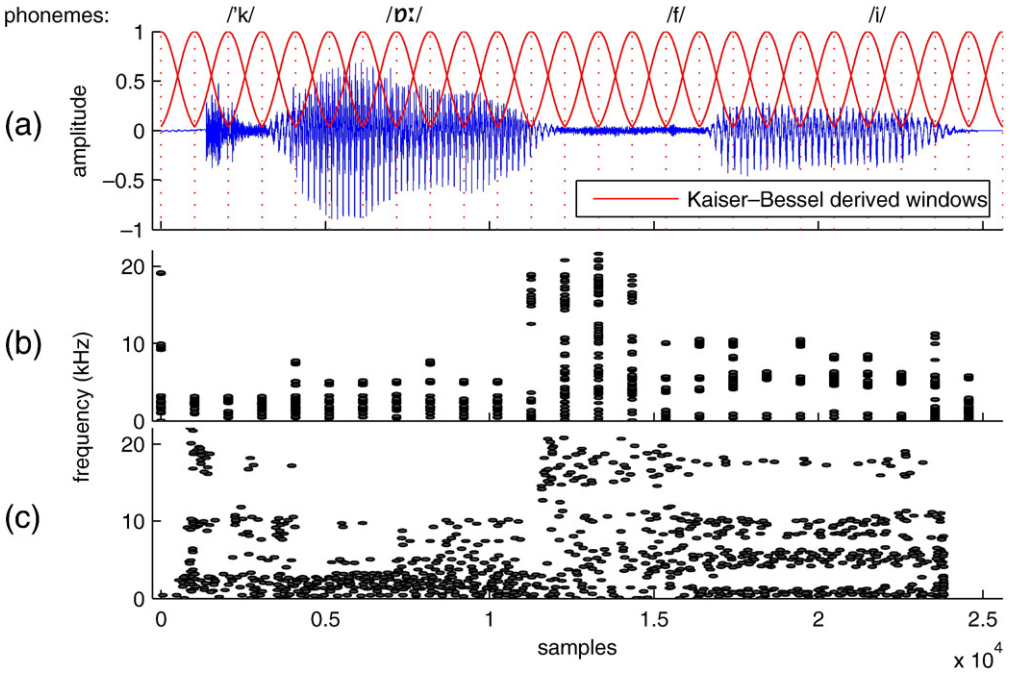

Fig. 1 – (a) Phoneme labeled example from Suzanne Vega's "Tom's Diner" where she sings "coffee". The segments of the overlapping MDCT blocks and the used KBD windows are shown in red. (b) Plot of the 1191 MDCT filterbank coefficients that are not quantized to zero at 16 kbps. (c) Plot of the 1018 MDCT matching pursuit coefficients used at 16 kbps.

In this context, it should be noted that the signal conversion into the frequency domain done by the human cochlea is not rigid in time. The occurrence of signal energy at a specific frequency, due to the frequency-to-place mapping along the basilar membrane, results in a deflection of the corresponding inner hair cells, thereby triggering spikes sent over the corresponding auditory nerves to the brainstem. This activation is solely threshold triggered and not externally clocked like in a DCT or MDCT filterbank.

This property of the human auditory system motivates the application of a shift-invariant signal model like matching pursuit (Mallat and Zhang, 1993) allowing arbitrary time positions. It assumes an additive signal model of the form

$$x[n] = \sum_{i=1}^{K} \alpha_i d_i[n] \quad (1)$$

with the signal $x \in \mathbb{R}^{N\times 1}$, the dictionary coefficients $\alpha_i \in \mathbb{C}$ and the dictionary atoms $D=[d_1 d_2 \ldots d_M] \in \mathbb{C}^{N\times M}$. The shift-invariance is achieved by constructing D using templates like gammatone atoms and adding all their possible shifts to the dictionary. Matching Pursuit is a greedy algorithm that first chooses the atom that best approximates the signal. The contribution of this atom is then subtracted from the signal and the process is iterated on the residual. So the task at the i-th iteration is to minimize the residual

$$r_{i+1}[n] = r_i[n] - \alpha_i d_{k_i}[n] \quad (2)$$

with $d_{k_i}[n] \in D$, k_i being the dictionary index of the atom chosen at the i-th iteration and α_i being the weight describing the contribution of the atom to the signal.

This signal coding paradigm also achieves a sparse signal representation as the increased time resolution results in an overcomplete representation of the signal space and the encoding of a signal is thereby not unique anymore. This overcompleteness allows the matching pursuit algorithm to search for the sparsest encoding in the infinite number of solutions. In contrast, the MDCT atoms form a basis for the signal space where only one unique representation for a signal exists.

In an initial audio coding experiment, we compared the performance of the matching pursuit approach with the traditional filterbank design using the masking model, scalefactor bands and adaptive quantization of the MPEG-4 AAC audio coding reference implementation. We selected the castanets.wav audio signal from the EBU-SQAM audio database (European Broadcasting Union, 1988) due to its transient properties, the TIMIT speech corpus representing the sound class of English speech and the often evaluated music test signal in audio coding, Suzanne Vega's "Tom's Diner" (svega.wav). We compared the coding quality of the MDCT filterbank (FB-MDCT) with the matching pursuit signal models using a MDCT (MP-MDCT) and a gammatone dictionary (MP-GAMMA). The results were evaluated using the objective difference grade (ODG) scale (ITU-R, 2001) computed with an objective prediction method of the perceived audio quality called PEMO-Q (Huber and Kollmeier, 2006) (for details see Experimental procedures). In Fig. 2, the number of used coefficients per second, the signal-to-noise ratio (SNR) and the ODG of the encoded signals at different bitrates are shown.

The matching pursuit algorithm encodes a signal until a given threshold is reached, which was set in this experiment to a fixed SNR for all bitrates (see Table 1). The MDCT filterbank in contrast always results in a perfect encoding if no further quantization is applied. In a next step, the matching pursuit respectively filterbank coefficients are encoded with a given bitrate using the masking model, scalefactor bands and adaptive quantization of the MPEG-4 AAC audio coding standard.

