GDANSK UNIVERSITY OF TECHNOLOGY
Faculty of Electronics,
Telecommunications and Informatics

Maciej Kulesza

A Novel Estimation Method of Spectral Tonality for Audio Coding Applications

PhD Dissertation

Supervisor:
prof. dr hab inż. Andrzej Czyżewski
Faculty of Electronics,
Telecommunications and Informatics
Gdansk University of Technology

Gdansk, 2010

0 ACRONYMS, ABBREVIATIONS, SYMBOLS AND TERMS

Acronym / Abbreviation	Meaning
AAC	Advanced Audio Coding
AAC-HE	Advanced Audio Coding – High Efficiency
ACELP	Algebraic Code-Excited Linear Prediction
ADPCM	Adaptive Differential Pulse Code Modulation
AM	Amplitude Modulation
AMR-WB	Adaptive Multi-Rate Wideband (Codec)
AMR-WB+	Extended Adaptive Multi-Rate Wideband (Codec)
AR	Autoregressive
AUC	Area Under Curve
CD	Compact Disc
CELP	Code-Excited Linear Prediction
DFT	Discrete Fourier Transform
DPCM	Differential Pulse Code Modulation
DSP	Digital Signal Processor
EBU	European Broadcast Union
FFT	Fast Fourier Transform
FM	Frequency Modulation
FTM	Frequency-derived Tonality Measure
ICC	Inter-channel Coherence
IID	Inter-channel Intensity Difference
IMDCT	Inverse Modified Discrete Cosine Transform
IP	Internet Protocol
IPD	Inter-channel Phase Difference
ISMR	Inverted Signal to Mask Ratio
ISO	International Organization for Standardization
ITU	International Telecommunication Union
ITU-R	International Telecommunication Union – Radiocommunication
LF	Low Frequencies
LP	Linear Prediction
L_{HF}	High Frequencies of the Left channel
L_{LF}	Low Frequencies of the Left channel
M1	Tonality detector combined with MPEG psychoacoustic model 1
MD	Mini Disc

MDCT	Modified Discrete Cosine Transformation
MELP	Mixed Excited Linear Prediction
MPEG	Motion Picture Experts Group
MP3	MPEG layer 3
M/S	Main/Side
MUSHRA	Multi-Stimulus with Hidden Reference and Anchor
MUX	Multiplexer
M_{HF}	High Frequencies of the Mono channel
M_{LF}	Low Frequencies of the Mono channel
NMR	Noise to Mask Ratio
NMT	Noise Masking Tone
ODG	Objective Difference Grade
OPD	Overall Phase Difference
PAC	Perceptual Audio Codec
PC	Personal Computer
PCM	Pulse Code Modulation
PE	Perceptual Entropy
PEAQ	Perceptual Evaluation of Audio Quality
PNS	Perceptual Noise Substitution
PS	Parametric Stereo
QMF	Quadratic Mirror Filter-bank
QIFFT	Quadratically Interpolated Fast Fourier Transform
ROC	Receiver Operating Characteristic
R_{HF}	High Frequencies of the Right channel
R_{LF}	Low Frequencies of the Right channel
SBR	Spectral Band Replication
SDG	Subjective Difference Grade
SFB	Scale-Factor Band
SFM	Spectral Flatness Measure
SLM	Sinusoidal Likeness Measure
SMR	Signal to Mask Ratio
SNR	Signal to Noise Ratio
SQAM	Sound Quality Assessment Material
SSR	Scalable Sampling Rate
STFT	Short Time Fourier Transform
S_{LF}	Low Frequencies of the Side channel (M/S coding)
TCX	Transform-Coded Excitation

TDAC	Time-Domain Aliasing Cancellation
TMN	Tone Masking Noise
TNS	Temporal Noise Shaping
UM	Unpredictability Measure
VoIP	Voice over IP

Symbol	Meaning
$a_-(k_{\mathrm{off}})$, $a_+(k_{\mathrm{off}})$	coefficients of the linear functions modeling characteristics of $u_-(c[k_{\max}])$ and $u_+(c[k_{\max}])$ ratios
$A[k,l]$	amplitude spectrum in the log scale
b	partition index
$b_-(k_{\mathrm{off}})$, $b_+(k_{\mathrm{off}})$	coefficients of the linear functions modeling characteristics of $u_-(c[k_{\max}])$ and $u_+(c[k_{\max}])$ ratios
$b_{\max}$	maximum partition number
$b_c[b,l]$	power ratio in partition
b_{alc}	number of bits to be allocated (experimental codec)
b_{alcAAC}, $b_{\mathrm{alcAAC}}[l]$	number of bits to be allocated with and without frame index (AAC)
b_{avb}	number of bits available for frame encoding (experimental codec)
b_{indx}	frequency band number
b_{val}	median bark value of the partition
c_{ns}	mean UM value obtained when analyzing white Gaussian noise
$c[b]$, $c[b,l]$	weighted UM of partition (with and without frame index)
$c[k]$, $c[k,l]$	UM of spectral component (with and without frame index)
$c'[k_{\max}]$	tonality of spectral peak derived from $\mathrm{ftm}[k_{\max}]$
$c''[k]$	hybrid tonality measure (UM and FTM)
$c_b[b,l]$	normalized $c_t[b,l]$
$c_t[b,l]$	unpredictability measure of partition convolved with spreading function
d	maximum relative detune of tonal component between two successive frames of analysis
d_{FM}	frequency modulation depth

$e[b,l]$	energy in partition (band)
$e_{cb}[b,l]$	partition energy convolved with spreading function
$e_n[b,l]$	normalized $e_{cb}[b,l]$
$e_{nl}[m,l_s]$	log energy of *m*-th scale factor band calculated basing on short FFT spectrum
$\overline{e_{nl}}[m,l_s]$	mean log energy in *m*-th scale factor band for 8 following short spectra
e_{HR}	mean square error of the scores given to the hidden reference signal in MUSHRA tests
$e_{\mathrm{part}}[m,l_s]$	energy of scale-factor band
F_{s}	sampling rate
$\mathrm{FTM}\left[k_{\mathrm{smax}}\left[j^{(l)}\right],l\right]$	FTM assigned to spectral maximum
$\mathrm{FTM}_{\mathrm{trk}}\left(T_{\mathrm{trk}}\left[r^{(l)}\right]\right)$	FTM of three-component tonal track
$\mathrm{ftm}\left[k_{\max}\right]$	inverted tonality measure $1-\mathrm{FTM}_{\mathrm{trk}}\left(T_{\mathrm{trk}}\left[r^{(l)}\right]\right)$ assigned to the spectral maxima $k_{\max}\left[i^{(l)}\right]$ detected within *l*-th frame
f_{c1}	cut-off frequency
f_{FM}	frequency modulation rate
f_{Mbias}	bias of frequency estimator employing QIFFT method
$f_{\mathrm{M}}\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right]\right]$	frequency of spectral peak estimated using QIFFT method
f_{p}	fundamental frequency
$f_w[k,l]$	phase spectrum (inherited from MPEG specification)
$\hat{f}_w[k,l]$	linearly predicted phase spectrum
$g\left[k_{\max}\right]$, $g\left[k_{\max}\left[i^{(l)}\right]\right]$	peakiness of spectral maxima (with and without frame index)
g_{thd}	peakiness threshold
$g_{\mathrm{tnl}}[k]$	parameter used for $m_g\left[k_{\max}\right]$ calculation
$h_{\mathrm{thd}}[k,l]$	hearing threshold for particular spectral bin
$h_{\mathrm{qthd}}[b]$	hearing threshold in quite in partition
$\mathrm{ISMR}[m]$	ISMR in scale-factor band
$i^{(l)}$	index of local spectral maximum detected within *l*-th frame

$j^{(l)}$	index of selected local spectral maximum detected within *l*-th frame
k	spectral index (either FFT or MDCT)
k_1, k_2	constants
$k_{\mathrm{high}}[b]$	spectral index corresponding to upper boundary of partition
$k_{\mathrm{high}}[b_{\mathrm{indx}}]$	spectral index corresponding to upper boundary of band
$k_{\mathrm{high}}[m]$	spectral index corresponding to upper boundary of scale-factor band
$k_{\mathrm{low}}[b]$	spectral index corresponding to lower boundary of partition
$k_{\mathrm{low}}[b_{\mathrm{indx}}]$	spectral index corresponding to lower boundary of band
$k_{\mathrm{low}}[m]$	spectral index corresponding to lower boundary of scale-factor band
$k_{\max}\left[i^{(l)}\right]$	spectral index corresponding to the *i*-th local maximum in *l*-th frame
$k_{\min-}\left[i^{(l)}\right]$	spectral index corresponding to the nearest spectral minimum laying below $i^{(l)}$-th spectral maximum
$k_{\min+}\left[i^{(l)}\right]$	spectral index corresponding to the nearest spectral minimum laying above $i^{(l)}$-th spectral maximum
k_N	spectral index in spectrum calculated with $Z_{\mathrm{p}} = 1$ (no zero-padding)
$k_{\mathrm{ngh}}[k]$	spectral index corresponding to component neighboring the spectral peak
$k_{\mathrm{off}}\left[j^{(l)}\right]$	bin offset related to the frequency estimation employing QIFFT method
k_{peak}	index of cross-correlation sequence maximum (between spectrum $X[k,l]$ and $W[k]$)
$k_{\mathrm{s\,max}}\left[j^{(l)}\right]$	spectral index corresponding to the *j*-th local maximum in *l*-th frame
$k_{\mathrm{s\,max}\,N}\left[j^{(l)}\right]$	spectral index corresponding to the selected spectrum maximum in the spectrum calculated with $Z_{\mathrm{p}} = 1$ (no zero-padding) derived from the index detected within spectrum calculated with $Z_{\mathrm{p}} > 1$
$K_{\mathrm{ngh}}[k]$	set of components laying on the both sides of the peak
$K_{\max}^{(l)}$	set of spectral local maxima detected within *l*-th spectrum
$\overline{\overline{K_{\max}^{(l)}}}$	total number of spectral local maxima detected within *l*-th frame
$K_{\mathrm{s\,max}}^{(l)}$	set of selected spectral local maxima detected within *l*-th spectrum

L	hop size of STFT analysis
L_1, L_2	thresholds for tonal track length
l	signal's frame number
l_s	short frame number being a sub-frame of long frame
m	index of scale-factor band
$m_{e-}[k_{\max}]$	energy relation between spectral bin corresponding to the peak and bin having index lower than peak
$m_{e+}[k_{\max}]$	energy relation between spectral bin corresponding to the peak and bin having index higher than peak
$m_g[k_{\max}]$	parameter related to the peakiness $g[k_{\max}]$
$m_{ge-}[k_{\max}]$, $m_{ge+}[k_{\max}]$	parameters used for tonality spreading
m_j	multiplier involved into $\Delta f_{\Phi}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$ estimation
$\mathrm{msgn}(\ )$	modified signum function
M	filter order
N	frame length
N_{sig}	signal length
n	index of time-domain signal sample
$n_b[b,l]$	energy threshold in partition
$\mathrm{NMT}[b]$	NMT in partition
$n_{\mathrm{part}}[m,l]$	Energy threshold in scale-factor band
$n_{\mathrm{sfb}*}$	number of scale-factor bands fulfilling one of the conditions: $t[m]<t_{\mathrm{thd}}$; $z_{\mathrm{sfb}}[m]<2$, $\sigma[m]<6$, or all of them simultaneously. The asterisk symbol should be then replaced with t, z, σ or PNS, respectively.
n_{sfbt}	number of scale-factor bands fulfilling condition $t[m]<t_{\mathrm{thd}}$
n_{sfbz}	number of scale-factor bands fulfilling condition $z_{\mathrm{sfb}}[m]<2$
$n_{\mathrm{sfb}\sigma}$	number of scale-factor bands fulfilling condition $\sigma[m]<6$
n_{sfbPNS}	number of scale-factor bands fulfilling all following conditions: $t[m]<t_{\mathrm{thd}}$; $z_{\mathrm{sfb}}[m]<2$, $\sigma[m]<6$
n_{asfb}	total number of scale-factor bands that are considered to be substituted with noise in particular sound sample
p	parameter of harmonic signal generator
$P^{(l)}$	total number of candidates to three-components tonal track
$p^{(l)}$	index of candidate to three-components tonal track

PE	PE calculated for particular frame (experimental codec)
$\mathrm{PE}_{\mathrm{AAC}}[l]$	PE calculated for particular frame (AAC)
$\mathrm{PNS}[\mathrm{m}]$	boolean indicating whether scale-factor band might be coded using PNS technique
$\mathrm{princarg}(\varphi)$	principle argument function
q_*	ratio between $n_{\mathrm{sfb}*}$ and n_{asfb}.The asterisk symbol should is replaced with t, z_{sfb}, σ or PNS, adequately to the $n_{\mathrm{sfb}*}$
q_t	ratio between n_{sfbt} and n_{asfb}
q_{zsfb}	ratio between n_{sfbz} and n_{asfb}
q_{σ}	ratio between $n_{\mathrm{sfb}\sigma}$ and n_{asfb}
q_{PNS}	ratio between n_{sfbPNS} and n_{asfb}
$r[k]$, $r[k,l]$	magnitude spectrum (with and without frame index)
$\hat{r}[k,l]$	linearly predicted magnitude spectrum
$r_a[k]$	spectrum smoothed using moving arithmetic-average filter
$r_{\cos}[k,l]$	term used for $c[k,l]$ calculation
$r_g[k]$	spectrum smoothed using moving geometric-average filter
$r_{\sin}[k,l]$	term used for $c[k,l]$ calculation
$r^{(l)}$	index of three-component tonal track detected within l-th frame
r_t	subject number (MUSHRA test)
r_{PElev}	constant used for pre-echo control in MPEG psychoacoustic model 2
s	iteration number ($\mathrm{SNR}[m,s]$ determining)
s_{cfc}	common (global) scale-factor
$s_{\mathrm{fc}}[k]$	scale-factor related to MDCT coefficient
$s_{\mathrm{fc}}[m]$	scale-factor related to scale-factor band
$s_{\mathrm{MSH}}[u_t, r_t, w_t]$	score given to the w_t-th sound sample in the u_t-th experiment by the r_t-th subject in MUSHRA test
$\mathrm{SFM}[b_{\mathrm{indx}}]$	SFM of band defined by the boundaries $k_{\mathrm{low}}[b_{\mathrm{indx}}]$ and $k_{\mathrm{high}}[b_{\mathrm{indx}}]$
$\mathrm{SFM}_{\mathrm{max}}$	Maximal value of SFM (usually equals –60 dB)

$\mathrm{SMR}\left[m,l\right]$	SMR in scale-factor band
$\mathrm{SMR}_{\mathrm{m}}$	logarithmic level differences from the minimum masking threshold to the masker level within the particular critical band
$\mathrm{SNR}[b,l]$	SNR in partition
$\mathrm{SNR}\left[m,s\right]$	SNR in scale-factor band
$\mathrm{sprdngf}(\ )$	spreading function
$\mathrm{TMN}\left[b\right]$	TMN in partition
t_{thd}	tonality threshold (scale-factor band)
$t[m]$	tonality of scale-factor band
$T_{\mathrm{cand}}\left[p^{(l)}\right]$	candidate to three-components tonal tracks
$T^{(l)}_{\mathrm{trk}}$	set of three-component tonal tracks
$T_{\mathrm{trk}}\left[r^{(l)}\right]$	three-component tonal track
$t_b[b],\ t_b[b,l]$	tonality of partition (with and without frame index)
u_t	experiment index (MUSHRA test)
$u_{-}(c[k_{\max}])$	$c[k_{\max}-1]/c[k_{\max}]$ ratio
$u_{+}(c[k_{\max}])$	$c[k_{\max}+1]/c[k_{\max}]$ ratio
$u_{-0}(k_{\mathrm{off}}),u_{-0.37}(k_{\mathrm{off}})$	$u_{-}(c[k_{\max}])$ for $c[k_{\max}]\to 0$ and $c[k_{\max}]=0.37$
$u_{+0}(k_{\mathrm{off}}),u_{+0.37}(k_{\mathrm{off}})$	$u_{+}(c[k_{\max}])$ for $c[k_{\max}]\to 0$ and $c[k_{\max}]=0.37$
v	length of the uniform quantizer coding words
$v[k_{\mathrm{peak}},l]$	sinusoidal likeness measure
w_t	index of sound sample used in MUSHRA test
$w[n]$	analysis window used for STFT calculation
w_{PE1}	constant used for $\mathrm{PE}_{\mathrm{AAC}}[l]$ calculation
w_{PE2}	constant used for $\mathrm{PE}_{\mathrm{AAC}}[l]$ calculation
w_{PE3}	constant used for b_{alc} calculation
$W[k]$	discrete Fourier transform of the $w[n]$
$x[k]$	MDCT coefficient
$x[n]$	signal sample