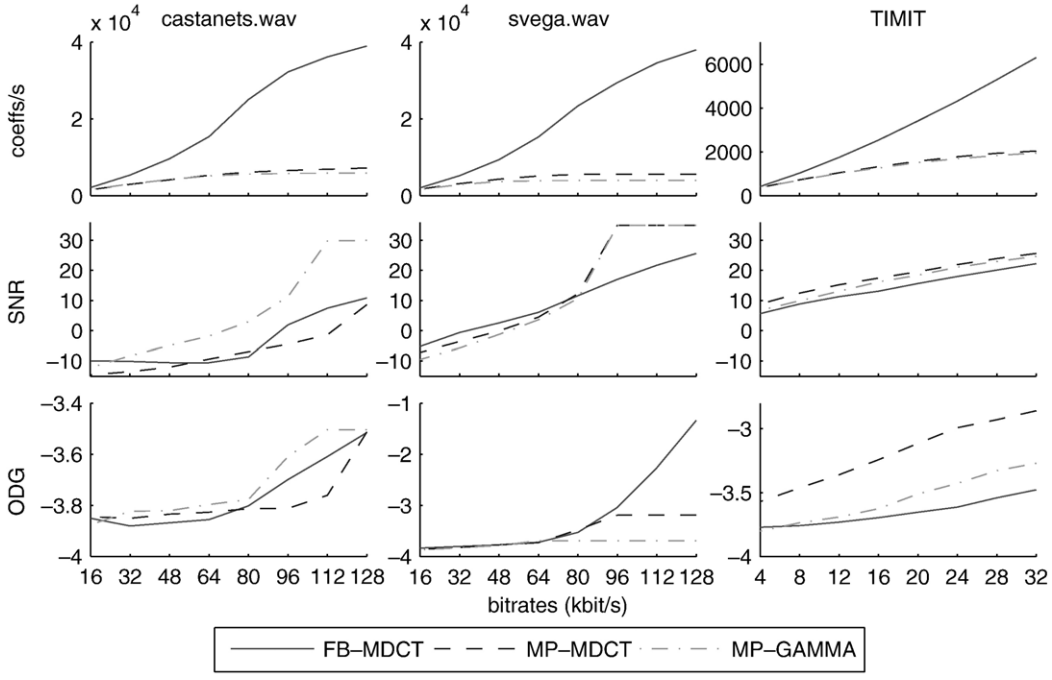

Fig. 2 – The upper row shows the average number of used coefficients per second, the middle row the average signal-to-noise ratio and the lower row the objective difference grade of the encoded signal at different bitrates.

Table 1 – Audio coding settings

Audio signal	Resolution	Sampling rate	Length	SNR threshold
castanets.wav	16 bit	48.0 kHz	7s 939 ms	30 dB
svega.wav	16 bit	44.1 kHz	20s 675 ms	35 dB
TIMIT	16 bit	16 kHz	5h 35 min	30 dB

Thereby the number of used coefficients per second is decreased whenever a coefficient is quantized to zero. This process can be understood by looking at the coefficients per second and SNR results shown in Fig. 2 for the Suzanne Vega song. For the highest bitrate, all coefficients of the matching pursuit signal models can be encoded in the given bit budget due to the initial sparse encoding. This results in a signal representation achieving the preset SNR. At reduced bitrates, quantization is needed to achieve the selected bitrate, first reducing the accuracy of the coefficients and later also reducing the number of used coefficients by quantizing small coefficient values to zero. The non-sparse coefficients of the filterbank signal model in contrast need to be quantized for all bitrates. For the castanets test signal, only the MP-GAMMA signal model results in a sparse representation where all coefficients fit into the bit budget at higher bitrates and achieve the preset SNR. The symmetric MDCT atoms cause for the very transient castanets signal pre-echo artifacts which reduce the SNR. For the TIMIT corpus, lower bitrates common for speech coding applications have been chosen, always resulting in a quantization of the coefficients and an SNR below the preset value. The sparsest encoding of the signal is always achieved by the MP-GAMMA-based audio encodings, followed by the MP-MDCT audio encoding and the MDCT filterbank-based approach.

For the castanets.wav signal, the MP-GAMMA audio coder achieves the highest SNR except for the lowest bitrate. This is also reflected in the audio quality. The MP-MDCT signal model encodes in general less signal energy than the FB-MDCT signal model resulting in a lower SNR. This does not directly show in the predicted audio quality, as for low bitrates the MP-MDCT dictionary achieves a better audio quality despite the lower SNR compared with the filterbank-based audio coder. Analyzing the audio coding at the lowest bitrates shows that for the gammatone dictionary 86% of the available bit budget is used to encode the position of the coefficients using the standard entropy encoder paradigm, resulting in a much stronger quantization of the coefficient amplitudes compared to the filterbank approach.

For the svega.wav signal, the SNR of the matching pursuit audio encoding is higher than the filterbank approach for high and moderate bandwidths and slightly lower for low bitrates. The perceived quality of the encoded audio signal is in contrast for high bitrates much better for the FB-MDCT audio coder and for moderate and low bitrates the ODG is almost identical between the three variations of the audio coder. The Suzanne Vega song also includes a significant amount of 'silent' frames having very low signal energy which are not encoded by the matching pursuit signal model due to the sparseness constraint. Analyzing the framewise ODG of the encoded signals shows that the difference in the ODG values between the matching pursuit and filterbank signal model is due to these frames, which can also be seen in Fig. 3 showing the SNR and ODG for the example in Fig. 1.

Here the FB-MDCT dictionary achieves the lowest perceived audio quality while achieving the highest SNR for low and moderate bitrates. This example shows that while the SNR is a valid measure for general signal coding problems, it is not as significant in audio coding applications as it does not account for psychoacoustic masking effects and does not measure the perceptual distortion. This can be understood by looking at the example in Fig. 1. The background noise at the end of the example is not encoded in the MP-MDCT signal model reducing the overall SNR of the encoding. The last phoneme /i/ in contrast is represented using the matching pursuit-based audio encoder also in the higher frequency bands above 15 kHz where the filterbank approach is not encoding any signal energy for this low bitrate, resulting in a perceptual degeneration of the audio signal. A sparse encoding of a signal results naturally in coefficients with higher coefficient values, which are then not quantized to zero compared to a filterbank approach. So the matching pursuit audio encoder is not only more accurate in time but also generally encodes more high-frequent features than the filterbank approach for a given bitrate.

For the TIMIT speech corpus, the SNR of the MP-GAMMA signal model is slightly lower than for the MP-MDCT for high bitrates and drawing near the SNR of the filterbank implementation for the lower bitrates. The FB-MDCT signal model always achieves the worst SNR. The best perceptual signal quality is always achieved by the MDCT matching pursuit signal model, followed by the MP-GAMMA signal model and the filterbank approach.

The poor performance of the gammatone-based audio coder is an unexpected result as gammatone windows have been

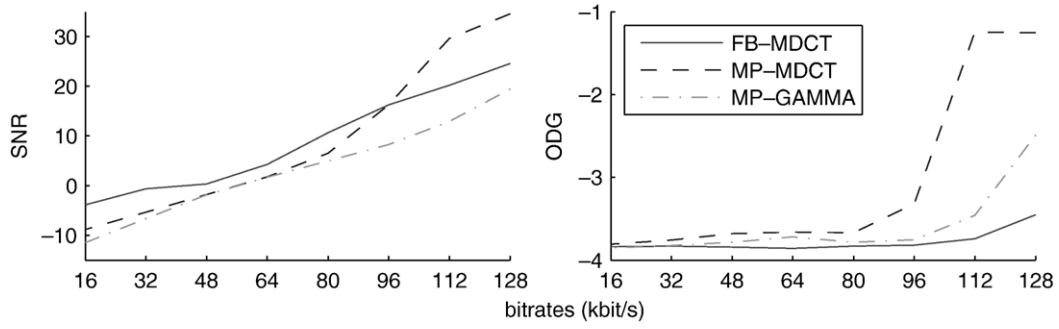

Fig. 3 – The average SNR and ODG for the example in Fig. 1.

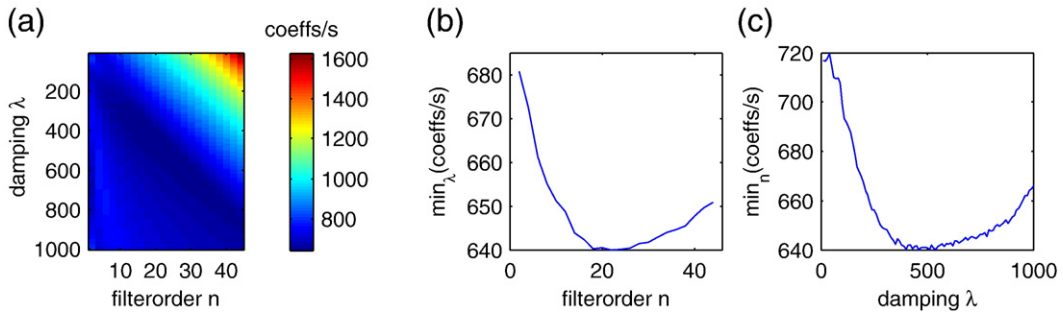

Fig. 4 – (a) Average number of coefficients needed per second for an SNR of 20 dB for the TIMIT speech corpus. (b) Minimal number of coeffs/s for a given filter order. (c) Minimal number of coeffs/s for a given bandwidth.

shown to be optimized to encode speech signals (Lewicki, 2002; Smith and Lewicki, 2006). Despite the fact that the gammatone signal model was always using the lowest number of coefficients to encode a signal to a given SNR, the robustness against quantization errors introduced by the MPEG4 audio encoding scheme showed to be lower compared to the MDCT atoms.

To analyze if using the human parameter values of the gammatone window is optimal for speech, we compared the achieved sparseness on variations of the gammatone window using the fast gammatone matching pursuit signal model. The encoding was stopped when an SNR of 20 dB was reached. We did a rigid scan on the parameter space of the gammatone function using the filter orders η=2, 4, 6,..., 38, 40, 42 and the damping factors λ=10, 20, 30,..., 980, 990, 1000, resulting in 2100 different encodings of the TIMIT database.

For the matching pursuit gammatone signal model, the minimal number of coefficients needed to encode an SNR of 20 dB for English speech is, as shown in Fig. 4, at the filter order n=22 and damping λ=460, resulting in 640.1 coeffs/s. We re-tested the audio coder with these optimized values.

As shown in Fig. 5, the optimized gammatone dictionary achieves now an SNR for the audio encoded TIMIT speech corpus which is almost identical to the MP-MDCT audio codec. This is not reflected in the perceived audio quality, where the MP-GAMMA albeit its sparser encoding is showing a higher impact of quantization errors on the perceived audio quality than the MDCT dictionary. Informal listening tests showed that the gammatone dictionary suffered from stronger musical tone artifacts compared to the MDCT dictionary. It should be kept in mind that this is an initial audio coding experiment. For example, there is a trade-off between the number of initial coefficients generated by the matching pursuit signal model and the consequently needed quantization of these coefficients to achieve the given bitrate. This has not been explored here, the stopping condition of the iterative matching pursuit algorithm was preset to a fixed SNR. More psychoacoustically motivated stopping rules could result in a better audio encoding quality. Additionally the lossless compression stage has been implemented using the standard entropy encoder paradigm to be able to directly compare the different signal models within the MPEG4-AAC audio coding scheme. We have shown that using a significance-tree coder (Strahl et al., 2005) brings advantages for sparse data and shows good performance for audio coding. Furthermore, the quantization algorithm can be optimized for a matching pursuit signal model (Goyal et al., 1998; Frossard et al., 2004).

To further investigate why the gammatone matching pursuit signal model results in a sparser encoding of the TIMIT database than the MDCT matching pursuit signal model for a given SNR, we adapted the gamma-window parameters slowly from an asymmetric to an approximately symmetric window while keeping the maximum of the window fixed (see Table 2). We analyzed its performance on the TIMIT speech database encoding up to an SNR of 20 dB.