$x_q[k]$	quantized MDCT coefficient
$x_{\mathrm{w}}[n,l]$	frame of the signal weighted by the $w[n]$
$x_{\mathrm{w,z}}[n,l]$	zero-padded $x_{\mathrm{w}}[n,l]$
$X[k,l]$	short time Fourier spectrum corresponding to the *l*-th frame
$X_{\mathrm{F1}}[k+k_1,l]$	magnitude spectrum smoothed using filter F1
$X_{\mathrm{F2}}[k+k_2,l]$	magnitude spectrum smoothed using filter F2
$y[n]$	realization of white Gaussian noise
$z_{\mathrm{bin}}[k]$	flatness assigned to spectral bin
$z_{\mathrm{sfb}}[m]$	flatness of scale-factor band
Z_{p}	zero-padding factor (expressed as a FFT length to the *N* ratio)
α	tonality derived from SFM
α_{p}	Tonality derived from the method proposed in US patent number 5,918,203
$\delta\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$	difference between $\Delta f_{\mathrm{M}}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$ and $\Delta f_{\Phi}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$
ε	threshold of phase difference
η	constant used during $\mathrm{SNR}[m,s]$ determining
$\sigma[m]$	standard deviation of log energy in scale-factor band
$\sigma^2_{\mathrm{ALL}}[r_t]$	variance of the scores given to the evaluated sound sample recording evaluated in MUSHRA test
φ_0	initial phase
$\Gamma[k_{\mathrm{peak}}]$	maximum of cross-correlation sequence (between spectrum $X[k,l]$ and $W[k]$)
$\Phi\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right]\right]$	phase spectrum
$\Delta f_{\mathrm{M}}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$	frequency jump related to the candidate to three-components tonal track calculated employing QIFFT estimator
$\Delta f_{\Phi}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$	frequency jump related to the candidate to three-components tonal track calculated employing phase-based estimator
$\Delta f_1\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$,	frequency jumps used for $\Delta f_{\min}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$ and $\Delta f_{\max}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$ calculation
$\Delta f_{\min}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$	minimal frequency jump corresponding to $T_{\mathrm{cand}}\left[p^{(l)}\right]$

$\Delta f_{\max}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$	maximal frequency jump corresponding to $T_{\mathrm{cand}}\left[p^{(l)}\right]$
$\Delta\Phi\left(k_{\mathrm{s}\max}\left[j^{(l+1)}\right], k_{\mathrm{s}\max}\left[j^{(l)}\right]\right)$	phase difference corresponding to $T_{\mathrm{cand}}\left[p^{(l)}\right]$
$\Delta^{2}\Phi\left(k_{\mathrm{s}\max}\left[j^{(l+1)}\right], k_{\mathrm{s}\max}\left[j^{(l-1)}\right]\right)$	second order phase difference corresponding to $T_{\mathrm{cand}}\left[p^{(l)}\right]$
$\Delta^{2}\phi\left(k_{\mathrm{s}\max}\left[j^{(l+1)}\right], k_{\mathrm{s}\max}\left[j^{(l-1)}\right]\right)$	phase offset corresponding to $T_{\mathrm{cand}}\left[p^{(l)}\right]$

0.1 DEFINITION OF TONALITY AND OTHER TERMS

The measure called *tonality* is used in this dissertation in order to allow quantitative comparison of the spectral bins or the frequency bands of the signal in terms of their noise-like or tone-like characteristics. In fact, it reflects the power ratio between tone-like and noise-like signal components that occupy a particular frequency band. The tonality falls into the [0,1] range, where 0 indicates that the spectral component or signal sub-band is totally noise-like. Reversely, when the tonality equals 1 the spectral bin corresponds to the pure sinusoid of a constant or modulated frequency or the signal subband comprises of one or more of them. The algorithms yielding continuous tonality measures within [0,1] range can be viewed as *scoring classifiers* instead of *discrete classifiers* (detectors) which provide only binary detection results regarding the characteristics of the spectral bins. Depending on the application, either scoring or discrete tonality classifier may be preferred. Obviously, every scoring classifier can be easily turned into discrete classifier when appropriate threshold to the results of tonality measuring is applied [46].

The remaining terms used in this dissertation are briefly defined below.

Term	**Definition / explanation**
3.5 kHz anchor	sound recording having band limited to the 3.5 kHz used during MUSHRA listening tests
Aliasing reduction	procedure implemented in the MP3 codec employed in order to reduce the aliasing related to the downsampling of the band-pass signals produced by the QMF filter-bank
Codec performance	properties of the codec like: provided subjective coding quality versus bit-rate, algorithm complexity, memory requirements, delay introduced by the encoding procedure, etc.
Coding quality	subjective quality of the recordings encoded using particular coding system in selected mode of its operation
Coding quality optimization	method of bits distribution to various codec encoding modules resulting in as high as possible subjective coding quality for

	pre-defined date rate
Data reduction	method allowing to encode particular signal or parameter, so that the disc space required for its storage or bandwidth of canal required for its transmission is reduced
Dynamic range of predictors	ability of the predictor to generate low prediction error in the signal bands containing low energy tonal components when other signal bands contain noise-like components of significantly higher energy
Digital effects	digital signal processing algorithms that can be applied to any audio signal in order to obtain desired effect (e.g. pitch shifting, reverb)
Efficient encoding	codec's ability to provide high coding quality while keeping the bit-rate as low as possible
False positive rate	attribute of discrete classifier - number of negative instances incorrectly classified as a positives divided by all negatives
Frequency masking	phenomenon related to the human auditory system which cause that the quieter sounds are imperceptible when the louder sound is presented to the listener at the same time (simultaneous masking)
Irrelevance of audio signal	signal components present within audio signal which are undetectable by the human auditory system
Lossless audio coding	coding methods producing sound samples of the decoded signal identical to the original ones
Lossy audio coding	coding methods producing sound samples of the decoded signal different to the original ones
Masking phenomena	occurs when the perception of one sound is affected by the presence of another sound (frequency/simultaneous and temporal masking can be distinguished)
Noise-like bands	bands of the signal containing no tone-like components
Partitions	bands of the signal used in the procedure of hearing threshold estimation by the MPEG psychoacoustic model 2
Perceptual audio coding	lossy coding methods introducing coding distortions so that they are least perceptible by the human auditory system
Perceptual entropy	minimum amount of bits that has to be transmitted in order to achieve the transparent audio quality
Perceptual model	algorithm allowing estimation of temporary hearing threshold
Perceptual quantization	quantization method where the quantization noise is shaped according to the properties of the human auditory system
Pre-echo distortions	distortions introduced by the codec when the signal energy changes rapidly within particular signal frame (e.g. when encoded frame is comprised of silent and transient)
Psychoacoustically controlled linear filter	all pole filter having frequency characteristic corresponding to the instantaneous hearing threshold
Redundancy of audio signal	refers to the predictability or statistical dependencies in the signal, which can be removed using lossless compression
Residual signal	difference between the original signal and its processed (encoded) representation

Short-time frequency response	corresponds to the filter characterized by the coefficients which are changed over time
Signal compression	procedure yielding the representation of the signal requiring fewer number of bits to be stored or transmitted
Signal perceptually weighted	signal filtered according to the instantaneous hearing threshold
Signal quality	subjective quality of the recordings encoded using particular system – usually determined basing on the results of the listening tests
Sound listening in critical way	listener ability to detect and classify particular coding artifacts and determine how they affect overall signal quality
Spreading function	determines the frequency range and the corresponding signal levels for particular signal component below which the another stimulus is inaudible
Stimulus	audio signals like pure tones, harmonic sounds, narrow-band and wide-band noise presented to the listener auditory system which cause deflection of the basilar membrane
Transparent coding	coding of the audio signal in such a way that the encoded and original signals are hardly distinguished by the listener
Transposition of spectrum	procedure of shifting the part of the signal spectrum from its original band to the other frequency range.
True positive rate	attribute of discrete classifier - number of positive instances correctly classified as a positives divided by all positives
Quantization of spectral samples	process of approximating a very large set of discrete spectral values by a relatively-small set of discrete symbols or integer values. In MPEG coding methods the MDCT coefficients, calculated basing on the PCM representation of original signal, are further quantized.
Quality measurement	procedure for determining the quality of audio signal employing digital signal processing algorithm
Voicing strength	parameter used in the MELP speech codecs in order to express the tonality of particular signal band

1 INTRODUCTION

Starting from the early 80's, when the Compact Disc (CD) was introduced to the market by Sony and Philips companies, the digital sound recordings have became available for the listeners all over the world. The CD was intended to replace the analog carriers for sound recordings like cassette tapes and gramophone discs. However, the business model behind the recordings distribution was not going to be changed – the listeners still had to visit their local music stores and buy selected albums burned permanently on the CDs. Although after 30 years the CD storage medium is still in common use, the way the audio recordings are distributed to the listeners has changed dramatically. This change is strongly related to the evolution and widespread of the digital transmission networks and Personal Computers (PCs). In fact, the PC has became the usual home equipment, and the internet connectivity has became a natural way of access to various digital resources. That is why, the listeners have started to search, download and transmit the audio recordings using internet instead of buying CDs in their local stores. Obviously, the legal issues emerged. The broadcasting companies have realized that the next step they should do is to switch from the analog to digital way of programs transmission. Among various technical issues related to the digital revolution affecting the way the audio material has been stored and distributed, the limited capacity of storage discs and bandwidth of transmission canals have became the one of the primary importance. The audio recordings have been stored on the CDs using Pulse Code Modulation (PCM) format. This format has been also used for transmission of speech signal with band limited to the 3.4 kHz and represented using 8-bit code words. However, the audio signal occupying frequency range from 20 Hz up to 20 kHz, encoded according to the PCM method with 16-bit code words, requires more than 700 kbps bandwidth in order to be transmitted. Providing transmission canals with such a bandwidth to the common users was a challenge in 90's and even later. The most straightforward method to overcome this limitation was to develop the novel methods of audio coding requiring lower transmission bandwidth.

The International Telecommunication Union (ITU) has standardized the coding algorithms allowing voice transmission with low bit-rate being the evolution of the PCM format [77][79]. These methods were the first step towards the reducing the bit-rate requirements for audio signal storage and transmission. The example of the audio codec operating similar to the PCM-based codecs standardized by the ITU is the APTX-100 [20]. One of the parameters allowing efficiency comparison of various coding techniques is the compression ratio. It is expressed as ratio between the amount of data required to encode the digital signal using the PCM method to the bit-stream produced by the coding algorithm. Since the methods encoding the time-domain samples provide relatively low compression ratio, in the 80's and 90's the various companies were focused on the research related to the high efficient methods encoding the frequency domain samples. Such codecs are usually called transform codecs, because the time-domain samples are transformed into the frequency-domain before encoding. The invented coding techniques were usually lossy, which means that the decoded sound samples are no longer identical to the samples of the original signal. The efficiency of lossy methods is not only related to the compression ratio, but also to the subjective quality of the audio material they provide assuming particular bit-rate. The simplest method to compare the efficiency of various lossy method is to compare the lowest bit-rate they require to encode the audio signal, so that the introduced distortions are imperceptible to the listener. It is also common to label such a coding scenario as transparent coding, which means that the listener can hardly distinguish the difference between the original and encoded signal.

The majority of lossy audio codecs explore the limitations of the human auditory system in order to provide transparent coding quality while keeping the bit-rate as low as possible [92]. The codecs belonging the lossy codecs family operating according to the above-mentioned scheme are: AC–1, AC–2 and AC–3. The AC-3 allows for data rate adjusting between 32 and 640 kbps depending on the number of encoded audio channels and other factors. These codecs were introduced by the Dolby Digital company starting from 1987 and successfully used for digital cinema sound applications [5][40]. The Sony company introduced the ATRAC codec optimized for the Mini Disc (MD) recorder requiring 140 kbps bit-rate for encoding of single channel. The Perceptual Audio Codec (PAC) and its multichannel version was introduced by the AT&T company [20]. Although various codecs were already invented in the beginning of 90's, they were usually combined with the devices or signal processing systems marketed by particular companies. Therefore, the Motion Picture Experts Group (MPEG) established in 1988 has undertaken the standardization of the compression algorithms for video and audio. In fact, the companies which have already developed their own audio codecs and other skilled bodies like Fraunhofer Institute were involved in the collaborative research aimed at the standardized, high-quality and efficient audio codec. The result of their work was the MPEG-1 standard defining codec comprised of three layers. While the first and second layers are in fact time-domain codecs, the third layer is the transform, perceptual audio codec. The third layer of the MPEG-1 is commonly called MP3 codec [67]. Along with the ongoing research in the audio coding field, it became obvious that the MP3 format is suboptimal and its efficiency may be significantly improved. Therefore in the MPEG-2 specification, the successor of the MP3 format known as an Advanced Audio Coding (AAC) was proposed [17][21][70]. Instead of MPEG-1 and 2, the MPEG-4 standard contains more than only well-defined codecs description. The MPEG-4 specifies number of coding tools starting from transform codec, through speech codec, to the codec employing advanced parametric coding techniques [53][71].

The general idea behind the transform, perceptual codecs like MP3, AAC and others is to keep the quantization noise below hearing threshold. It is assumed that if this requirement is met, the transparent coding quality is obtained [17]-[21][66]-[71][86][130]. Obviously, in order to properly shape the quantization noise during coding procedure it is required to model the masking phenomena occurring in the human auditory system. This is what the psychoacoustic model is intended to be used for [67][97]. The psychoacoustic model being a part of considered codecs usually simulates only some basic processes related to the simultaneous masking. Although the various scenarios of masking are described in the literature, the psychoacoustic model usually distinguishes between tone-like and noise-like signal components to simulate Tone-Masking-Noise (TMN) and Noise-Masking-Tone (NMT) scenarios [70][87][109]. This is an important issue, because the tone-like and noise-like stimuli have completely different masking properties [54][58][119][142]. The spectrum of audio signals usually contains also modulated tonal components which can be produced by a singer or an instrumentalist using the vibrato effect or other technique in order to achieve desired musical expression. Although frequency and amplitude modulated stimuli evoke different sensation to the listener comparing to the stationary ones, it can be assumed that up to some modulation ratios they have similar simultaneous masking properties [154][174]. When these kind of components dominate in some critical bands of the processed signal, improper classifying them as noise-like ones may affect the hearing threshold estimate, resulting in the deterioration of coding quality [94][97]. The important issue is that the AAC encoder employs almost identical psychoacoustic model to the one already used in the MP3 encoder. Both of them incorporate the same method for distinguishing tone-like and noise-like components which fails when the tone-like signal components are modulated in frequency or/and amplitude.

It must be pointed out, that there are a lot of psychoacoustic models already implemented and successfully combined with the experimental audio codecs [9][10][16][31][84][120][121][131][132][161]. Some of them do not require distinguishing between noise-like and tone-like components. However, these models are not in common use due to their relatively high complexity and ongoing research related to the novel modeling scheme [9][16][131][132]. Therefore, in this dissertation the model defined in MPEG standard is considered, since it is commonly used in various audio processing applications [17][19][20][53][64][66][67][68][70][71].

It is well known that the MPEG standards define only the general concept of the codecs and provide basic implementation of described codecs [38][39]. The most important part of considered standards is the specification of bit-stream format ensuring compatibility between various implementations of the same codec type. The efficiency of the MP3 and AAC codecs, implemented according to the specifications given in the standards is poor due to the numerous reasons. It may be deduced that one of them is related to the limited efficiency of method used for distinguishing tone-like and noise-like components [39][94][97]. This was probably the reason why developers involved into MPEG standard preparation patented alternative method for tonal components detection [62].

The author of this dissertation decided to develop a novel method for tonality estimation which might operate efficiently in the case of constant frequency and modulated tonal components. The proposed method is intended to be the substitute for method used in the MPEG psychoacoustic model combined with AAC encoder.

The novel coding algorithms including AAC are usually hybrid. They combine techniques known from parametric approach to signal coding with classic transform coding methods known from codecs like MP3 [8][13][22][52][61][111][137][144]. The reason for that is related to the increase demand for codecs allowing to efficiently encode speech, general audio and mixed content [2][122][144][168]. In these codecs the role of tonality estimation method may not be limited only to be the part of the psychoacoustic model. The MPEG-4 standards introduced the Perceptual Noise Substitution (PNS) technique allowing to increase the audio coding efficiency further [61][148]. In this method the signal subbands containing only noise-like components are detected by the encoder and further synthesized with locally generated noise by the decoder instead of being quantized and encoded in the usual way. It is to be shown in this dissertation that proposed tonality estimation algorithm may be used also as a basis for PNS module implementation.

1.1 THESIS STATEMENTS

This dissertation defends the following theses:

1. It is possible to estimate tonality of unmodulated or frequency-modulated sinusoidal components of audio signals through the comparison of their instantaneous frequency variations determined employing both: an estimator processing spectral amplitude samples and estimator processing spectral phase samples.
2. The distortions introduced during perceptual audio coding may be effectively limited employing tonality estimation algorithm proposed in this dissertation.

1.2 RESEARCH AIMS

The research presented in this dissertation is focused on a novel method for tonality estimation of spectral components. The primary aims of the research described in this dissertation are as follows:

1) to develop a novel algorithm providing adequate tonality estimates for constant and modulated sinusoidal components;
2) to combine the proposed algorithm with the MPEG psychoacoustic model 2, in order to verify whether replacing the standard tonality estimator with the proposed one leads to more reliable estimate of hearing threshold;
3) to verify whether proposed tonality estimation algorithm may be used as a basis for detector of the signal bands containing only noise-like component that can be encoded according to the PNS technique;
4) to reveal the importance of tonality estimation to the audio coding efficiency.

2 SELECTED METHODS FOR AUDIO CODING

The audio coding algorithms may be classified into different categories depending on the assumed criterion. When considering their ability to provide representation of the decoded signal identical to the signal before encoding, the lossless and lossy methods may be distinguished [50][56][87][90][112][113][140]. Assuming that the transmission medium does introduce any distortions, in the lossless methods there is no difference between the original and decoded versions of the signal. However, the ratio between the number of bits required to store the original signal and the number of bits needed to store the encoded representation (compression ratio) is limited in such methods and usually is around 2:1 [50][55][90][139]. Consequently, the lossless methods are generally used for audio signal repositories, as they can be easily played back or processed. Contrarily, the compression ratio provided by lossy methods is significantly higher than in lossless methods. However, the lossy methods modify the time-domain or frequency-domain samples of the signal, so that they are no longer identical to the samples of original signal. Among various lossy methods, the most widespread are algorithms employing modeling of the phenomenon related to the human auditory system. The idea behind such algorithms is to encode the audio signal, so that the introduced distortions are not perceived by the listener even if the time-domain waveform is not preserved [21][123]. This group of coding algorithms is especially useful for transmission or broadcasting applications. Lossy audio formats like MP3 or AAC are also the basic formats for portable players, mobile phones or other devices equipped with processor allowing their decoding and playback.