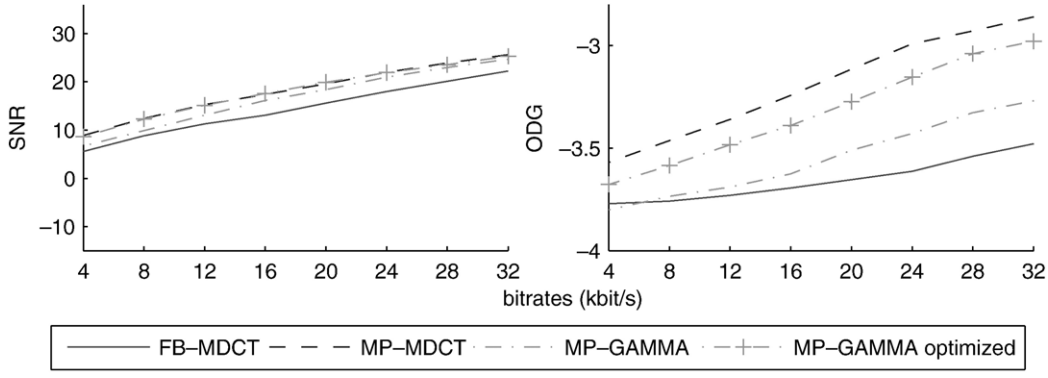

Fig. 5 – The left column shows the SNR, the right the ODG for the TIMIT sound corpus using the different encoding schemes at different bitrates.

Table 2 – Coefficients per second needed to achieve an SNR of 20 dB for gammatone windows with different skewness and the TIMIT sound corpus

Signal model	n	λ	Skewness	coeffs/s
MP-MDCT			0.0	693.2
MP-GAMMA	2	31.25	−1.74	799.0
	4	93.9	−0.71	727.5
	6	156.25	−0.09	682.0
	8	218.9	0.17	663.5
	10	281	0.35	653.0
	12	344	0.49	648.9
	14	406	0.61	651.2
	16	469	0.74	651.5
	18	531	0.82	655.5

The number of coefficients per second decreased from the approximate symmetric gammatone window with a skewness of −0.09 to a minimum at a skewness of 0.49. This is consistent with the earlier derived optimal gammatone parameters $n=22$, $\lambda=460$, which result into a skewness of 0.45. This indicates that a positive skew resulting in an asymmetry is one of the important properties of the gammatone dictionary that leads to an increased sparseness for speech compared to the symmetric Kaiser–Bessel derived window.

2.2. *Physiological signal model*

We conducted a further experiment using a very large dictionary of gammatone atoms with center frequencies ranging from 15.625 Hz to 8000 Hz, increased in 15.625 Hz steps, and damping parameters $\lambda=2\pi b\mathrm{ERB}(f_c)$ ranging from 100 to 5950, increased in steps of 50, resulting in a dictionary size of 966,656,000 atoms per second.

Fig. 6a shows the center-frequency distribution of the selected atoms, which follows the 1/f law normally found for natural signals (Bell and Sejnowski, 1996). The selected gammatone bandwidth parameter b was mostly chosen as 0.19 which differs from the human value of 1.019 (Irino, 1995). The selected bandwidths for every frequency band are shown in Fig. 6b, where the size of the datapoint and its color represent the amount of signal energy that is encoded using this parameters. Atoms encoding less than one percent but more than one tenth of a percent of the TIMIT sound corpus are plotted in gray. The bandwidth parameter used to encode the most energy in the according frequency band is marked by a red square. The matching pursuit algorithm selected mainly atoms with bandwidths below 100 Hz. Also the maximal bandwidth was frequently chosen. It can be noted that the bandwidths encoding the most energy of the signal per frequency band mainly stay below the human bandwidths (Zwicker, 1961). This highly overcomplete dictionary encodes the TIMIT database with an average of 543 coeffs/s.

We further tested if an encoding of the English speech database into a sparse representation limited to 21 different bandwidths for all frequencies with the human values

$$\lambda = 193,\ 262,\ 331,\ldots,\ 3477,\ 4169,\ 4998$$

(Zwicker, 1961) would result in any physiologically known parameter values.

Fig. 7a shows again a frequency distribution following the 1/f law and most of the signal energy was now encoded using $b=0.342$. Fig. 7b shows a similar distribution of the selected atoms like in Fig. 6b. Again, mainly atoms with small bandwidth are preferred. And the most selected bandwidth per frequency is again widening at higher frequencies but staying below the human bandwidth. This fixed bandwidth dictionary encodes the TIMIT database with an average of 607 coeffs/s.

The selection of mainly long dictionary atoms having a small filter bandwidth compared to the human auditory filters is coherent with the signal structure of the TIMIT speech corpus. We computed an average phoneme length of 72.8 ms for the TIMIT database and the filter lengths mainly selected by the matching pursuit algorithm are the two longest atoms with 116 ms and 77.3 ms, as they result in the highest correlation with the signal. The occasional selection of short atoms