Regarding codecs ability to encode either general audio signals or only human voice, coding algorithms may be divided into the speech codecs and general audio codecs [52][71][89][122]. In the majority of telecommunication applications only the human voice is transmitted. Although the compression ratio is important in such applications, the key issue is also the delay introduced by the encoder and decoder. In order to provide conditions allowing natural conversation, the delay introduced by the coding algorithms and the transmission medium should be reasonable low [30][56][115][169]. Therefore, there is a separated group of coding algorithms allowing very efficient speech encoding, introducing low delay. However, these methods are not well suited for encoding polyphonic audio material. In this case the subjective coding quality is usually poor.

Another criterion allowing classification of coding methods into various groups is the width of the signal band they are able to preserve. Historically, speech codecs are mostly narrow-band codecs. In order to assure a high intelligibility of speech, it is enough to encode only limited band. That is why these codecs encode the speech signal occupying frequencies below 4 kHz [30][89][105]. However, the nowadays speech coding algorithms are able to encode the speech signal up to 7 kHz or even higher, as it allows more natural conversation [3][93][95][126][161]. The general audio codecs usually encode the entire signal band starting from 20 Hz or even lower up to 20 kHz [21][40]. However, the encoded signal band may be limited if the required compression ratio is tend to be high (for example higher than 20:1). In fact, newly developed codecs can usually alter the width of encoded signal band basing on the user requirements, bit-rate constrains or content of the signal [44][52][93][126][137][144][146][153].

Although there is a great variety of coding algorithms, they have some common features related to the architecture they are based on. While some basic algorithms operate on the time-domain samples (waveform codecs) of the entire-band signal, other split the signal into the sub-bands and then encode sub-band samples, separately [20][66][75][76][79][80]. Contrarily to the methods processing the time-domain samples, frequency-domain codecs transform block of the time-domain samples into the frequency-domain in order to encode them appropriately. These codecs are usually called transform codecs [130]. Instead of quantizing the time-domain or frequency-domain samples (e.g. MDCT coefficients which are real numbers), the codec may employ the particular signal model and it encodes only the parameters of this model. This is usually the case for speech codecs, where the signal is assumed to be a mixture of noise and harmonic signal, filtered by the frequency response of the human voice tract [30][89]. The speech and audio signal may be also viewed as a sum of modulated sinusoids of appropriately adjusted instantaneous frequencies, phases and amplitudes. Codecs employing such a sinusoidal model encode parameters of sinusoids instead of signal samples. This approach has been explored by many researches and varieties of such codec architectures were proposed [7][99][117][149][150][151]. The parametric approach is also employed for multichannel coding. The significant reduction of bit-rate required to encode multichannel recording may be obtained by extracting the correlated and uncorrelated parts of the multichannel signals. Since the correlated components of the multichannel recording are common to some extent for few channels, they can be encoded jointly, along with some additional parameters [11][12][24].

Nevertheless various classification methods outlined above, novel coding algorithms are usually hybrid. They can operate in different modes allowing optimization of the bit-rate to signal quality ratio. The codecs installed in the mobile terminals are not longer required only to encode speech efficiently but also general audio and mixed content. Therefore, the transform codec architecture and parametric approach are commonly combined together within the single codec structure [96][137][144]. Furthermore, the general audio codecs incorporate techniques allowing to lower the bit-rate requirements while introducing almost unperceived distortions to the signal. The techniques like Spectral Band Replication (SBR) or Perceptual Noise Substitution (PNS) use limitations of the human auditory system and they usually significantly change the signal structure [37][61][71][107][146][167].

The in-depth description of all possible codec structures is beyond the scope of this dissertation. Instead of this, the architectures of selected codecs is described in the following subsections. The principles of operation of these selected methods may be helpful in revealing the usefulness of the novel method for tonality estimation what is the main subject of this dissertation.

2.1 WAVEFORM CODECS

Waveform codecs attempt to encode the input signal, so that the decoded waveform is as similar as possible (e.g. assuming root mean squared error criterion) to the original waveform. In the following subsections some selected types of waveform codecs are briefly described.

2.1.1 Pulse Code Modulation (PCM)

The Pulse Code Modulation (PCM) refers to the process of quantizing the signal samples, so that the amplitude and time are represented in a discrete form [30][56]. While the sampling interval between two corresponding samples is inverse to the sampling rate, the number of uniform quantization levels is determined by the number of available bits. This discrete-time signal is usually used as an input signal to the advanced coding algorithms. However, using uniform quantizer for speech signal is suboptimal. Therefore, speech signal is usually converter into discrete-time signal according to the μ-law or A-law characteristic before uniform quantizing [75]. Since the signal characteristic is modified, the quantization is non-uniform, indeed. Such an operation results in higher performance of codec.

2.1.2 Differential PCM

The contiguous samples of many audio signals are correlated to each other. Thus, Differential PCM (DPCM) encoder integrates the predictor. Instead of quantizing the raw samples, the prediction error is quantized in this method. The predicted samples are subtracted from the samples of original signal resulting in the prediction error signal. The prediction error signal has significantly lower variance and dynamic range comparing to the original signal. Consequently, the coding quality provided by the pure PCM method may be obtained using DPCM with reduced bit-rate [30][56].

2.1.3 Adaptive Differential PCM

In the DPCM encoder both the predictor and quantizer are fixed (time-invariant). It was observed however, that the coding efficiency may be increased when these modules adapt to the time-varying behavior of the input signal [30][56]. Instead of direct quantization of the error signal as it is in DPCM, in the ADPCM encoder, the error signal is normalized before quantizing and the gain information is transmitted along with quantized error samples and the predictor's parameters [79][80]. The adaptation may be performed backward or forward. When the coding delay is of primary importance, usually the backward adaptation is used. However, encoders employing backward adaptation are far more sensitive to the transmission errors.

2.1.4 Perceptual Predictive Codec

The role of the ADPCM encoder is to reduce the redundancy of the signal being processed. This is obtained by adapting the predictor and quantizer parameters. The interesting concept of improving this coding scheme is to combine with it the psychoacoustic pre- and post-filtering modules. The block diagram of such codec is presented in Fig. 2.1.

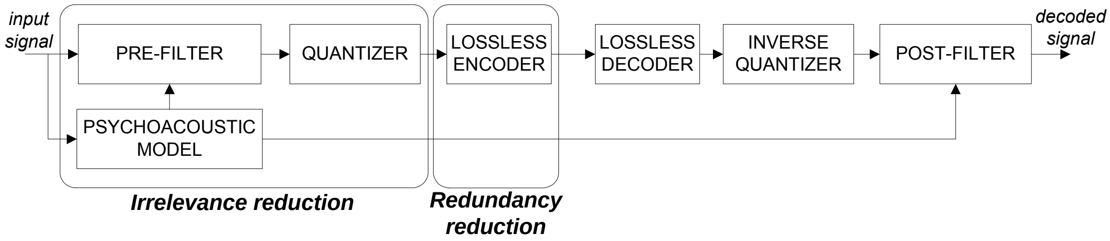


Fig. 2.1 Block diagram of perceptually improved ADPCM codec [147]

It is based on the separating irrelevance (pre/post-filtering) and redundancy (predictive approach similar to the ADPCM) reduction into independent functional units. The input signal is pre-filtered using linear filter of characteristics adapting to the instantaneous hearing threshold in order to reduce irrelevance within the signal. In the next stage the preprocessed signal is encoded according to the lossless method. These operations are inversely applied in the decoder, resulting in the decoded signal. It was proved that presented coding system provides high coding quality at the lower rate than for the standard ADPCM codec bit-rate [147]. Additionally, it inherits the advantages of the ADPCM codec related to the relatively low delay introduced by algorithms of such a type [30].

2.1.5 Subband Codecs

Contrary to the algorithms encoding the time-domain samples of the signal directly, in subband codecs the input signal is first split into the number of subband signals. Next, every subband signal is encoded separately employing for instance one of the methods described in the previous subsections or using some more advanced techniques. The motivation for splitting the signal into the subbands is that the noise introduced into the particular subband does not leak into the other subbands. Therefore, the subband signals may be quantized with the lowest possible number of levels allowing preserving the coding quality resulting in relatively high compression ratio (e.g. 8:1). This concept was the basis for MPEG 1 Layer 1 and 2 codecs [18][20][34][66]. Furthermore, splitting the wideband signal into the subband signals of equal bands gives the opportunity to encode the wideband signal using multiple narrowband codecs. The narrowband codecs may operate parallel to each other and encode subband signals [76]. This approach to signal coding was also the basis for transform coders described in the following subsections.

2.2 TRANSFORM CODECS

In the transform codecs the input sample recordings are processed on the block-by-block basis. The time-to-frequency transformation is applied to every block. In order to avoid quantization of complex spectrum, the MDCT transform is usually used, because the MDCT coefficients are real numbers. The spectral samples are quantized and encoded [21][173].

2.2.1 MPEG 1 Layer 3

The block diagram of the MPEG 1 Layer 3 encoder (called usually MP3) is presented in Fig. 2.2.

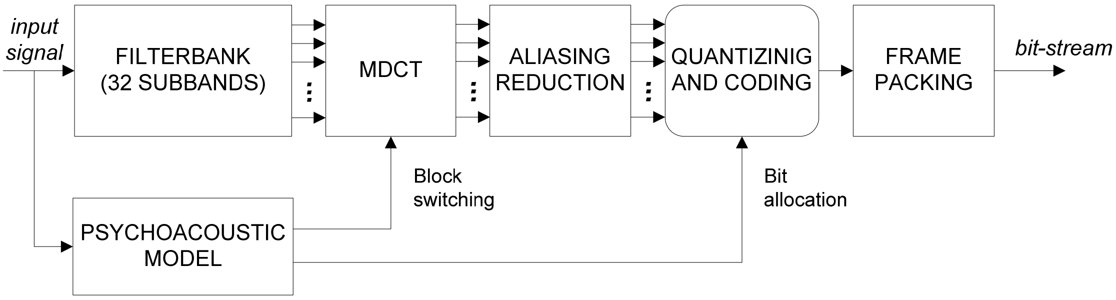


Fig. 2.2 Block diagram of the MP3 encoder [66]

The block of the input signal is fed into the Quadratic Mirror Filter (QMF) bank comprising of 32 equal band filters. Next, the Modified Discrete Cosine Transformation (MDCT) is applied to the these subband signals resulting in 18 spectral samples. Since the subband signals are critically sampled and the filter characteristics overlap, the aliasing effect occurs. Since the frequency characteristics of the filters are known the aliasing may be reduced applying aliasing reduction butterfly procedure to the MDCT coefficients of neighboring subbands [18][66]. The filter-bank of the MP3 encoder is hybrid indeed, as it comprises of the set of QMF filters and MDCT. If the transient is detected then 18 spectral samples are divided into 3 groups and each group is encoded separately. This technique allows for the reduction of the pre-echo distortions [6][35][130][145]. Otherwise, 18 samples are encoded jointly. The MDCT coefficients are than nonlinearly quantized and encoded, so that the introduced distortions are either below the hearing threshold or are minimized depending on the bit-rate constrains. The quantization and encoding process is iterative and involves two nested loops [66][123]. Further, the quantized MDCT coefficients are encoded using Huffman code and the bit-stream is formatted. Among all of the layers of MPEG-1, the

MP3 provides the highest compression ratio. However, it was found that MP3 encoder may deteriorate the quality of signals due to the limited efficiency of transients encoding [17]. Furthermore, it is not well suited to the multi-channel coding, and requires quite high bit-rate for transparent coding. Regarding these and other limitations, the more advanced method for audio coding has been developed by the researchers unified within MPEG.

2.2.2 *MPEG-2*

Although MPEG-2 Advanced Audio Coding (AAC) supports up to the 48 channels, here the default monophonic configuration is described. Since the architecture of this coding system is the basis for experiments carried by the author, this codec is described in great detail. The block diagram of the MPEG-2 AAC encoder is presented in Fig. 2.3.

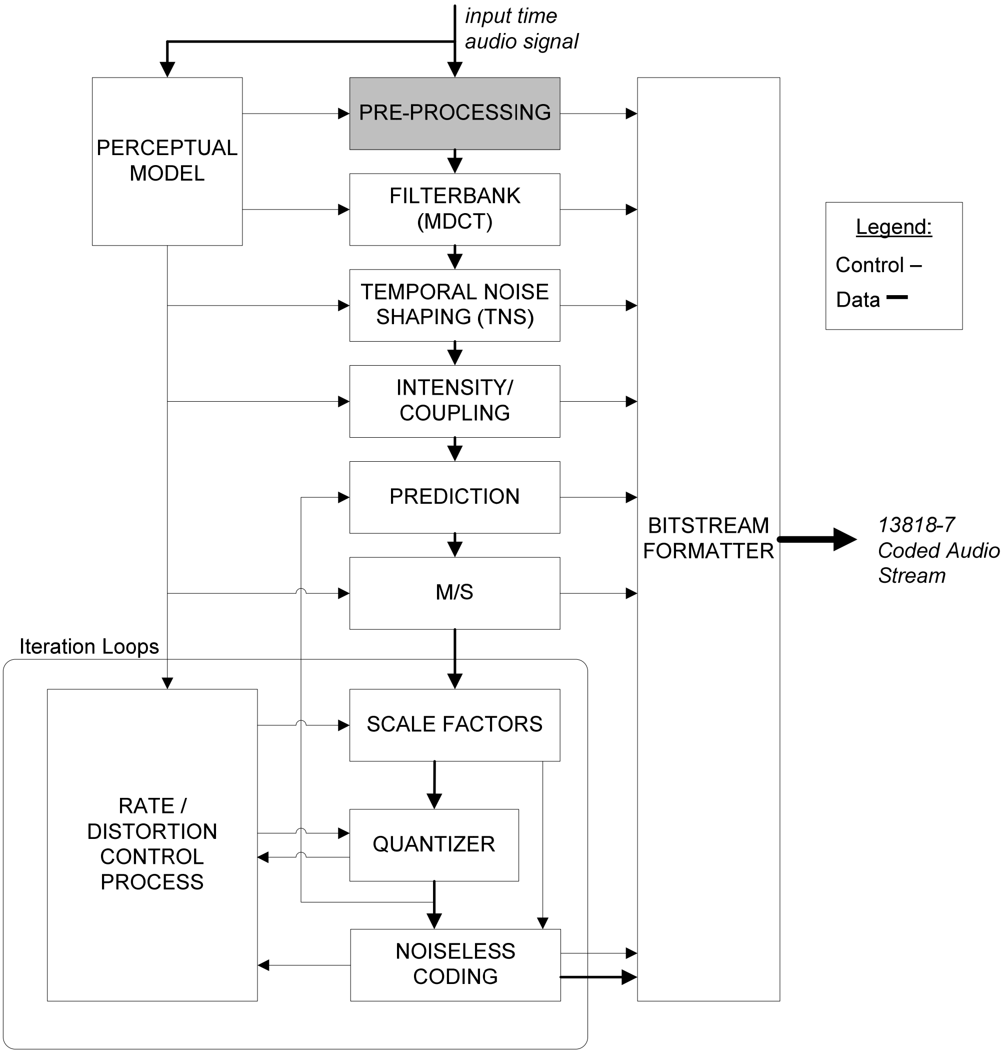


Fig. 2.3 Block diagram of the MPEG-2 AAC encoder [17]

The AAC system offers three profiles: main profile, low-complexity profile and scalable sampling rate (SSR) profile. Here only the main profile is described, as this is the basic codec configuration providing best coding quality at any given data rate. The gain control being the pre-processing stage is applied to the signal only in the SSR profile [17].

The input signal is transformed to the frequency domain using MDCT. At sampling rate of 48000 samples/s the frequency resolution of 23 Hz (2048 samples – long block) and the time resolution of 2.6 ms (256 – short block) may be obtained [102]. In order to control the temporal shape of the quantization noise within each block of transform, the Temporal Noise Shaping (TNS) technique is used. The TNS module filters the spectral samples so that the target spectral coefficients are replaced by prediction residual [60]. The final bit-rate may be reduced for multichannel recordings by encoding only the energy envelopes of the paired channels. This is obtained by employing intensity stereo coding. Obviously, this technique is not used when monophonic signals are processed. Since the stationary signals containing tonal components usually require high data rate in order to provide transparent coding, the spectral samples are predicted in AAC basing on the number of neighboring blocks. The encoder determines for every band whether coding of prediction error provides higher data compression ratio than coding of the MDCT coefficients directly. Basing on this, encoder selects the normal or predicted mode of operation. For stereo signals, the left and right channels are transformed into a mid channel (M) and a side channel (S). While the mid channel is the sum of the left and right channels, the side channel is the difference of the left and right channels [20][21][22][70].

The AAC encoder similarly to the MP3 performs the quantization of the spectral coefficients grouped into the scale-factor bands. The MDCT coefficients within these bands are quantized according to the global and local scale-factors according to the following formula:

$$x_q[k] = \left\lfloor \left| x^{0.75}[k] \right| 2^{0.1875\,(s_{\mathrm{cfc}} - s_{\mathrm{fc}}(k))} + 0.4054 \right\rfloor \tag{2.1}$$

where $x[k]$, k=1, …, 2048 (long block) stands for MDCT coefficients and s_{cfc} is the common (global) scale-factor, $s_{\mathrm{fc}}[k]$ are the scale-factors assigned to individual MDCT coefficient within particular scale-factor band [152][164]. The scale-factors are adjusted, so that the distortions introduced by nonlinear quantizing are minimized and the resulting bit-rate is lower or equal to the rate required by the user (rate/distortion control process in the Fig. 2.3). This is obtained by two nested iteration loops. The task of inner iteration loop shown in Fig. 2.4 is changing the scale-factors until the MDCT coefficients may be encoded with assumed data rate.

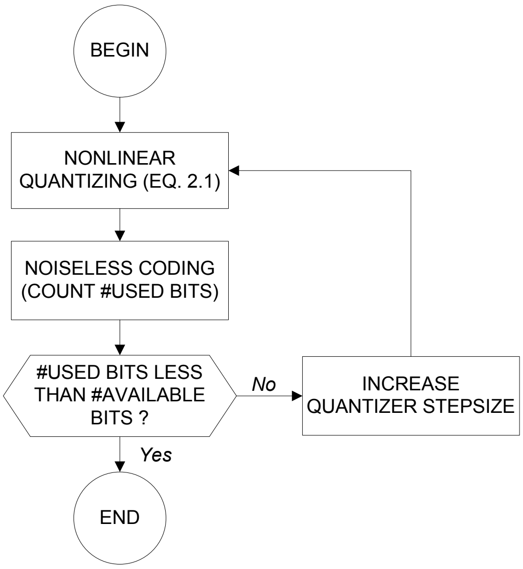

BEGIN
NONLINEAR QUANTIZING (EQ. 2.1)
NOISELESS CODING (COUNT #USED BITS)
#USED BITS LESS THAN #AVAILABLE BITS ?
No
INCREASE QUANTIZER STEPSIZE
Yes
END

Fig. 2.4 Simplified inner iteration loop of the AAC [17]

The task of outer iteration loop (see Fig. 2.5) is amplifying the MDCT coefficients composing particular scale-factor band, so that the demands of the psychoacoustic model are fulfilled as far as possible [17][66][70]. The process terminates when all scale-factors are amplified or there are no scale-factor band with distortions exceeding the allowed distortions. There are also other conditions which cause termination of the outer loop. Since iterative process presented in Figs. 2.4 and 2.5 is time consuming, the simple methods allowing to determine the scale-factors have been already proposed. However, these methods do not provide highest possible coding quality assuming particular data rate [4].

All scale-factors are quantized in 1.5 dB step. While the global scale-factor is encoded using 8 bits, all remaining scale-factors are encoded differentially. The spectral coefficients are grouped and interleaved and encoded using Huffman technique. Finally, the bit-stream is formatted according to the specification given in AAC standard [70][71].

The AAC system was developed as a successor of the MP3 algorithm. Generally, it allows reduction of the required bit-rate for audio encoding twice, comparing to the MP3 method. This is obtained by employing longer block length, spectral coefficients prediction and other techniques. Furthermore, incorporating the TNS method results in reduction of the pre-echo distortions introduced by the encoder for the blocks containing transients [60]. The MPEG-4 specification defines the parametric coding techniques like Perceptual Noise Substitution (PNS) , Spectral Band Replication (SBR) or Parametric Stereo (PS) which may be combined with the AAC system in order to even further improve its performance [22][24][28][37][53][61][100][107][148][167]. The AAC algorithm is supported by music players produced by the various companies.

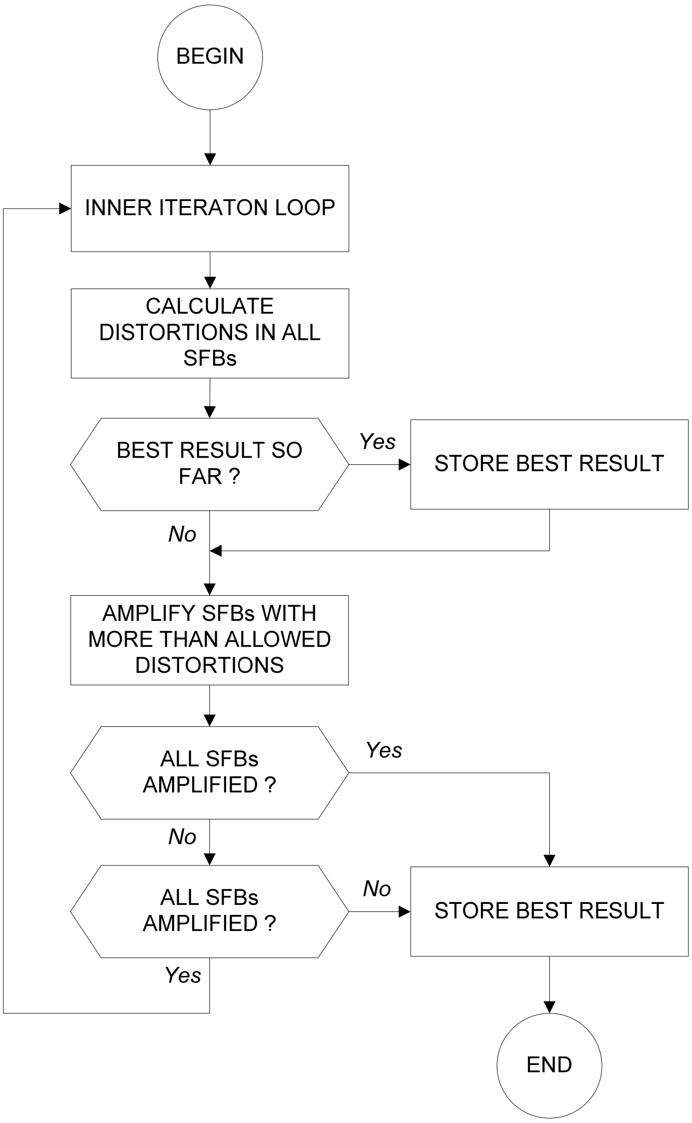


Fig. 2.5 Simplified outer iteration loop of the AAC [17][70]

2.2.3 Other algorithms

The MPEG specifications contain only general description of the codecs and the details related to the bit-stream. Since the implementation details are not included, there are plenty of algorithms which are consistent with MPEG bit-stream but different in details. In fact, the way the particular modules of the codec are implemented are of the prominent importance for its performance. Various codecs consistent with MP3 bit-stream have been recently compared employing subjective tests [135]. The MP3 and AAC principles of operation were also the basis for other coding algorithms like open source Ogg Vorbis [125]. Furthermore, the transform approach to audio coding was employed in the codecs designed for communication purposes which were intended to efficiently encode both the general audio signals, and speech [122].

2.3 PARAMETRIC CODECS

The waveform and transform codecs encode either the time-domain or frequency-domain sample series of audio signal. Contrarily, the parametric codecs try to decompose the incoming signal, in such way that it can be synthesized basing on the set of the predefined parameters [7][29][30][41][56][99][117]. The encoder rule is to estimate these parameters, so that the synthesized signal would sound as similar as possible to the original one. In two following subsections the descriptions of the speech codecs and the codecs employing sinusoidal model are presented.

2.3.1 Speech codecs

The low bit-rate parametric speech codecs usually employ the speech production model presented in Fig. 2.6.

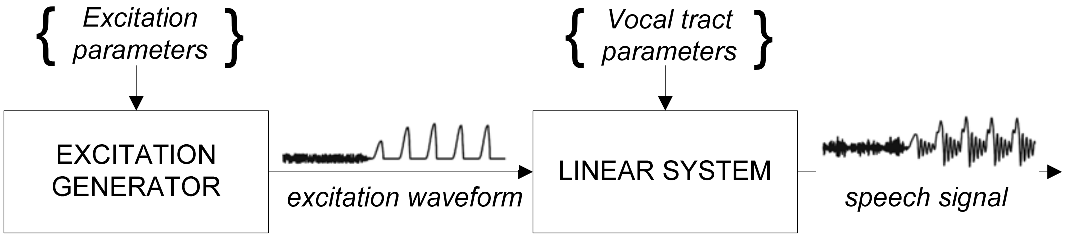


Fig. 2.6 Speech production model [108]

It is assumed in the model, that the excitation generator represents the various modes of sound generation in the vocal tract. The frequency response of the linear system is shaped according to frequency response of the human vocal tract [30][108].

The simplest excitation generator can produce only noise in case of the unvoiced speech segments and periodic impulses related to the pitch and loudness of the voiced speech. However, codecs employing such a simple generator deteriorate the signal quality significantly, because they do not encode the transients of the speech appropriately [30][56][89]. Although the general concept behind the up-to-date speech codecs is almost identical to the concept of codec mentioned above, they differ mainly in the excitation model and method of its parameter encoding. Among various codec architectures, the most commonly used is the CELP (Code Excited Linear Prediction).

In CELP codecs the excitation is determined by the encoder on the analysis-by-synthesis basis (closed loop analysis) using set of predefined excitations stored in the codebook. In fact, the representation of the code-words stored in the encoder codebook and the way the excitation is synthesized vary in CELP-based codecs significantly. The locally synthesized frame is subtracted from the original one resulting in the residual signal. Moreover, the residual signal is psychoacoustically weighted by the curve corresponding to the hearing threshold. The encoder selects the code-word allowing to synthesize the particular speech frame, so that the introduced distortions are minimized [30][56][78][82][143]. Usually, the residual signal is weighted by the curve derived from the spectrum envelope. However, it was proved recently that the coding quality may be improved by replacing this simple mechanism with the more advanced one. Coding quality may be improved when the weighting curve is derived from the hearing threshold generated by the psychoacoustic model [161].

As it was mentioned previously the excitation signal is filtered according to the frequency characteristic of the vocal tract. Usually, the all-pole filter reflecting the characteristic of the vocal tract is used for estimating the speech spectrum envelope. Since the limited number of formants of speech signal can be distinguished in the frequency range up to the 4 kHz, it is enough to represent the spectrum envelope using 10 linear predication (LP) coefficients [78][81][82]. For wideband speech codecs preserving the signal band up to the 7 kHz, 14 or more LP coefficients are required [158].

The CELP-based codecs are sometimes classified as hybrid codecs because of the method they generate and encode the excitation. However, in this dissertation they are treated as a parametric coders. The term ‘hybrid codec’ is used here to the codecs incorporating modules operating according to the different coding schemes – for example codecs combining transform and parametric approaches to the signal coding.

2.3.2 Sinusoidal model based codecs

The general idea behind sinusoidal model based codecs is that the audio signal may be represented as a sum of the modulated sinusoids of time varying amplitudes, frequencies and phases. These instantaneous waveforms may be quantized and transmitted to the decoder instead of the time-domain or frequency-domain samples. Historically, the pure sinusoidal model was applied to the speech signal [117]. It evolved later, resulting in the more advanced codec architectures, allowing efficient encoding of the general audio signals. Furthermore, signal decomposed into the group of sinusoids can be easily manipulated before synthesis which makes these representation very attractive for the applications of the digital effects like pitch shifting, time scale modification and others [116][134][149][150][151][162][173]. The interesting codec architecture employing sinusoidal modeling is presented in the Fig. 2.7.

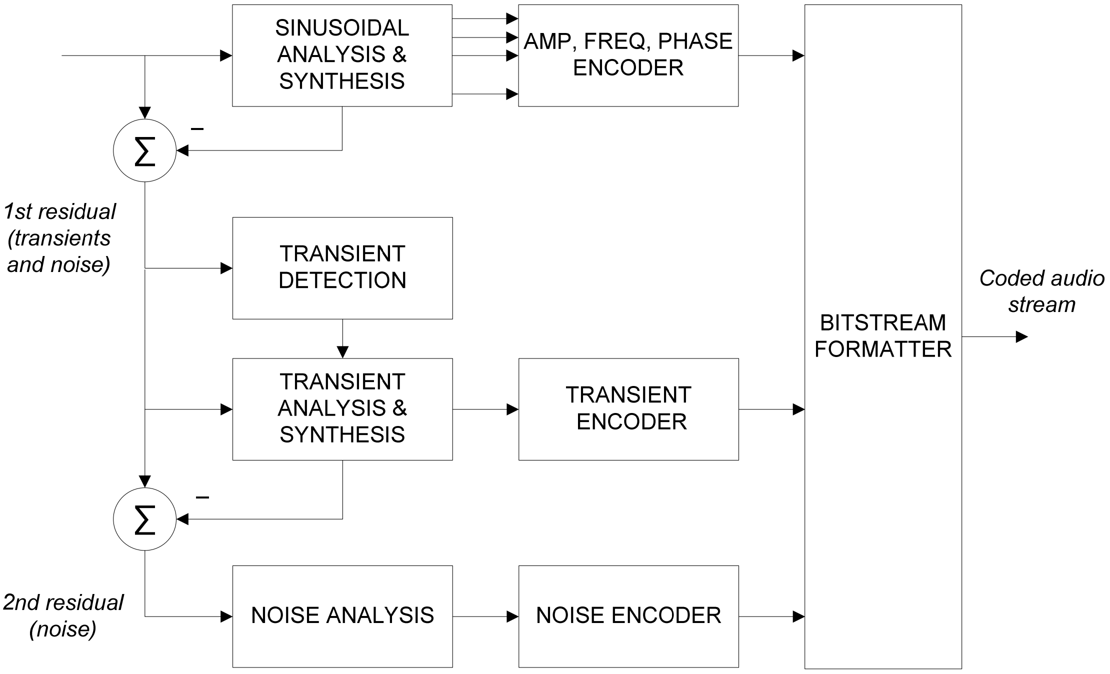


Fig. 2.7 Codec employing sinusoids+noise+transient model [163]

The input signal is first analyzed in order to detect the sinusoids and to estimate their parameters. Next, the sinusoids are synthesized and subtracted from the input signal resulting in first residual containing transients and noise. The transients are then extracted from the first residual signal and encoded. The second residual signal contains only noise [163]. Since the signal is represented as a sum of sinusoids, transients and noise, this model is called sines+noise+transients [99][162]. It may be improved further by employing harmonic analysis technique or psychoacoustic model in order to omit totally masked sinusoids during encoding [111]. Although the concept of such codec is straightforward, the complex analysis must be employed in order to classify the signal components and estimate their parameters. Therefore, such a system usually requires off-line processing [1][15][25][29][41][88][103][104][116][134][141][149][150][151].

2.4 HYBRID CODECS

Since the parametric approach to signal coding allows significant data rate reduction comparing to the transform codecs, the novel algorithms are usually hybrid. A few hybrid codecs may operate either in parametric or transform mode depending on the content of audio program (speech or polyphonic music) [137][144]. Others employ dedicated parametric modules allowing reduction of data rate when it is required [22]. The commonly used technique is parametric encoding of the multichannel recordings, because pairs of the channels (e.g. left and right) are usually highly correlated each to the other. In this subsection the brief description of selected hybrid codecs is presented.

2.4.1 MPEG-4

Instead of the single codec the MPEG-4 specification contains audio coding framework comprised of various coding algorithms. Generally, the structure of AAC codec defined within MPEG-2 is used. The AAC architecture is extended by the parametric coding methods like Perceptual Noise Substitution (PNS), Spectral Band Replication (SBR) and others resulting in AAC-HE v2 codec (AAC – High Efficiency). When using PNS technique the signal subbands containing noise-like components are filled with locally generated noise by the decoder. Consequently, the data rate is reduced, because the MDCT coefficients of noise-like subbands are no longer quantized and encoded [61][148]. In the SBR the upper part of the signal spectrum is reconstructed from the lower part of the spectrum instead of direct encoding (see Fig. 2.8).

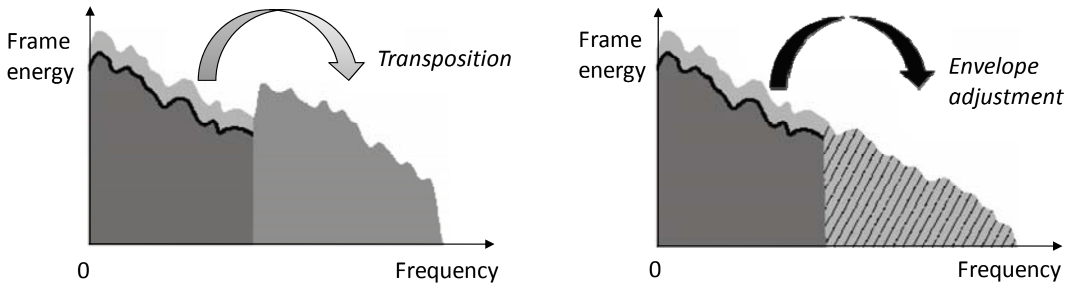


Fig. 2.8 Reconstruction of the upper part of spectrum using SBR technique (left – bandshifting of lower part of the spectrum, right – envelope adjustment) [44]

Firstly, the lower part of the spectrum is transposed to the higher frequency range. Next, its envelope is adjusted so that it is similar to the envelope of the original spectrum. Since only the parameters related to the spectrum envelope and also a few more are needed for reconstruction of upper part of spectrum, the substantial data rate reduction is obtained [42][44][49][175]. The MPEG-4 specification defines also codec module employing sinusoidal model and other parametric coding methods [68][70][71].

2.4.2 Extended Adaptive Multi-Rate Wideband codec

The AMR-WB+ codec is an extended version of the Adaptive Multi-Rate Wideband (AMR-WB) codec [144]. The AMR-WB+ extends the AMR-WB by supplementing it by bandwidth extension, stereo encoding and switching between different coding methods [77]. Contrarily to the AMR-WB codec being indeed a pure speech codec, the AMR-WB+ is intended to encode speech and general audio content with high quality. In Fig. 2.9 the architecture of the AMR-WB+ encoder is presented.

The major part of the encoder is the module responsible for encoding of the lower frequencies (LF) of the signal, while the higher frequencies (HF) are encoded using parametric method. The lower frequencies of the signal are encoded on the analysis-by-synthesis basis using algebraic CELP (ACELP) or transform-coded excitation (TCX) technique. Every frame of the signal is encoded both using ACELP and transform-coded excitation techniques and then distortions introduced by these two methods are compared each to the other. Finally, the method providing lower distortions is selected and the parameters related to it are encoded [144][159]. This codec was the first important step toward the unified coding algorithm allowing efficient encoding of the speech, music and mixed content.