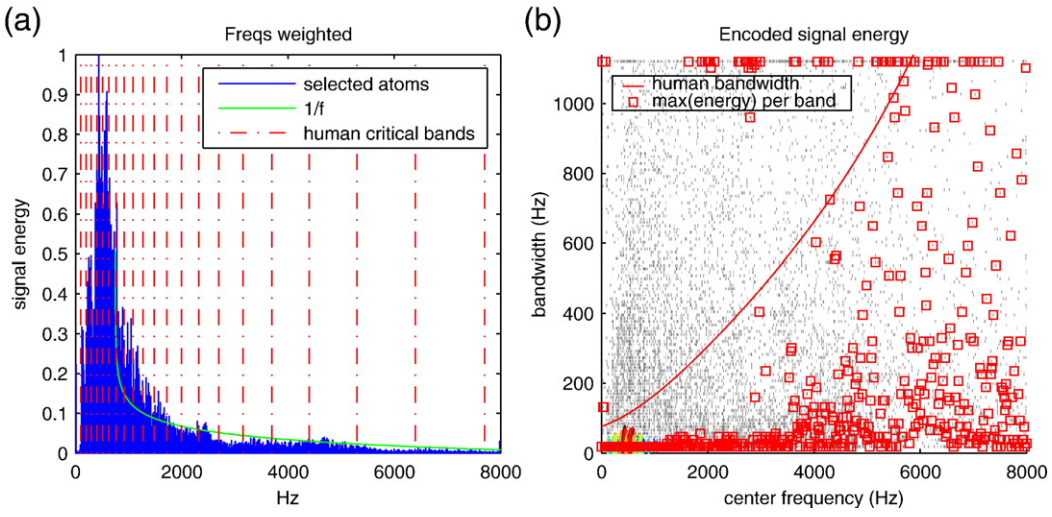

Fig. 6 – (a) Frequency distribution of selected atoms. (b) Encoded signal energy per bandwidth and frequency of selected atoms to encode 20 dB SNR of the TIMIT speech corpus using $\lambda\in$100, 150..., 5900, 5950.

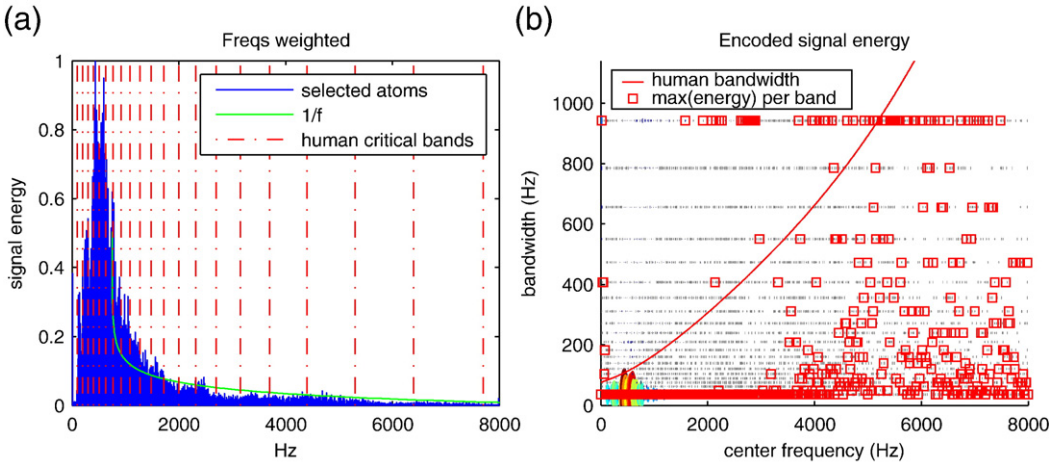

Fig. 7 – (a) Frequency distribution of selected atoms. (b) Encoded signal energy per bandwidth and frequency of selected atoms to encode 20 dB SNR of the TIMIT speech corpus using human $\lambda \in$ 193, 262..., 4169, 4998.

having a long filter bandwidth can be attributed to short signal parts like consonants and to artifacts generated by the matching pursuit algorithm due to its iterative signal decomposition.

3. Discussion

The hypothesis driving the present study is an application of Barlow's efficient coding hypothesis (Barlow, 1961) for audio signal coding. The main efficiency measure in biological systems is the number of spikes needed to transmit a representation of the perceived signal (Laughlin and Sejnowski, 2003). This corresponds in a computer signal model with the number of coefficients used to encode a signal or in other words, how sparse a signal encoding is. One way of increasing the sparseness of an encoding is to increase the correlation of the analyzing filter respectively dictionary atoms with the signal class. We could verify previous results (Lewicki, 2002; Smith and Lewicki, 2006) showing that gammatone atoms have an increased correlation with the English speech, achieving a higher sparseness compared to MDCT atoms. One of the main properties leading to the increased sparseness is the asymmetric time envelope of a gammatone atom as shown in Table 2. This can be understood by the fact that most natural sounds are asymmetric in time, exhibiting a short transient followed by an exponential damped oscillation. This also yields benefits regarding the matching pursuit. The algorithm picks the most energetic atom and for a dictionary with symmetric atoms, it will choose an atom that has also support before the actual start of the attack of the sound. Subtracting this atom from the signal will result in a pre-echo artifacts, creating an artificial signal component just before the transient. The asymmetry of the gammatone window prevents such pre-echo artifact.

Furthermore, the envelope asymmetry indicates that the sparseness constraint is more important for the neurosensory system than a signal analysis achieving a perfect time-frequency resolution. A dictionary atom cannot be arbitrarily concentrated in both time and frequency. Gabor (1946) has shown that, given Heisenberg's uncertainty principle, a symmetric modulated Gaussian window achieves optimal joint time-frequency resolution. For the visual system, such two-dimensional Gabor atoms have been derived as expansion function that generate a nearly sparse signal model and have also been verified in the visual cortex (Olshausen et al., 1996). Compared to a Gabor atom, a gammatone atom has an enlarged analyzing window area in the time-frequency plane of a factor of $\sqrt{\frac{2n-1}{2n-3}}$, n being the filter order of the gammatone (Solbach et al., 1998). The gammatone dictionary showing the highest correlation with the TIMIT database and thereby achieving the sparsest encoding has a higher filter order and thus a better joint time-frequency resolution than a signal model using the human physiological parameters. Gammatone functions with non-human parameters for the TIMIT speech corpus have also been derived by Smith (2006), optimizing a randomly initialized dictionary using a gradient search algorithm. In this study, we applied a full search over the parameter space of the gammatone function, showing that only one local minimum exists for the TIMIT speech corpus. The non-human parameters can be explained by the fact that sparseness is not the only constraint that shaped the auditory system. Consequently, a pure matching pursuit model is not a valid correspondence to the human auditory filter, explaining why the parameters achieving the sparsest encoding for the TIMIT sound corpus do not resemble the human physiological data.