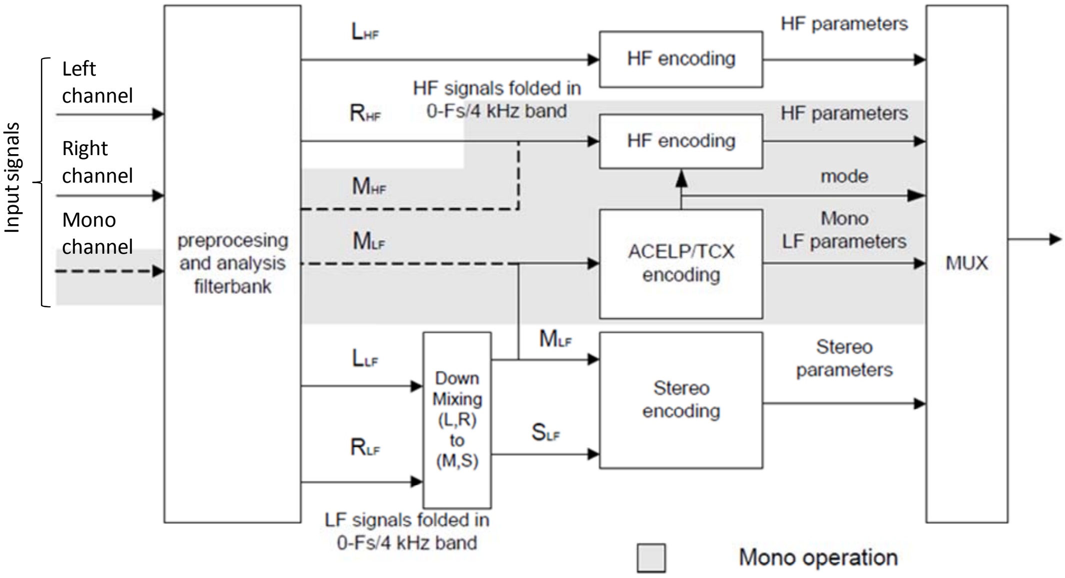


Fig. 2.9 Structure of the AMR-WB+ encoder [159]

The disadvantage of the AMR-WB+ is the high complexity of the encoder. Additionally, due to the high delay introduced by the encoder it is best suited to the broadcast applications rather than to the real-time communication.

2.4.3 Scalable wideband speech and audio coder - G729.1

The G729.1 coder is the scalable wideband (50-7000 Hz) speech and audio coder interoperable with narrow-band speech codec G729 [82]. This codec is based on three-stage structure with 12 embedded layers and generates bit-stream at 8 to 32 kbps rate (see Fig. 2.10).

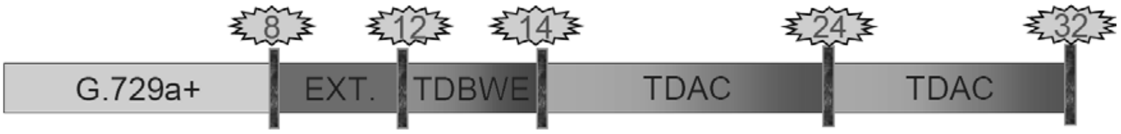


Fig. 2.10 Modes of G729.1 operation for various data rates in kbps

Layer 1 and 2 are in full compliance with G729 bit-stream at 8 and 12 kbps rate, respectively. At 14 kbps rate the signal frequencies up to 7 kHz are encoded. Layers from 4 to 12 provide predictive transform coding referred to the TDAC (Time domain aliasing cancellation). For bit-rates from 14 to 32 kbps the weighted ACELP error signal in the 50-4000 Hz range is encoded in order to improve the quality of signal [137]. The G729.1 is the hybrid codec designed for packetized wideband voice (VoIP) and videoconference applications.

2.4.4 Low bit-rate unified speech and audio coding - MPEG RM0

The family of the MPEG codecs was recently enriched by the unified speech and audio codec named RM0. This codec combines techniques similar to the ones employed in the AAC-HE v2 codec like PNS and SBR with analysis-by-synthesis approach to signal coding. The input signal is classified as a speech-like or music-like and encoded by the one of the two core codecs: frequency domain and linear prediction domain codecs. The listening tests revealed that the MPEG RM0 operates at least as good as AMR-WB+ and AAC-HE v2 or even better [122].

2.4.5 Parametric multichannel coding

The transform and hybrid codecs commonly encode the stereo or paired channels using parametric approach [11][12][23][24]. In the AAC-HE v2 encoder set of three parameters for each of up to 34 subbands are determined [71]:

- the inter-channel intensity difference (IID)
- the inter-channel and overall phase difference (IPD and OPD)
- the inter-channel coherence (ICC)

Basing on the decoded mono signal and the above-mentioned parameters, the stereo image is retrieved by the decoder. The AAC-HE v2 encoder requires only up to 9 kbps in order to preserve the stereo image of the audio signal.

3 FUNCTION OF TONALITY ESTIMATION IN AUDIO CODING APPLICATIONS

Because this dissertation is devoted to the novel method for tonality estimation of spectral components, the selected applications where such a measure plays an important role are described in this Section.

3.1 HEARING THRESHOLD ESTIMATION

3.1.1 Simultaneous masking

The simultaneous masking phenomenon usually called frequency masking is of great importance for lossy coding algorithms. When low level signal e.g. pure tone (maskee) and stronger signal e.g. band limited noise (masker) are close to each other in frequency, the weaker signal may be inaudible due to the simultaneous masking phenomenon. Although the research related to hearing auditory system has been conducted for many years, there are still some aspects of frequency masking which are not explained in great detail [43][85][88][118]. Nevertheless, the basic effects occurring in the basilar membrane when single stimulus is presented are well known [65]. It is well known that the auditory system can be described as a bandpass filter bank. The filter bandwidths are in order of 100 Hz for low frequencies, and up to 5000 Hz for high frequencies. Up to 15.5 kHz frequency 24 critical bands are defined [124]. It is assumed in perceptual audio coding that the distortions introduced in particular frequency band are perceived only in this band. The stimuli is the signal (either masker or maskee) which bends the basilar membrane in the human inner ear. The masking characteristic depends on the stimulus frequency, amplitude and whether it is noise-like or tone-like [58]. The stimulus generates the excitation along basilar membrane that is modeled by the spreading function and a corresponding masking threshold (Fig. 3.1). While the slope of the spreading function is steep for the frequencies lower than the frequency of the stimulus, it is shallow for the frequencies above the stimulus frequency [54][91][174]. Signal-to-Mmask Ratio (SMR) and minimal Signal-to-Mask Ratio (SMR_m) denote the logarithmic level differences from the maximum and minimum masking threshold to the masker level within particular critical band, respectively.

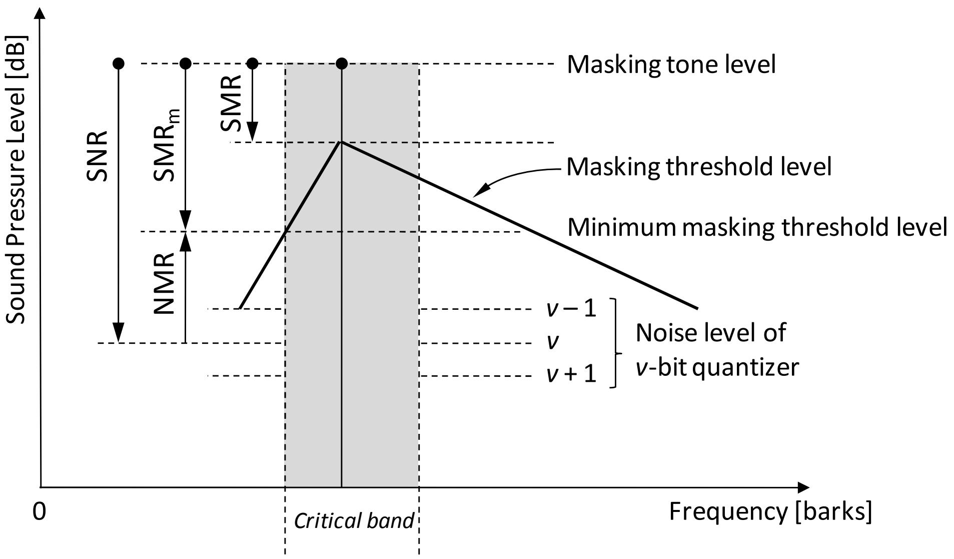


Fig. 3.1 Schematic representation of frequency masking [124][130]

While the SMR for the noise-like components is usually assumed to be around 6 dB regardless the critical band number, the SMR for tone-like components is significantly higher and strongly depends on the frequency of the stimulus [58][101][119][174]. In the perceptual codecs quantization noise is calculated as the sum of squared differences between original amplitudes of spectral components belonging to the particular frequency band (corresponding to critical band) and the quantized ones. When the masking tone is quantized using uniform *v*-bit quantizer, quantization noise might be introduced at level denoted in the Fig. 3.1 as *v*. Then, the Signal-to-Noise Ratio (SNR) is the distance between signal level and introduced quantization noise level. Similarly, Noise-to-Mask Ratio (NMR) is the difference between SNR and $\mathrm{SMR_m}$. The quantization noise is not perceived by the listener when the NMR is higher than 0 dB.

In order to provide transparent coding quality the information related to the required SNR for every critical band must be determined. This task is performed by the psychoacoustic model [9][10][18][19][20][21][31][57][86][87][120][121][123]. Although various psychoacoustic models have been developed up to now, the MP3 and AAC codec being the basis for experiments described in this dissertation employ MPEG psychoacoustic model 2. Therefore, the detailed description of this psychoacoustic model is presented in the next subsection. The symbols notation were inherited from the MPEG standard.

3.1.2 MPEG psychoacoustic model 2

The steps of the procedure yielding estimation of hearing threshold according to the MPEG psychoacoustic model 2 are as follows [18][67][70]:

1. Get the new frame of samples;
2. Multiply the frame by the Hann window and calculate the complex spectrum using FFT (Fast Fourier Transform). Calculate the r and f_w representing the magnitude and phase of the spectral components, respectively.
3. Calculate predicted values of magnitude and phase, basing on two preceding spectra as given by:

$$\hat{r}[k,l] = 2r[k,l-1] - r[k,l-2] \tag{3.1}$$

$$\hat{f}_w[k,l] = 2f_w[k,l-1] - f_w[k,l-2] \tag{3.2}$$

where k is spectral bin number, and l represents frame number.

4. Calculate unpredictability measure c

$$c[k,l] = \frac{\sqrt{r_{\cos}^2[k,l] + r_{\sin}^2[k,l]}}{r[k,l] + |\hat{r}[k,l]|} \tag{3.3}$$

where

$$r_{\cos}[k,l] = r[k,l]\cos(f_w[k,l]) - \hat{r}[k,l]\cos(\hat{f}_w[k,l]) \tag{3.4}$$

$$r_{\sin}[k,l] = r[k,l]\sin(f_w[k,l]) - \hat{r}[k,l]\sin(\hat{f}_w[k,l]). \tag{3.5}$$

5. Calculate the energy and unpredictability in threshold calculation partitions (predefined signal bands)

$$e[b,l] = \sum_{k=k_{\text{low}}[b]}^{k_{\text{high}}[b]} r^2[k,l] \tag{3.6}$$

$$c[b,l] = \sum_{k=k_{\text{low}}[b]}^{k_{\text{high}}[b]} r^2[k,l]c[k,l] \tag{3.7}$$

where b=1, 2, …, $b_{\max}$ ($b_{\max} = 71$ assuming F_s=48 kHz) is partition number, $k_{\mathrm{low}}[b]$and $k_{\mathrm{high}}[b]$ are the partition boundaries.

6. Convolve the partitioned energy and unpredictability with spreading functions

$$e_{cb}[b,l] = \sum_{bb=1}^{b_{\max}} e[bb,l]\,\mathrm{sprdngf}\,(b_{\mathrm{val}}[bb], b_{\mathrm{val}}[b]) \tag{3.8}$$

$$c_t[b,l] = \sum_{bb=1}^{b_{\max}} c[bb,l]\,\mathrm{sprdngf}\,(b_{\mathrm{val}}[bb], b_{\mathrm{val}}[b]) \tag{3.9}$$

where sprdngf() is spreading function, b_{val} is the median bark value of the partition and $b_{\max}$ is the maximum partition number for particular sampling rate. The $e_{cb}[b,l]$ and $c_t[b,l]$ must be then renormalized:

$$e_n[b,l] = \frac{e_{cb}[b,l]}{\sum_{bb=0}^{b_{\max}} \mathrm{sprdngf}\,(b_{\mathrm{val}}[bb], b_{\mathrm{val}}[b])} \tag{3.10}$$

$$c_b[b,l] = \frac{c_t[b,l]}{e_{cb}[b,l]} \tag{3.11}$$

7. Convert $c_b[b,l]$ to the $t_b[b,l]$ and limit its range, so that 0<$t_b[b,l]$ 1

$$t_b[b,l] = -0{,}299 - 0{,}43\ln(c_b[b,l]) \tag{3.12}$$

8. Calculate the required SNR in each partition

$$\mathrm{SNR}[b,l] = t_b[b,l]\,\mathrm{TMN}[b] + [1 - t_b[b,l]]\,\mathrm{NMT}[b] \tag{3.13}$$

where TMN[b] and NMT[b] are tone masking noise and noise masking tone values, respectively. The NMT[b] is constant regardless the partition index and is equal to 5.5 dB in the MP3 encoder and 6 dB in the AAC encoder. While the TMN[b] depends strongly on the partition index in the MP3 encoder, it is constant and equal to 18 dB in the AAC.

9. Calculate the power ratio

$$b_c[b,l] = 10^{\frac{-\mathrm{SNR}[b,l]}{10}} \tag{3.14}$$

10. Calculate the actual energy threshold

$$n_b[b,l] = e_n[b,l]b_c[b,l] \tag{3.15}$$

While in the MP3 encoder the energy threshold is spread across the individual FFT bins, in the AAC this step is omitted.

11. Apply pre-echo control and compare estimated threshold to the threshold in quite

$$n_b[b,l] \leftarrow \max\left(h_{\mathrm{qthd}}[b], \min\left(n_b[b,l], n_b[b,l-1]r_{\mathrm{PElev}}\right)\right) \tag{3.16}$$

where $h_{\mathrm{qthd}}[b]$ is threshold in quite, r_{PElev} is equal to 2 for long blocks, and equal to 1 for short blocks, max() and min() are the functions yielding maximal and minimal value, respectively.

12. Calculate perceptual entropy (Eq. 3.17)

$$\mathrm{PE}_{\mathrm{AAC}}[l] = \sum_{b=1}^{b_{\max}} \left[\left(k_{\mathrm{low}}[b] - k_{\mathrm{high}}[b]\right) \log_{10}\left(\frac{n_b[b,l]}{e[b,l]+1} \right) \right] \tag{3.17}$$

where $b_{\max}$ is the maximum partition number.

13. Basing on the perceptual entropy decide whether single long block or group of short blocks should be encoded. The MPEG specification does not include the details related to the implementation of the long to short block switch. In the open-source implementations of the psychoacoustic model, the short blocks are encoded when the perceptual entropy exceeds specified threshold (eg. 1800) [45].

14. Calculate the SMR(m) in each scale-factor band

$$\mathrm{SMR}[m,l] = 10 \log_{10} \frac{e_{\mathrm{part}}[m,l]}{n_{\mathrm{part}}[m,l]} \tag{3.18}$$

where $e_{\mathrm{part}}[m,l]$ and $n_{\mathrm{part}}[m,l]$ are given by

$$e_{\text{part}}[m,l] = \sum_{k=k_{\text{low}}[m]}^{k_{\text{high}}[m]} r^2[k,l] \tag{3.19}$$

$$n_{\text{part}}[m,l] = \\ \min\left(h_{\text{thd}}[k_{\text{low}}[m],l], h_{\text{thd}}[k_{\text{low}}[m]+1,l], \ldots, h_{\text{thd}}[k_{\text{high}}[m],l]\right)\left(k_{\text{high}}[m]-k_{\text{low}}[m]+1\right) \tag{3.20}$$

Respectively, where $k_{\text{low}}[m]$ and $k_{\text{high}}[m]$ are the low and high boundaries of m-th scale-factor band and $h_{\text{thd}}[k,l]$ is the threshold derived from

$$h_{\text{thd}}[k,l] = \left\{ \frac{n_b[b,l]}{k_{\text{high}}[b]-k_{\text{low}}[b]+1}; \quad k_{\text{low}}[b] \le k \le k_{\text{high}}[b] \right. \tag{3.21}$$

15. Calculate the bit allocation out of perceptual entropy

$$bit_{\text{alcAAC}}[l] = w_{\text{PE1}}\,\text{PE}_{\text{AAC}}[l] + w_{\text{PE2}}\sqrt{\text{PE}_{\text{AAC}}[l]} \tag{3.22}$$

where w_{PE1}=0.3 and w_{PE2}=6 for long blocks, w_{PE1}=0.6 and w_{PE2}=24 for short blocks.

Considering (3.13) it can be noticed, that the reliability of the hearing threshold estimate provided by psychoacoustic models and further coding efficiency directly relates to the performance of the employed tonality estimation method, as the difference in masking between totally noisy and totally tonal components is assumed to be 12 dB in case of the AAC encoder.

3.2 SINUSOIDAL MODELING

In order to distinguish the stochastic and deterministic part of the signal it is required to apply criterion allowing to detect tonal components within signal spectra. Since the stochastic part of the signal is usually modeled using filtered white Gaussian noise, wrong classification of the tonal components as noise-like ones would lead to deterioration of signal quality after synthesis. Additionally, in application based on the sinusoidal model it is required to match the tonal components detected within single spectra into the tonal tracks. A few methods have been proposed in order to accomplish this task [32][116][117][150]. The algorithm proposed in this dissertation may be viewed as an element extending the sinusoidal modeling framework.

3.3 PERCEPTUAL NOISE SUBSTITUTION

One of the techniques allowing increasing the coding quality while preserving the bit-rate requirements is the Perceptual Noise Substitution (PNS). Regardless the waveforms of the noise-like signals are different, they sound alike providing they have the same statistics. Thus, there is no need to encode them in the same manner as the tone-like signals. Instead of quantizing the spectral coefficients of the noise-like components, the pure noise-like signal bands may be filled by the decoder with weighted, locally generated, narrow-band of white Gaussian noise. The bits amount saved by the PNS in the encoder may be used for more accurate encoding of the tonal bands, increasing the perceived coding quality [61][148].

Obviously, improper classification of the tone-like bands as a noise-like ones leads to annoying coding artifacts (unpleasant signal distortions). Thus, the key issue in the PNS technique is the reliable detection of the noise-like bands which may be substituted with the noise in the decoder. Detection of such signal bands is performed usually by the dedicated method operating in parallel to the tonality estimation algorithm used for hearing threshold estimation. One way to obtain this is to split the signal into the bands, and then feed each signal band into the individual predictor. The noise-like bands would have high prediction error indicating that the PNS may be applied to them [61][148]. In the other approaches the mean and variance energy ratios are analyzed or other time domain signal parameters are examined [47][127].

3.4 SPECTRAL BAND REPLICATION

Although the concept of SBR is straightforward, simple bandshifting of the part of the spectrum occupying lower frequencies to the high frequency region and its envelope adjusting is not enough to provide sufficient signal quality. The content of the upper part of the spectrum may significantly differ from the content of lower part. Actually, the subbands in the higher part of the spectrum may contain strong tonal components which are not present in the low frequency region. In this case, the SBR encoder must detect this tonal component and add its parameters to general SBR bit-stream [71]. However, in the subbands occupying higher frequency region tonal components usually have relatively low power comparing to the noise power. In fact, the noise-like components usually dominate in these subbands Therefore the SBR decoder must accurately combine transposed components coming from the lower part of the spectrum with locally generated noise. The tonality estimation method, proposed in this dissertation is not intended to be directly combined with the SBR tool. However, after some modifications it may be used as an element of the SBR or other module allowing regeneration of the high frequency content of the audio signal.

4 SELECTED METHODS FOR TONALITY ESTIMATION

Various approaches to the classification of the spectra components in terms of their tonality have been already proposed. Some of them are restricted only to speech or other simply structured signals [83][155][166][170]. The multi-pitch analysis may be used in order to decompose audio signal into a set of the harmonic sequences [29]. In this dissertation the author focuses mainly on the comprehensive methods which can be used during the analysis of any audio signal in real-time using digital signal processors. These techniques usually employ the short time Fourier transform (STFT) analysis, where as a result of time discretization and quantization, one deals with time-domain frames and spectral (DFT) bins, respectively. In the intra-frame methods, the decisions regarding tonality are taken basing on an analysis results related to a single frame sequence. Contrarily, in the inter-frame method, time evolution of frequency-domain or time-domain samples is inspected.

4.1 AMPLITUDE SPECTRUM ANALYSIS

Generally, the stationary sinusoidal components are viewed as local maxima in signal spectrum. The above statement is valid when the frequency resolution of analysis is higher than the frequency distance between two neighboring sinusoids. The straightforward methods for tonality estimation of spectral components are based on the amplitude spectrum analysis. Three selected methods employing such an analysis are presented in following subsections.

4.1.1 Sinusoidal Likeness Measure

Let the $W[k]$ be the Fourier transform of the analysis window $w[n]$ (n stands for the time-domain sample index). It is possible to inspect the signal spectrum in order to find the spectrum parts corresponding to the shape of the band limited $W[k]$. In order to do this the cross-correlation sequence signal spectrum $X[k,l]$ and window spectrum $W[k]$ is calculated (l stands for index of signal frame). The maxima of cross-correlation sequence indicates the presence of tone-like spectral components. The tonality measure basing on the above-mentioned cross-correlation sequence is commonly named Sinusoidal Likeness Measure (SLM) and is defined as follows [141]:

$$v\left[k_{\text{peak}},l\right]=\frac{\left|\Gamma\left[k_{\text{peak}}\right]\right|}{\left|W\left[k_{\text{peak}}\right]\right|\left|X\left[k_{\text{peak}},l\right]\right|} \tag{4.1}$$

where $\Gamma[k_{\text{peak}}]$ is the maximum of the cross-correlation sequence at frequency corresponding to bin index k_{peak} and $|W[k_{\text{peak}}]|$ and $|X[k_{\text{peak}}]|$ are derived from norms of $W[k]$ and $X[k,l]$. Since the SLM expresses the similarity between the group of bins related to the spectrum of sinusoid of constant frequency and the analyzed signal spectrum, this measure fails in case of tonal components of fast varying frequency or amplitude.

4.1.2 Spectral Flatness Measure

Since most of instruments including human voice produce harmonic sounds, their amplitude spectra have plenty of separated peaks. Instead of analyzing every single peak, the tonality of signal bands may be determined employing the method called Spectral Flatness Measure (SFM). In this method the ratio between geometric and arithmetic means is calculated for every predefined spectrum band as given by

$$\text{SFM}\left[b_{\text{indx}}\right]=20\log_{10}\left(\frac{\sqrt[1+k_h\left[b_{\text{indx}}\right]-k_l\left[b_{\text{indx}}\right]]{\prod_{k=k_l\left[b_{\text{indx}}\right]}^{k_h\left[b_{\text{indx}}\right]}\left|X\left[k,l\right]\right|}}{\frac{1}{1+k_h\left[b_{\text{indx}}\right]-k_l\left[b_{\text{indx}}\right]}\sum_{k=k_l\left[b_{\text{indx}}\right]}^{k_h\left[b_{\text{indx}}\right]}\left|X\left[k,l\right]\right|}\right) \tag{4.2}$$

where b_{indx} stands for frequency band number and $k_l\left[b_{\text{indx}}\right]$, $k_h\left[b_{\text{indx}}\right]$ are the spectral indices representing frequency band boundaries. The tonality factor of a particular band is expressed by

$$\alpha=\min\left(\frac{\text{SFM}\left[b_{\text{indx}}\right]}{\text{SFM}_{\max}},1\right) \tag{4.3}$$

where $\text{SFM}_{\max}$ is usually adjusted to –60 dB. The SFM parameters were originally calculated for entire bandwidth of the signal. Such a global SFM was used in order to determine the hearing threshold according to psychoacoustic model proposed by Johnston [86][87]. Instead of calculating one tonality measure for entire spectrum, MPEG-7 audio description standard suggests calculating similar parameters in ¼ octave bands [59][69].

4.1.3 Method defined within US patent number 5,918,203

The researchers involved into MPEG audio specifications preparation proposed the following additional method for tonality estimation (see Fig. 4.1).

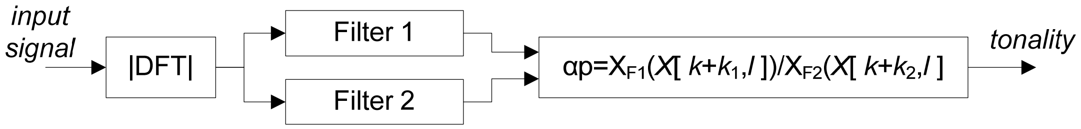


Fig. 4.1 Tonality estimation method defined in patent US 5,918,203 [62]

This method may be viewed as a generalized version of the algorithm yielding SFMs. In the proposed method the amplitude spectrum is smoothed by two filters of different properties resulting in $X_{\mathrm{F1}}[k+k_1,l]$ and $X_{\mathrm{F2}}[k+k_2,l]$, respectively (F1 stands for Filter 1 and F2 for Filter 2). The k_1 and k_2 are constants depending on the filters parameters. The tonality α_p is expressed through the ratio between the results produced by two above-mentioned filters [62]. Note, that a modification of this method is used by the author as an element of the detector incorporated into PNS module, which is further described in Section 8.

4.1.4 Method combined with MPEG psychoacoustic model 1

The MPEG psychoacoustic model 1 employs simple, heuristic, binary detector of tone-like components. In fact, this method does not provide the measure of tonality for spectral components. All local maxima of spectrum are classified as tonal, providing they meet the following criterion

$$A[k,l] - A[k + k_{\mathrm{ngh}}, l] \geq 7 \tag{4.4}$$

where $A[k,l]$ is the amplitude spectrum $|X[k,l]|$ in the log scale (base equal to 10) and k_{ngh} defines a set of the spectral indices corresponding to the components laying on the both sides of the peak. While the k_{ngh} range is narrow for peaks of lower frequencies it tends to be broader for the spectral components occupying higher frequency regions [67]. This method was originally defined for the spectrum generated by the FFT applied to the frame consisting of 1024 time-domain signal samples.

4.2 PHASE-AMPLITUDE SPECTRA ANALYSIS

On the contrary to the noise-like spectral components, the phase of the tone-like components changes deterministically. Therefore, either only the phase spectrum may be analyzed or the joint analysis of phase and amplitude spectra may be carried out in order to determine the tonality of spectral components. A few examples of such a methods are briefly described below.

4.2.1 Unpredictability measure

The instantaneous phase of the constant frequency sinusoids increases linearly in time [128][173]. Therefore, the phase changes corresponding to the particular spectral component may be analyzed in order to determine its tonality. The tonality measure basing on this approach is called unpredictability measure (UM). In this method both: the amplitude and phase values of every spectral component, are linearly predicted basing on two preceding spectra. Next, the predicted values are compared with the values coming from the processed signal frame. The differences between predicted and true phase and amplitude values are directly mapped into the tonality measure [67][70][93]. Since the formulae allowing calculating UM have been already given by (3.3), (3.4) and (3.5) they are omitted here.

4.2.2 Tonality detection using TDAC filter-bank

Instead of analyzing the time evolution of phase like it is in the UM method, it is possible to distinguish the tone-like and noise-like components by analyzing only a single signal spectrum [48][173]. In this method the phase difference between two neighboring spectral MDCT components is examined

$$-\varepsilon \le \mathrm{ctg}\left[phase\{x[k]\} - phase\{x[k-1]\}\right] \le \varepsilon \qquad (4.5)$$

where ε=0.05 for long signal frames consisting of 2048 samples, and ε=0.1 for frames consisting of 512 samples (assuming 48 kHz sampling rate). Although MDCT coefficients are real numbers, they are usually calculated using FFT algorithm. Therefore, it is possible retrieve the phase values employing dedicated procedure. The condition given in (4.5) tends to be met in a random way for signals containing mainly low energy noise. In order to increase the efficiency of tonal components detection, some additional criteria related to the amplitudes of spectral components are applied [48]. Although this method would be an interesting substitute for the UM method, it has been proved that it operates correctly only for unmodulated sinusoids. Consequently, it does not overcome the main drawback of the method used in the MPEG psychoacoustic models. It should be noticed that this method provides only binary results related to the components tonality (every component is either tone-like or noise-like). Therefore, it should be viewed as a substitute for the tonal component detector used in the MPEG psychoacoustic model 1 which provides less reliably hearing threshold estimates than model 2. The advantage of this method is its ability to detect short tonal components, e.g. lasting less than 20 ms.

4.3 TIME-DOMAIN ANALYSIS

Contrarily to the methods described in subsections 4.1 and 4.2, there are methods allowing examination of the entire signal tonality or tonality of its subbands basing on the analysis of the time-domain samples. Although these methods do not assign tonality estimates to the spectral components, they are briefly described here in order to make the list of considered methods complete.

4.3.1 Auto-correlation

In the speech codecs the auto-correlation or other function allowing to determine signals auto-similarity is usually used for pitch period estimation [56]. Additionally, speech codecs usually perform binary classification of signal frames into voiced and unvoiced ones. Although such a classification involves analysis of a group of parameters, the maximum of cross-correlation function calculated for contiguous frames of signal is usually one of them [158]. The Mixed Excited Linear Prediction (MELP) codec employs the module responsible for determining the parameter called voicing strength which may be viewed as a simplified tonality estimate. In the MELP codec the speech is synthesized usually in five subbands [30]. Every subband is assumed to be the mixture of the harmonic signal and noise. The voicing strength determines the proportion of harmonic signal level to noise level. The considered parameter is estimated basing on the normalized auto-correlation value. Additionally, in order to determine the voicing strength of the upper subbands also the flatness of signal envelope is analyzed. There is also a patented implementation of the noise detector incorporated into the PNS module employing similar approach to the signal tonality estimation [127].

4.3.2 *Prediction*

The tonality may be estimated by comparison of the actual time-domain samples with predicted ones. Due to the limited dynamic range of the predictors, signal is usually first split into subbands and down-sampled using a polyphase quadrature filter bank as it is presented in Fig. 4.2.

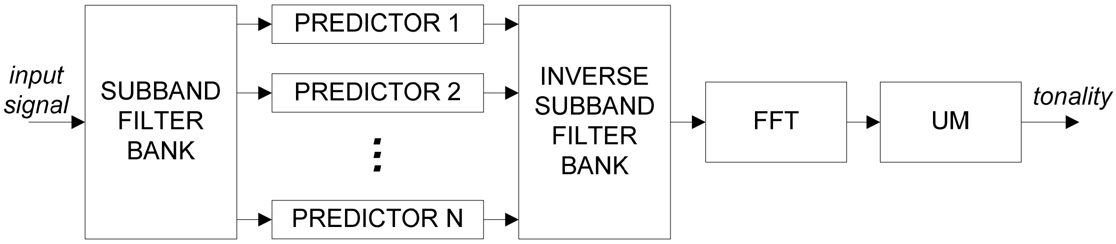


Fig. 4.2 Tonality estimation by prediction in subbands [148]

After this operation, each subband signal is fed into the individual predictor. Next, the inverse filter bank is used in order to combine the signals from the predictors into the entire frequency band signal. The tonality of individual spectral bins may be determined by applying FFT and employing the UM method. This algorithm was proposed as one of the methods allowing detection of the noise-like components and was successfully combined with the PNS module [148].

5 NOVEL METHOD FOR TONALITY ESTIMATION OF SPECTRAL COMPONENTS

5.1 ALGORITHM OUTLINE

The major properties of the proposed algorithm for tonality estimation of spectral components are as follows:

- it exploits both: inter-frame and intra-frame approach simultaneously and incorporates a module responsible for matching single tonal components into the tonal tracks (partials);
- it assigns non-zero tonality measures only to the spectral local maxima, while all other spectral bins tend to have a noise-like characteristic (the tonality spreading over bins neighbouring to the spectrum maxima is described in subsection 7.1.2 and 7.1.3);
- it allows efficient detection of modulated tonal components.

The proposed algorithm detects the local maxima of magnitude spectra corresponding to three contiguous frames of a signal and then matches them to the candidates for tonal tracks (partials). Further the frequency jumps related to the local spectra maxima belonging to these candidates are estimated employing the magnitude-based and phase-based methods. The verification of the candidates tonality is based on the distance between frequency jumps derived from the magnitude-based and phase-based estimators. This distance is also used to estimate so called frequency-derived tonality measure (FTM) for spectral bins belonging to track candidates being verified.

The first frequency jumps estimator employs well known technique of polynomial fitting (quadratically interpolated FFT – QIFFT) to the spectrum maximum and its two neighboring bins [1][36][88][98]. The second proposed estimator is non-standard and was specially developed in order to meet the following requirements:

- it yields inadequate instantaneous frequency values when the spectrum bins involved into the estimation procedure do not correspond to the tonal components (the frequency distance between values obtained using quadratic interpolation and phase-based method should be abnormally high – i.e. higher than half of the frequency resolution of spectral analysis);

- it allows accurate instantaneous frequency estimation of frequency modulated tonal components

A similar approach was previously proposed as an element of sinusoidal modeling framework, however the method proposed here is different [103][133][134]. The block diagram illustrating processing steps of the proposed algorithm is presented in Fig. 5.1. The detailed description of the algorithm along with the definitions of all symbols used are given in the following subsections.