Another important effect is the physiological size constraints. Sparseness is achieved by an overcomplete signal model, whose atoms have overlapping analysis windows in the time-frequency domain. To increase the sparseness of a given signal model, the overlap of theses areas needs to be increased. It has been shown that the increase in length of the auditory epithelia during phylogeny is greater than the increase in the upper frequency, especially in birds and mammals (Manley, 2000), but a momentous increase in frequency resolution is impeded by the size constraint of the hair cell and the cochlear length itself. Consequently, the sparse encoding of an auditory stimulus at the stage of the cochlear is achieved mainly by the high time resolution of its shift-invariant signal model. Analogous to the visual system, where the edge detector filters predicted by a sparseness constraint can be found in the primary visual cortex (Olshausen et al., 1996), it can be assumed that in the auditory system the

sparse coding paradigm will also have an increased influence in the later stages of the auditory pathway compared to the early stages.

Except for Smith and Lewicki (2006), all audio coding applications using a gammatone signal model (Ambikairajah et al., 2001; Feldbauer et al., 2005; Toshio Irino, 2006) applied human parameters and a block-based filterbank model. In general, it has been shown that audio coding applications using a sparse signal model like matching pursuit can have advantages compared to critical sampling signal models like filterbanks or wavelet analysis (Gribonval, 2001; Davies and Daudet, 2006; Krstulovic and Gribonval, 2006; Smith and Lewicki, 2006). Using the union of a MDCT and modified discrete sine transform (MDST) as a signal model, Davies et al. have shown that a twofold oversampling in the frequency domain results for a transient guitar solo test signal in higher SNRs compared to a MDCT signal model (Davies and Daudet, 2006). Smith et al. showed for English speech alike an increased SNR compared to a wavelet or Fourier transform using a signal model that is highly overcomplete in time (Smith and Lewicki, 2006). Replacing the normal quantization with a psychoacoustic masking model, scalefactor bands and adaptive quantization, we measured also an increased SNR and perceived audio quality compared to the block-based signal models for the transient castanets signal and the English speech corpus. The sparseness constraint affects the distribution of the signal energy to few coefficients with high coefficient amplitudes and many coefficients with near zero or zero amplitude. This leads to a distribution of the quantization errors which are mostly either below the absolute hearing threshold or at high sound pressure levels, which is advantageous due to the human logarithmic scale of sound intensity. Additionally fewer coefficients are quantized to zero compared to a filterbank approach, which preserves more of the original signal structure in the quantization process. The synthesis of the atomic signal decomposition introduces also artificial patterns like musical tones, generating a disturbing tonal percept due to equidistant structures on the frequency scale. These artifacts need to be addressed using a postprocessing step in the audio coding design.

4. Conclusion

A matching pursuit gammatone signal model for the English speech using the TIMIT database has been analyzed, showing that, compared to MDCT dictionaries, gammatone dictionaries achieve a sparser encoding for the TIMIT database, indicating that the gammatone atoms are expansion functions that are higher correlated with the English speech class. We also showed that a shift-invariant matching pursuit signal model has advantages in audio coding applications and that a gammatone matching pursuit signal model results in better perceived audio quality for very transient signals due to their asymmetric filter shape. A full search over the gammatone filter parameter space showed that the human auditory system cannot be directly compared to a matching pursuit signal model and that the optimal parameters are not identical to the human physiological values. We showed that the asymmetric filter shape of the cochlear filter can be predicted assuming a sparseness constraint on the signal coding.

5. Experimental procedures

5.1. Gammatone signal model

The gammatone signal model describing the human auditory filter response is defined as (Patterson et al., 1988)

$$g_t(t) = at^{n-1}e^{-\lambda t}e^{2\pi i f_c t} = at^{n-1}e^{-2\pi b\mathrm{ERB}(f_c)t}e^{2\pi i f_c t} \tag{3}$$

with the amplitude a, the filter order n and $\lambda = 2\pi b\mathrm{ERB}(f_c)$ being the damping factor where b defines the proportion to the equivalent rectangular bandwidth (ERB) of the auditory filter which is defined for moderate sound pressure levels (Moore et al., 1990) as $\mathrm{ERB}(f_c) = 24.7 + 0.108 * f_c$ for a center frequency f_c. For humans, the parameters $n = 4$ and $b = 1.019$ have been derived using notched-noise masking data (Irino, 1995).

5.2. Fast matching pursuit for gammatone signal model

Every real-valued atom $d_{\omega,\phi}$ with the frequency ω and the phase ϕ can be associated with a complex atom d_ω and its conjugate $\overline{d_\omega}$. It is

$$d_{\omega,\phi} = \frac{K_{\omega,\phi}}{2}\left(e^{i\phi}d_\omega + e^{-i\phi}\overline{d_\omega}\right) \tag{4}$$

with $K_{\omega,\phi}$ being a normalization factor. The set of atoms $d_{\omega,\phi}$ where only the phase varies lies in the subspace that is spanned by d_ω and $\overline{d_\omega}$. So the orthogonal projection $P_{V_\omega}r_i$ of the residuum r_i onto this subspace $V_\omega = \mathrm{span}\{d_\omega, \overline{d_\omega}\}$ results in a vector lying in the direction of the real atom $d_{\omega,\phi}$ having the optimal phase. This variation is called Molecular Matching Pursuit (Gribonval, 1999) as selecting the best real atom $d_{\omega,\phi}$ is equivalent to finding the best di-atomic molecule V_ω with

$$\sup_{\omega,\phi}|\langle r_i, d_{\omega,\phi}\rangle|^2 = \sup_\omega \sup_\phi |\langle r_i, d_{\omega,\phi}\rangle|^2 = \sup_\omega ||P_{V_\omega} r_i||^2 \tag{5}$$

Using the biorthogonal basis $d_\omega^{\otimes}, \overline{d_\omega^{\otimes}}$ of V_ω with

$$d_\omega^{\otimes} = \frac{1}{1 - |\langle \overline{d_\omega}, d_\omega\rangle|^2}\{d_\omega - \langle d_\omega, \overline{d_\omega}\rangle \overline{d_\omega}\} \tag{6}$$

the orthogonal projection on a di-atomic molecule is computed by

$$P_{V_\omega} r_i = \langle r_i, d_\omega\rangle d_\omega^{\otimes} + \langle r_i, \overline{d_\omega}\rangle \overline{d_\omega^{\otimes}} \tag{7}$$

and it follows

$$||P_{V_\omega} r_i||^2 = \frac{2\mathrm{Re}\{|\langle r_i, d_\omega\rangle|^2 - \langle d_\omega, \overline{d_\omega}\rangle \langle r_i, d_\omega\rangle^2\}}{1 - |\langle \overline{d_\omega}, d_\omega\rangle|^2} \tag{8}$$

The orthogonal projection of the real-valued signal on the space spanned by a complex gammatone atom and its conjugate transpose can be computed completely in the frequency domain using the fast Fourier transformation (FFT), resulting in complexity of $\mathcal{O}(N \log N)$ instead of $\mathcal{O}(N^2)$ per matching pursuit iteration with N being the length of the analyzed signal part. The results in this paper have been computed using the free available Matching Pursuit Toolkit (Gribonval and Krstulovic, 2005) which conducts an initial analysis of the signal and only recomputes in the next iteration the changed signal part, resulting in an overall complexity of $\mathcal{O}(L \log L) + K \cdot (2N-1)\ \mathcal{O}(N \log N)$ with L being

the signal length, K the number of iterations and N the atom length.

It is

$$\langle r_i, d_\omega \rangle = \sum_{t=0}^{N-1} r_i[t] t^{n-1} e^{-\lambda t} e^{-2\pi i \frac{\omega t}{N}} dt = \mathcal{FFT}_\omega \left(r_i[t] t^{n-1} e^{-\lambda t} \right)$$

and

$$\langle d_\omega, \overline{d_\omega} \rangle = \sum_{t=0}^{N-1} \left(t^{n-1} e^{-\lambda t} \right)^2 e^{2\pi i 2 \frac{\omega t}{N}} dt = \mathcal{FFT}_{-2\omega} \left(\left(t^{n-1} e^{-\lambda t} \right)^2 \right).$$

In the audio coding experiment, we omitted the phase information of the gammatone signal model for a valid comparison with the MDCT filterbank signal model. For only real-valued atoms, we have $\langle d_\omega, \overline{d_\omega} \rangle = 1$ simplifying the projection to

$$\langle r_i, d_\omega \rangle d_\omega = \pm \alpha d_\omega$$

Our gammatone atom implementation will be available on the official MPTK web page.

5.3. *Audio coding*

We used the perceptual model, scalefactor bands and adaptive quantization algorithms from the MPEG4 AAC reference implementation (Moving Picture Experts Group, 1999; Painter and Spanias, 2000). The final noiseless coding stage has been adapted for the sparse overcomplete matching pursuit signal models by adding a run-length encoding step before the entropy encoder similar to the encoding step in the JPEG standard.
We used the following signals and settings:

For the Suzanne Vega music sample svega.wav, an increased SNR threshold of 35 dB was necessary to achieve a sufficient coding quality due to its more complex signal structure.

We predicted the perceived audio quality of the encoded audio signals relative to the uncoded signal using a model of auditory perception (PEMO-Q) (Huber and Kollmeier, 2006). The estimated perceived audio quality is mapped to a single quality indicator, the Objective Difference Grade (ODG) (ITU-R, 2001). This is a continuous scale from 0 for "imperceptible impairment", −1 for "perceptible but not annoying impairment", −2 for "slightly annoying impairment", −3 for "annoying impairment" to −4 for "very annoying impairment".

We tested the common bitrates 128, 112, 96, 80, 64, 32, 16 kbps for music and 32, 28, 24, 20, 16, 12, 8, 4 kbps for speech. The matching pursuit signal models were restrained to real-valued atoms to allow a valid comparison to the real-valued MDCT filterbank of the AAC reference implementation. The initial MP-GAMMA signal model used a filter order of 4 and a damping factor of 1000 corresponding to the human filter bandwidth at 1.2 kHz. The skewness of an atom waveform was computed by $y = \frac{E(x-\mu)^3}{\sigma^3}$.

REFERENCES

Aharon, M., Elad, M., Bruckstein, A., 2006. K-SVD: an algorithm for designing of overcomplete dictionaries for sparse representation. IEEE Trans. Signal Proc. 54 (11), 4311–4322.

Ambikairajah, E., Epps, J., Lin, L., 2001. Wideband speech and audio coding using gammatone filter banks. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP01).

Attwell, D., Laughlin, S., 2001. An energy budget for signaling in the grey matter of the brain. J. Cereb. Blood Flow Metab. 21 (10), 1133–1145.

Barlow, H., 1961. Possible principles underlying the transformation of sensory messages. In: Rosenbluth, W. (Ed.), Sensory Communication. MIT Press, Cambridge, pp. 217–234.

Bell, A., Sejnowski, T., 1996. Learning the higher order structure of a natural sound. Netw. Comput. Neural Syst. 7 (2), 261–266.

Brandenburg, K., Kunz, O., Sugiyama, A., 2000. MPEG-4 natural audio coding. Signal Process., Image Commun. 15 (4–5), 423–444.

Chen, S. S., 1995. Basis Pursuit. Ph.D. thesis, Stanford University.

Chen, S.S., Donoho, D.L., Saunders, M.A., 1998. Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. 20 (1), 33–61.

Chi, T., Ru, P., Shamma, S., 2005. Multiresolution spectrotemporal analysis of complex sounds. J. Acoust. Soc. Am. 118 (2), 887–906.

Dau, T., Püschel, D., Kohlrausch, A., 1996a. A quantitative model of the effective signal processing in the auditory system: I. Model structure. J. Acoust. Soc. Am. 99 (6), 3615–3622.

Dau, T., Püschel, D., Kohlrausch, A., 1996b. A quantitative model of the effective signal processing in the auditory system: II. Simulations and measurements. J. Acoust. Soc. Am. 99 (6), 3623–3631.

Davies, M., Daudet, L., 2006. Sparse audio representations using the MCLT. Signal Proc. 86 (3), 457–470.

Davis, G., 1994. Adaptive Nonlinear Approximations. Ph.D. thesis, New York University.

Donoho, D., Tsaig, Y., 2005. Recent advances in sparsity-driven signal recovery. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP05).