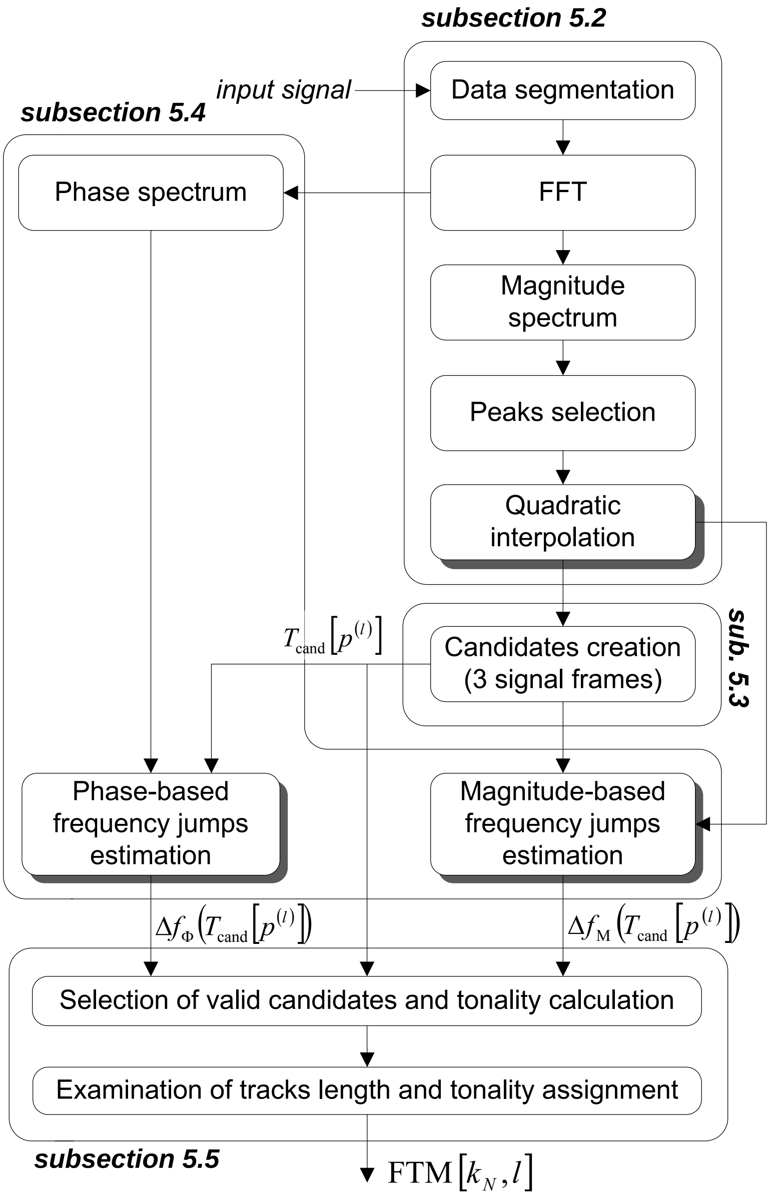

subsection 5.2
input signal
Data segmentation
subsection 5.4
Phase spectrum
FFT
Magnitude spectrum
Peaks selection
Quadratic interpolation
$T_{\mathrm{cand}}[p^{(l)}]$
Candidates creation (3 signal frames)
sub. 5.3
Phase-based frequency jumps estimation
Magnitude-based frequency jumps estimation
$\Delta f_{\Phi}(T_{\mathrm{cand}}[p^{(l)}])$
$\Delta f_{\mathrm{M}}(T_{\mathrm{cand}}[p^{(l)}])$
Selection of valid candidates and tonality calculation
Examination of tracks length and tonality assignment
subsection 5.5
$\mathrm{FTM}[k_N, l]$

Fig. 5.1 Block diagram of the proposed algorithm for tonality estimation of spectral bins [97]

5.2 MAGNITUDE SPECTRUM ANALYSIS

Firstly, the input signal is divided into short segments, further called frames, in conformity to the STFT analysis concept. Every frame is weighted using the von Hann window [173]:

$$x_{\mathrm{w}}[n,l]=x[n+lL]w[n] \tag{5.1}$$

where: n=0, 1, ..., N–1, l denotes frame number, L is hop size of analysis, $x[n]$ is analyzed signal, and $w[n]$ is von Hann window of length N. Secondly, every frame of the signal is optionally zero-padded to its double length before applying the FFT procedure [1]:

$$x_{\mathrm{w,z}}[n,l]=\begin{cases} x_{\mathrm{w}}[n,l], & N-1\geq n\geq 0 \\ 0, & Z_{\mathrm{p}}N-1\geq n>N-1 \end{cases} \tag{5.2}$$

where Z_{p} is a zero-padding factor (expressed as a FFT length to the frame length ratio) that equals 1 (no zero-padding) or 2 in presented implementation.

The motivation for zero-padding of the signal frame before FFT calculation is the reduction of estimator bias resulting in an improved accuracy of frequency estimation. Basing on experimental results presented in literature, the maximum frequency bias of the QIFFT assuming the von Hann window is up-bounded in the following way [1]

$$f_{\mathrm{Mbias}}\leq\frac{F_s}{N}\left(\frac{1}{4Z_p}\right)^3 \tag{5.3}$$

where F_{s} is sampling rate as before. For zero-padding factor equal to 2 and frame length equivalent to approximately 32 ms (for instance: F_{s}=32 kSa/s, N=1024) the bias of considered frequency estimator calculated according to (5.3) is less than 0.07 Hz. Using the zero-padding factor higher than 2 seems to be impractical as it would result in a significant increase of the computational complexity, assuring only slight increase of the frequency estimation accuracy. Thus, in the investigated method for tonality measuring every frame of the input signal is optionally zero-padded to its doubled length.

Next, the short time Fourier spectrum is calculated [128]:

$$X[k,l] = \sum_{n=0}^{Z_\mathrm{p}N-1} x_{\mathrm{w,z}}[n,l] \cdot e^{-\frac{j2\pi kn}{Z_\mathrm{p}N}} \tag{5.4}$$

where k=0, 1, 2, …, $Z_\mathrm{p}N$–1 is a spectral bin number. Further on the local maxima of l-th magnitude spectra are detected to form a set:

$$K_{\max}^{(l)} = \{k : |X[k,l]| > |X[k-1,l]| \quad \wedge \quad |X[k,l]| > |X[k+1,l]|\} \tag{5.5}$$

The spectral index corresponding to the i-th local maximum in l-th frame is denoted as

$$k_{\max}\left[i^{(l)}\right] \in K_{\max}^{(l)}, \quad i^{(l)} = 0,\ 1,\ \ldots,\ \overline{\overline{K_{\max}^{(l)}}} - 1 \tag{5.6}$$

where $\overline{\overline{K_{\max}^{(l)}}}$ stands for total number of spectral local maxima detected within l-th frame. In order to eliminate spectral maxima of low energy that tend to have noise-like characteristic, the parameter expressing their peakiness is calculated according to the formula

$$g\left[k_{\max}\left[i^{(l)}\right]\right] = A\left[k_{\max}\left[i^{(l)}\right], l\right] - \frac{A\left[k_{\min-}\left(i^{(l)}\right), l\right] + A\left[k_{\min+}\left(i^{(l)}\right), l\right]}{2} \tag{5.7}$$

where $k_{\min-}\left[i^{(l)}\right]$ and $k_{\min+}\left[i^{(l)}\right]$ are the indices of the closest spectra minima on the both sides of the peak and $A\left[k_{\max}\left(j^{(l)}\right)\right] = 20\log\left(\left|X\left[k_{\max}\left(j^{(l)}\right), l\right]\right|\right)$ [150].

All peaks having $g\left[k_{\max}\left[i^{(l)}\right]\right]$ below the threshold g_thd are assumed to be noisy and are excluded from further processing. In order to select the appropriate g_thd value the normalized histogram of $g\left[k_{\max}\left[i^{(l)}\right]\right]$ values for the realization of the white Gaussian noise sampled at F_s=44100 Hz was analyzed using the frame length equivalent to the 46.4 ms (2048 samples). Considering normalized histogram presented in Fig. 5.2, the g_thd equal to 9 dB was selected as it allows rejecting about 30% of the noisy maxima. Unfortunately, the tonal components of low power will be also rejected with the assumed threshold. As the tonality of such components is low, treating them as noise-like ones has rather insignificant influence on the operation of the psychoacoustic model.

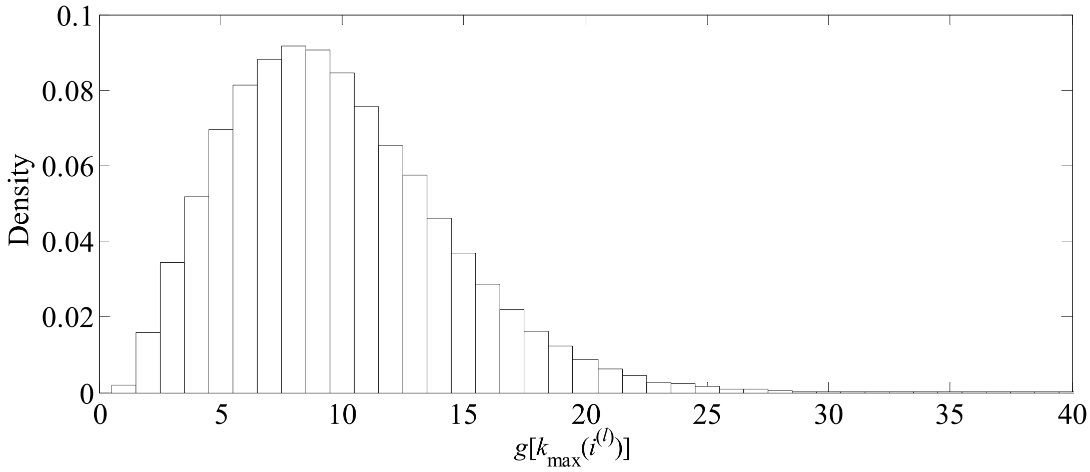


Fig. 5.2 Normalized histogram of the spectral peak heights for the 46 ms frames of sampled white Gaussian noise

The indices of the spectral maxima satisfying the criterion defined below are further considered as candidates for tonal components [109]:

$$K_{\text{s max}}^{(l)} = \left\{ k_{\max}\left[i^{(l)}\right] : g\left[k_{\max}\left[i^{(l)}\right]\right] > g_{\text{thd}} \right\} \tag{5.8}$$

where g_{thd}=9 dB and "s" is added to the subscript of set $K_{\max}^{(l)}$ in order to indicate that it stands for the selected maxima in l-th frame instead of all the maxima defined in Eq. (5.5). Consequently, the spectral indices corresponding to the j-th local maximum in l-th frame are denoted as

$$k_{\text{s max}}\left[j^{(l)}\right] \in K_{\text{s max}}^{(l)}, \quad j^{(l)} = 0,\ 1,\ \ldots,\ \overline{\overline{K_{\text{s max}}^{(l)}}} - 1 \tag{5.9}$$

Although the criterion defined in (5.8) is quite weak, it provides a substantial reduction of computational complexity of the algorithm stages described further. This is justified because the processing load strongly depends on the number of tonal components candidates. Moreover, the rejection of noisy peaks minimizes the probability of misclassifying noise-like components as tone-like ones.

In the next stage of processing, the quadratic interpolation is applied to all selected spectral maxima. Firstly, the fractional part – so called bin offset of spectral indices is determined [15][138].

$$k_{\mathrm{off}}\left[j^{(l)}\right]=\frac{1}{2}\frac{A\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right]-1\right]-A\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right]+1\right]}{A\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right]-1\right]-2A\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right]\right]+A\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right]+1\right]} \tag{5.10}$$

The instantaneous frequency of the spectrum peak detected in *l*-th frame of the signal is then given by

$$f_{\mathrm{M}}\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right]\right]=\frac{k_{\mathrm{s\,max}}\left[j^{(l)}\right]+k_{\mathrm{off}}\left[j^{(l)}\right]}{Z_{\mathrm{p}}N}F_{\mathrm{s}} \tag{5.11}$$

where *N* is the length of signal frame, and F_{s} is the sampling rate.

5.3 CREATION OF CANDIDATES FOR THREE-COMPONENT TONAL TRACKS

It is assumed that the maxima in three successive spectra may belong to the same tonal track if the following criteria are fulfilled

$$(1+d)f_{\mathrm{M}}\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right]\right]>f_{\mathrm{M}}\left[k_{\mathrm{s\,max}}\left[j^{(l-1)}\right]\right]>(1-d)f_{\mathrm{M}}\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right]\right] \tag{5.12}$$

$$(1+d)f_{\mathrm{M}}\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right]\right]>f_{\mathrm{M}}\left[k_{\mathrm{s\,max}}\left[j^{(l+1)}\right]\right]>(1-d)f_{\mathrm{M}}\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right]\right] \tag{5.13}$$

where $d\in(0,1)$ stands for the maximum relative detune of the tonal component between two successive frames of analysis. Although the *d* value can be expressed in relation to the maximal assumed pitch variation speed against absolute time, it was found experimentally that *d*=0.1 is suitable for the analysis of audio signals if the frame length corresponds to approximately 30 ms of the signal and the hop size is selected between ¼ and ¾ of the frame length.

If the following maxima frame triad $\left[k_{\mathrm{s\,max}}\left[j^{(l-1)}\right],\ k_{\mathrm{s\,max}}\left[j^{(l)}\right],\ k_{\mathrm{s\,max}}\left[j^{(l+1)}\right]\right]$ fulfills the criterion defined in (5.12) and (5.13), the candidate for the three-component tonal track, denoted further as a $T_{\mathrm{cand}}\left[p^{(l)}\right]$ is created, where $p^{(l)}=0,\ 1,\ \ldots,\ P^{(l)}-1$ is a candidate index and $P^{(l)}$ is the total number of candidates created in *l*-th frame of analysis. Additionally, if more than one component in the previous *l*–1 or following *l*+1 spectra meet the maximum detune criterion, the three-component tonal track candidate is created using the components of minimum frequency distance (Fig. 5.3).

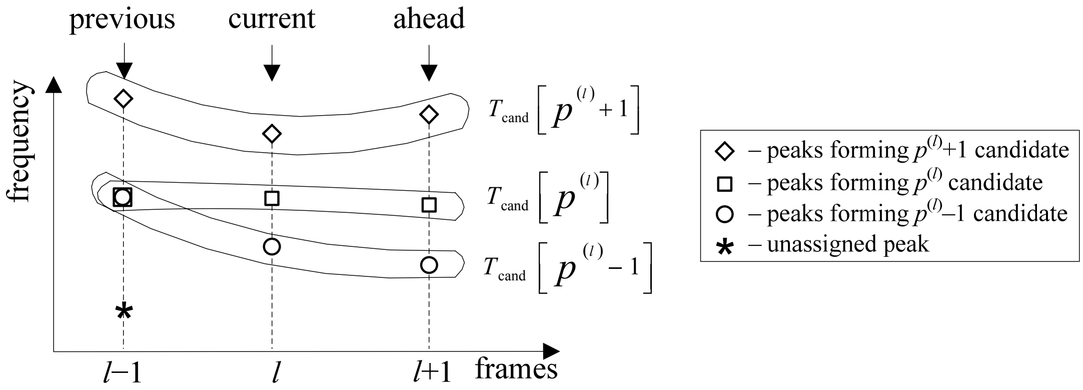

previous
current
ahead
frequency
$T_{\mathrm{cand}}\left[p^{(l)}+1\right]$
$T_{\mathrm{cand}}\left[p^{(l)}\right]$
$T_{\mathrm{cand}}\left[p^{(l)}-1\right]$
l–1
l
l+1
frames
◇ – peaks forming $p^{(l)}$+1 candidate
□ – peaks forming $p^{(l)}$ candidate
○ – peaks forming $p^{(l)}$–1 candidate
* – unassigned peak

Fig. 5.3 Creation of tonal track candidates

5.4 FREQUENCY JUMP ESTIMATION

Let $T_{\mathrm{cand}}\left[p^{(l)}\right]$ be a candidate for a three-component tonal track comprised of the $\left[k_{\mathrm{s\,max}}\left[j^{(l-1)}\right], k_{\mathrm{s\,max}}\left[j^{(l)}\right], k_{\mathrm{s\,max}}\left[j^{(l+1)}\right]\right]$ set of spectral maxima indices. The frequency jump measured halfway between centers of frames l–1 and l to the halfway between centers of frames l and l+1 corresponding to the $T_{\mathrm{cand}}\left[p^{(l)}\right]$ is estimated twice using the magnitude-based and phased-based methods resulting in the $\Delta f_{\mathrm{M}}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$ and $\Delta f_{\Phi}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$ estimates, respectively (Fig. 5.4). The procedures for frequency jumps estimation are described in the two following subsections.

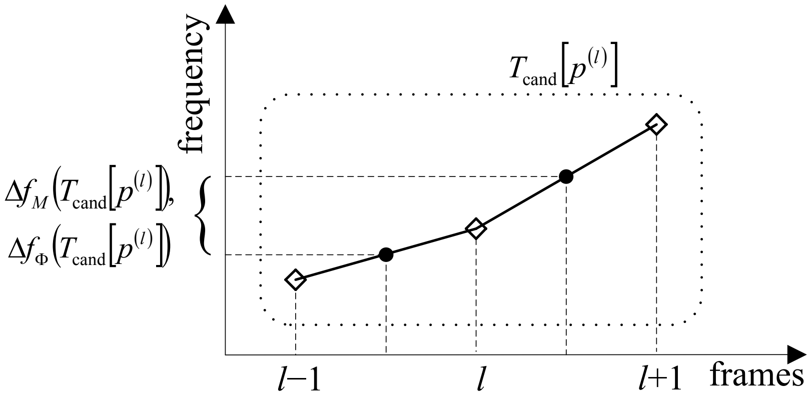

frequency
$T_{\mathrm{cand}}\left[p^{(l)}\right]$
$\Delta f_{M}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right),$
$\Delta f_{\Phi}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$
l–1
l
l+1
frames

Fig. 5.4 Frequency jumps corresponding to the tonal track candidate

5.4.1 Magnitude-based method

Basing on the magnitude-based frequency estimator defined in (5.11) and assuming linear interpolation of frequencies in halfway of frame centers (Fig. 5.3) the frequency jump related to the $T_{\text{cand}}\left[p^{(l)}\right]$ is given by:

$$\begin{gathered}
\Delta f_{\mathrm{M}}\left(T_{\text{cand}}\left[p^{(l)}\right]\right)=\frac{f_{\mathrm{M}}\left[k_{\mathrm{s\,max}}\left[j^{(l+1)}\right]\right]-f_{\mathrm{M}}\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right]\right]}{2}+f_{\mathrm{M}}\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right]\right]- \\
\frac{f_{\mathrm{M}}\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right]\right]-f_{\mathrm{M}}\left[k_{\mathrm{s\,max}}\left[j^{(l-1)}\right]\right]}{2}-f_{\mathrm{M}}\left[k_{\mathrm{s\,max}}\left[j^{(l-1)}\right]\right]= \\
\frac{f_{\mathrm{M}}\left[k_{\mathrm{s\,max}}\left[j^{(l+1)}\right]\right]-f_{\mathrm{M}}\left[k_{\mathrm{s\,max}}\left[j^{(l-1)}\right]\right]}{2}
\end{gathered} \tag{5.14}$$

5.4.2 Phase-based method

The method for frequency jumps estimation presented below is based on the well-known technique of instantaneous frequency estimation employing phase vocoder [25][27][51][106][136]. Various phase-based frequency estimators were evaluated and compared in [104]. Also the attempt was made to combine phase vocoder with a generalized sinusoidal model, which includes phase and amplitude modulations [14]. Usually, the hop size of STFT analysis equal to one sample is used in these methods in order to allow the estimation of instantaneous frequency in the full Nyquist band [104]. In this case, the above mentioned estimators yield adequate instantaneous frequency values (consistent with the magnitude-based estimates) even if the maximum of the analyzed spectrum corresponds to a noise-like component. This property makes these estimators useless in the application presented in this article, because it is assumed here that frequency jumps derived from the phase and magnitude spectra analysis are inconsistent to each other when the spectral bins belonging to the particular tonal track candidate are indeed noise-like. The algorithm for tonality estimation presented in this dissertation intends to operate when the hop size of the analysis ranges from approximately ¼ to ¾ of the frame length. Regarding the defined hop size range, the maximal instantaneous frequency that can be estimated employing long term phase vocoder concept is limited due to the phase indetermination problem [15][41]. In order to overcome this difficulty, a special dedicated procedure is applied to the phase angles used here to calculate the frequency jumps instead of instantaneous frequencies. Furthermore, the classical phase-based estimators presume that the quasi-sinusoidal component does not alter its frequency within two contiguous steps of analysis and thus the phase values used for frequency estimation come from spectra bins of identical indices. Contrary to these methods, the frequency jumps estimator presented below is able to operate correctly even if the bin indices corresponding to tonal track candidate are different to each other. Additionally, it yields frequency jumps inconsistent with the estimates provided by the magnitude-based estimator defined in (5.14) when the candidate for tonal track comprises of at least one noise-like spectra bin.