Donoho, D., Elad, M., Temlyakov, V., 2006. Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Trans. Inf. Theory 52 (1), 6–18.

European Broadcasting Union, 1988. Sound quality assessment material recordings for subjective tests. URL http://sound.media.mit.edu/mpeg4/audio/sqam/.

Feldbauer, C., Kubin, G., Kleijn, W., 2005. Anthropomorphic coding of speech and audio: a model inversion approach. EURASTP J. Appl. Signal Process. 2005 (9), 1334–1349.

Frossard, P., Vandergheynst, P., Figueras i Ventura, R., Kunt, M., 2004. A posteriori quantization of progressive matching pursuit streams. IEEE Trans. Signal Proc. 52 (2), 525–535.

Gabor, D., 1946. Theory of communications. J. Inst. Electr. Commun. Eng. 93 (III-26), 429–457.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallet, D., Dahlgren, N., 1990. The DARPA TIMIT Acoustic–Phonetic Continuous Speech Corpus CDROM. NTIS order number PB91-505065.

Goodwin, M.M., 1998. Adaptive Signal Models: Theory, Algorithms, and Audio Applications. Kluwer Academic Publishers, Boston.

Gorodnitsky, I., Rao, B., 1997. Sparse signal reconstruction from limited data using FOCUSS: are-weighted minimum norm algorithm. IEEE Trans. Signal Proc. 45 (3), 600–616.

Goyal, V., Vetterli, M., Thao, N., 1998. Quantized overcomplete expansions in IR N: analysis, synthesis, and algorithms. IEEE Trans. Inf. Theory 44 (1), 16–31.

Gribonval, R., 1999. Approximations non-linéaires pour l'analyse des signaux sonores. Ph.D. thesis, Université Paris IX Dauphine.

Gribonval, R., 2001. Fast matching pursuit with a multiscale dictionary of Gaussian chirps. IEEE Trans. Signal Proc. 49 (5), 994–1001.

Gribonval, R., Krstulovic, S., 2005. MPTK, The Matching Pursuit Toolkit. URL http://mptk.gforge.inria.fr/.

Hoyer, P., 2002. Non-negative sparse coding. Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing, pp. 557–565.
Huber, R., Kollmeier, B., 2006. PEMO-Q: a new method for objective audio quality assessment using a model of auditory perception. IEEE Trans. Audio, Speech Lang. Process. 14 (6), 1902–1911.
Irino, T., 1995. An optimal auditory filter. Proceedings of the IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 198–201.
ITU–R, 2001. Recommendation BS. 1387-1, Methods for Objective Measurements of Perceived Audio Quality.
Krstulovic, S., Gribonval, R., 2006. MPTK: matching pursuit made tractable. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP06).
Laughlin, S.B., Sejnowski, T.J., 2003. Communication in neuronal networks. Science 301 (5641), 1870–1874 (September).
Lewicki, M., 2002. Efficient coding of natural sounds. Nature Neurosci. 5 (4), 356–363.
Mallat, S., Zhang, Z., 1993. Matching pursuit in a time-frequency dictionary. IEEE Trans. Signal Proc. 41 (12), 3397–3415.
Malvar, H., 1999. A modulated complex lapped transform and its applications to audioprocessing. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP99).
Manley, G.A., 2000. Cochlear mechanisms from a phylogenetic viewpoint. Proc. Natl. Acad. Sci. 97 (22), 11736–11743 (October).
Moore, B., Peters, R., Glasberg, B., 1990. Auditory filter shapes at low center frequencies. J. Acoust. Soc. Am. 88, 132–140.
Moving Picture Experts Group, Jul 1999. MPEG-4 Audio Version 2 (Final Committee Draft 14496-3 AMD1). ISO/IEC/JTC1/SC29/WG11 N2803.
Neff, R., Zakhor, A., 1997. Very low bit-rate video coding based on matching pursuits. IEEE Trans. Circuits Syst. Video Technol. 7 (1), 158–171.
Olshausen, B., Field, D., 2004. Sparse coding of sensory inputs. Curr. Opin. Neurobiol. 14 (4), 481–487.
Olshausen, B., et al., 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381 (6583), 607–609.
Oppenheim, A., Schafer, R., 1989. Discrete-Time Signal Processing. Prentice-Hall Inc., Upper Saddle River, NJ, USA.
Painter, T., Spanias, A., 2000. Perceptual coding of digital audio. Proc. IEEE 88 (4), 451–515.
Patterson, R., 2000. Auditory images: how complex sounds are represented in the auditory system. Acoust. Sci. Technol. 21 (4), 183–190.
Patterson, R., Moore, B., 1986. Auditory filters and excitation patterns as representations of frequency resolution. In: Moore, B. (Ed.), Frequency Selectivity in Hearing, pp. 123–177.
Patterson, R., Nimmo-Smith, I., Holdsworth, J., Rice, P., 1988. An efficient auditory filterbank based on the gammatone function. APU Rep. 2341.
Smith, E., 2006. Efficient auditory coding. Ph.D. thesis, Carnegie Mellon University.
Smith, E., Lewicki, M., 2005. Efficient coding of time-relative structure using spikes. Neural Comput. 17, 19–45.
Smith, E., Lewicki, M., 2006. Efficient auditory coding. Nature 439 (7079), 978–982.
Solbach, L., Wöhrmann, R., Kliewer, J., 1998. The complex-valued continuous wavelet transform as a preprocessor for auditory scene analysis. In: David, F., Rosenthal, H.G.O. (Eds.), Computational Auditory Scene Analysis. Lawrence Erlbaum Associates, pp. 273–291.
Strahl, S., Zhou, H., Mertins, A., 2005. An adaptive tree-based progressive audio compression scheme. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA05), pp. 219–222.
Toshio Irino, R.D.P., 2006. Dynamic, compressive gammachirp auditory filterbank for perceptual signal processing Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP06), pp. 133–136.
Zwicker, E., 1961. Subdivision of the audible frequency range into critical bands (Frequenzgruppen). J. Acoust. Soc. Am. 33, 248.