All three phase values corresponding to the spectra maxima belonging to the $T_{\text{cand}}\left[p^{(l)}\right]$ are involved into the procedure for frequency jump estimation (Fig. 5.5). First, the 2nd order phase difference corresponding to $T_{\text{cand}}\left[p^{(l)}\right]$ is calculated according to the following formula:

$$\begin{aligned}
&\Delta^2\Phi\left(k_{\text{s max}}\left[j^{(l+1)}\right], k_{\text{s max}}\left[j^{(l-1)}\right]\right) = \\
&\Delta\Phi\left(k_{\text{s max}}\left[j^{(l+1)}\right], k_{\text{s max}}\left[j^{(l)}\right]\right) - \Delta\Phi\left(k_{\text{s max}}\left[j^{(l)}\right], k_{\text{s max}}\left[j^{(l-1)}\right]\right) = \\
&\Phi\left[k_{\text{s max}}\left[j^{(l+1)}\right]\right] - \Phi\left[k_{\text{s max}}\left[j^{(l)}\right]\right] - \left(\Phi\left[k_{\text{s max}}\left[j^{(l)}\right]\right] - \Phi\left[k_{\text{s max}}\left[j^{(l-1)}\right]\right]\right) = \\
&\Phi\left[k_{\text{s max}}\left[j^{(l-1)}\right]\right] - 2\Phi\left[k_{\text{s max}}\left[j^{(l)}\right]\right] + \Phi\left[k_{\text{s max}}\left[j^{(l+1)}\right]\right]
\end{aligned} \tag{5.15}$$

where $\Phi\left[k_{\text{s max}}\left(j^{(l)}\right)\right] = \arctan\left(\dfrac{\operatorname{Im}\left(X\left[k_{\text{s max}}\left(j^{(l)}\right), l\right]\right)}{\operatorname{Re}\left(X\left[k_{\text{s max}}\left(j^{(l)}\right), l\right]\right)}\right)$.

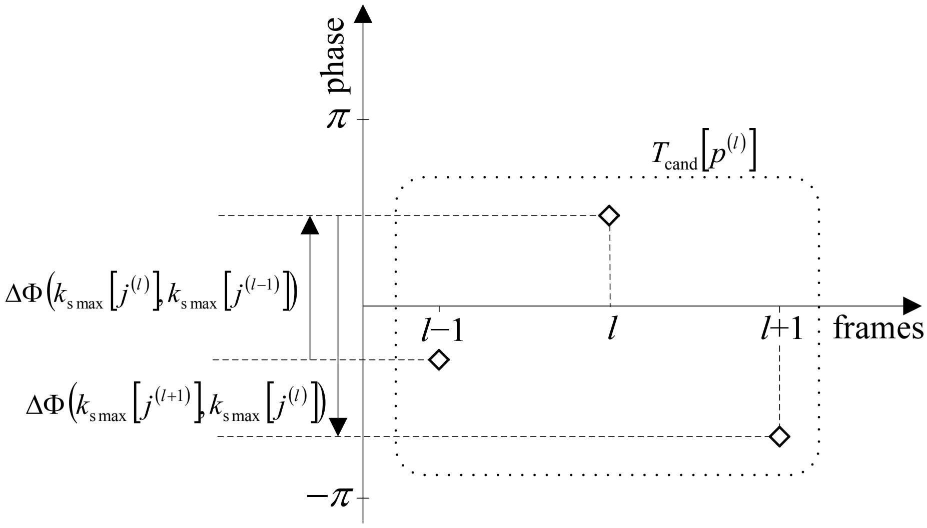

phase
π
$-\pi$
$T_{\text{cand}}\left[p^{(l)}\right]$
$\Delta\Phi\left(k_{\text{s max}}\left[j^{(l)}\right], k_{\text{s max}}\left[j^{(l-1)}\right]\right)$
$\Delta\Phi\left(k_{\text{s max}}\left[j^{(l+1)}\right], k_{\text{s max}}\left[j^{(l)}\right]\right)$
$l-1$
l
$l+1$
frames

Fig. 5.5 Phase jumps corresponding to the tonal track candidate

If the quasi-sinusoidal component changes its frequency and is represented by spectral bins of various indices in the contiguous spectra, the following phase offset must be involved into the frequency jump estimation procedure:

$$\begin{aligned}
&\Delta^2\phi\left(k_{\mathrm{s\,max}}\left[j^{(l+1)}\right], k_{\mathrm{s\,max}}\left[j^{(l-1)}\right]\right) = \\
&\frac{\pi(N-1)}{Z_{\mathrm{p}}N}\left(k_{\mathrm{s\,max}}\left[j^{(l+1)}\right] - k_{\mathrm{s\,max}}\left[j^{(l)}\right] + k_{\mathrm{s\,max}}\left[j^{(l-1)}\right] - k_{\mathrm{s\,max}}\left[j^{(l)}\right]\right) = \\
&\frac{\pi(N-1)}{Z_{\mathrm{p}}N}\left(k_{\mathrm{s\,max}}\left[j^{(l-1)}\right] - 2k_{\mathrm{s\,max}}\left[j^{(l)}\right] + k_{\mathrm{s\,max}}\left[j^{(l+1)}\right]\right)
\end{aligned} \tag{5.16}$$

If the $k_{\mathrm{s\,max}}\left[j^{(l-1)}\right]$, $k_{\mathrm{s\,max}}\left[j^{(l)}\right]$ and $k_{\mathrm{s\,max}}\left[j^{(l+1)}\right]$ indices remain in the linear relation the phase offset defined above is nullified. The frequency jump related to the $T_{\mathrm{cand}}\left[p^{(l)}\right]$ is given by:

$$\begin{aligned}
&\Delta f_{\Phi}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right) = \\
&\frac{F_s}{2\pi L}\left(\operatorname{princarg}\left(\Delta^2\Phi\left(k_{\mathrm{s\,max}}\left[j^{(l+1)}\right], k_{\mathrm{s\,max}}\left[j^{(l-1)}\right]\right)\right) + \Delta^2\phi\left(k_{\mathrm{s\,max}}\left[j^{(l+1)}\right], k_{\mathrm{s\,max}}\left[j^{(l-1)}\right]\right)\right)
\end{aligned} \tag{5.17}$$

where $\operatorname{princarg}(\varphi) = (\varphi + \pi)\operatorname{mod}(-2\pi) + \pi$ is the principal argument function mapping the input phase φ into the (–π, π) range [41][173].

However, when tonal components of fast varying frequency are analyzed, the frequency jump derived from phase processing (5.16) may not be adequate due to the phase ambiguity problem [15][41]. In order to overcome this difficulty, the minimal and maximal frequency jumps are calculated basing on the numbers of spectra bins constituting a particular candidate for tonal track [98]:

$$\begin{aligned}
&\Delta f_1\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right) = \\
&\frac{F_s}{Z_{\mathrm{p}}N}\left(\frac{k_{\mathrm{s\,max}}\left[j^{(l+1)}\right] - k_{\mathrm{s\,max}}\left[j^{(l-1)}\right]}{2} - \operatorname{msgn}\left(k_{\mathrm{s\,max}}\left[j^{(l+1)}\right] - k_{\mathrm{s\,max}}\left[j^{(l-1)}\right]\right)\right)
\end{aligned} \tag{5.18}$$

$$\begin{aligned}
&\Delta f_2\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right) = \\
&\frac{F_s}{Z_{\mathrm{p}}N}\left(\frac{k_{\mathrm{s\,max}}\left[j^{(l+1)}\right] - k_{\mathrm{s\,max}}\left[j^{(l-1)}\right]}{2} + \operatorname{msgn}\left(k_{\mathrm{s\,max}}\left[j^{(l+1)}\right] - k_{\mathrm{s\,max}}\left[j^{(l-1)}\right]\right)\right)
\end{aligned} \tag{5.19}$$

$$\Delta f_{\min}\left(T_{\text{cand}}\left[p^{(l)}\right]\right) = \min\left(\Delta f_1\left(T_{\text{cand}}\left[p^{(l)}\right]\right), \Delta f_2\left(T_{\text{cand}}\left[p^{(l)}\right]\right)\right) \qquad (5.20)$$

$$\Delta f_{\max}\left(T_{\text{cand}}\left[p^{(l)}\right]\right) = \max\left(\Delta f_1\left(T_{\text{cand}}\left[p^{(l)}\right]\right), \Delta f_2\left(T_{\text{cand}}\left[p^{(l)}\right]\right)\right) \qquad (5.21)$$

where msgn( ) is modified *signum* function which returns a value equal to 1 for non-negative arguments (higher or equal to 0 value) and a value equal to –1 for negative arguments. Next it is checked whether for any multiplier

$$m_j = \begin{cases} 1,\ 2,\ ...,\ 6; & \Delta f_{\Phi}\left(T_{\text{cand}}\left[p^{(l)}\right]\right) < \Delta f_{\min}\left(T_{\text{cand}}\left[p^{(l)}\right]\right) \\ -1,\ -2,\ ...,\ -6; & \Delta f_{\Phi}\left(T_{\text{cand}}\left[p^{(l)}\right]\right) > \Delta f_{\max}\left(T_{\text{cand}}\left[p^{(l)}\right]\right) \end{cases} \qquad (5.22)$$

the following criterion is fulfilled:

$$\Delta f_{\max}\left(T_{\text{cand}}\left[p^{(l)}\right]\right) > \Delta f_{\Phi}\left(T_{\text{cand}}\left[p^{(l)}\right]\right) + m_j \frac{F_{\text{s}}}{L} > \Delta f_{\min}\left(T_{\text{cand}}\left[p^{(l)}\right]\right) \qquad (5.23)$$

If multiplier m_j exists the frequency jump derived from the phase spectrum processing is updated as given by

$$\Delta f_{\Phi}\left(T_{\text{cand}}\left[p^{(l)}\right]\right) \leftarrow \Delta f_{\Phi}\left(T_{\text{cand}}\left[p^{(l)}\right]\right) + m_j \frac{F_{\text{s}}}{L} \qquad (5.24)$$

Otherwise m_j=0 is assumed. The range for m_j was determined experimentally and is constant here regardless of the instantaneous frequencies of spectral peaks constituting the candidate for tonal track. However, a slight improvement (lower false positive rate of tonal components detection) may be obtained by making the m_j multiplier range narrower for lower frequencies and wider for higher frequencies.

5.5 MEASURING SPECTRAL COMPONENT TONALITY

5.5.1 Measuring tonal track candidate tonality

It can be assumed that a particular number of created candidates for three-component tonal tracks represent the sets of noise-like components. In order to eliminate these candidates and then select only truly tonal tracks the difference between frequency jumps estimated using (5.14) and (5.17) or (5.24) is determined for every candidate:

$$\delta\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)=\Delta f_{\mathrm{M}}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)-\Delta f_{\Phi}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right) \tag{5.25}$$

Among all candidates only these fulfilling criterion defined below are selected;

$$\mathbf{T}_{\mathrm{trk}}^{(l)}=\left\{T_{\mathrm{cand}}\left[p^{(l)}\right]:\left|\delta\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)\right|<\frac{F_{\mathrm{s}}}{NZ_{\mathrm{p}}}\right\} \tag{5.26}$$

The valid r-th three-component tonal track created within l-th step of analysis is then denoted as

$$T_{\mathrm{trk}}\left[r^{(l)}\right]\in\mathbf{T}_{\mathrm{trk}}^{(l)}, r^{(l)}=0,\ 1,\ \ldots,\ \overline{\overline{\mathbf{T}_{\mathrm{trk}}^{(l)}}} \tag{5.27}$$

and the frequency-derived tonality measure (FTM) is assigned to it according to the formula:

$$\mathrm{FTM}_{\mathrm{trk}}\left(T_{\mathrm{trk}}\left[r^{(l)}\right]\right)=1-\frac{NZ_{\mathrm{p}}}{F_{\mathrm{s}}}\left|\delta\left(T_{\mathrm{trk}}\left[r^{(l)}\right]\right)\right| \tag{5.28}$$

5.5.2 *Assigning tonality to spectral bins*

In order to increase the reliability of decision regarding spectra bins tonality (the incorrect classification of noise-like components as tonal ones), it seems reasonable to analyze whether the components of the valid tonal tracks are the elements of a longer tonal track appearing in four or more contiguous spectra. Thus, the tonality measure $\mathrm{FTM}_{\mathrm{trk}}\left(T_{\mathrm{trk}}\left[r^{(l)}\right]\right)$ is assigned to the $k_{\mathrm{smax}}\left[j^{(l)}\right]$ spectral bin resulting in $\mathrm{FTM}\left[k_{\mathrm{smax}}\left[j^{(l)}\right],l\right]$, if $T_{\mathrm{trk}}\left[r^{(l)}\right]$ is at least L_1-th three component tonal track belonging to the same long tonal track. The L_1 parameter should be reasonably selected as its affects the detection delay (L_1=1 is used in all experiments).

The three-component tonal track $T_{\mathrm{trk}}\left[r^{(l)}\right]$ composed of the maxima triad $\left[k_{\mathrm{s\,max}}\left[j^{(l-1)}\right], k_{\mathrm{s\,max}}\left[j^{(l)}\right], k_{\mathrm{s\,max}}\left[j^{(l+1)}\right]\right]$ which is not continued with any $T_{\mathrm{trk}}\left[r^{(l+1)}\right]$ is assumed to be dying. Furthermore, if the $T_{\mathrm{trk}}\left[r^{(l)}\right]$ is the last valid three-component track being a part of the longer track consisting of at least L_2 three component tonal tracks, the tonality measure $\mathrm{FTM}_{\mathrm{trk}}\left(T_{\mathrm{trk}}\left[r^{(l)}\right]\right)$ is assigned to $k_{\mathrm{smax}}\left[j^{(l+1)}\right]$. The

$L_2 = \left\lfloor \dfrac{N}{L} + 0.5 \right\rfloor + 1$ is the experimentally determined parameter and obviously $L_2 > L_1$.

The symbol $\lfloor \ \rfloor$ stands for rounding to the nearest integer towards minus infinity. The procedure described above is depicted in Fig. 5.6.

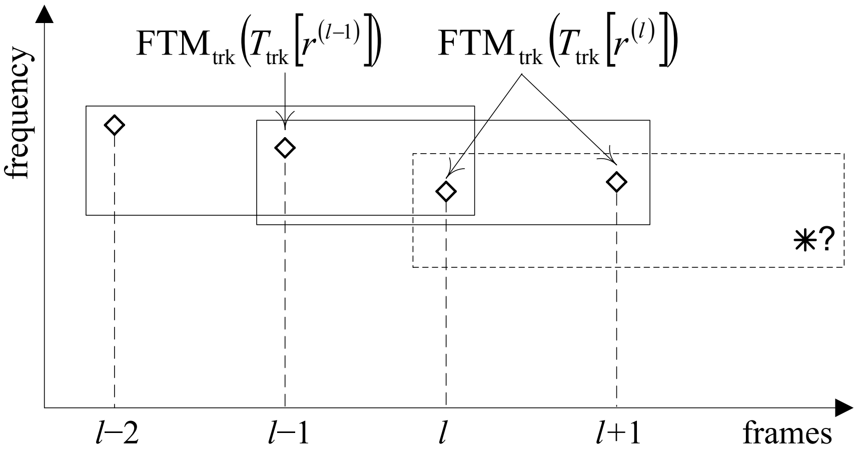

frequency
$\mathrm{FTM}_{\mathrm{trk}}\left(T_{\mathrm{trk}}\left[r^{(l-1)}\right]\right)$
$\mathrm{FTM}_{\mathrm{trk}}\left(T_{\mathrm{trk}}\left[r^{(l)}\right]\right)$
*?
$l-2$
$l-1$
l
$l+1$
frames

Fig. 5.6 Assignment of tonality measures to spectral bins (solid rectangle – valid tracks, dashed rectangle – invalid or not existing track)

It can be seen from Fig. 5.6 that the tonality measure $\mathrm{FTM}_{\mathrm{trk}}\left(T_{\mathrm{trk}}\left[r^{(l)}\right]\right)$ is assigned to the peak detected within the (l+1)-th spectrum, even if it is not a part of a valid three-component tonal track or if this track does not exist in the (l+1)-th step of the analysis.

Regarding the application requirements the described algorithm may return tonality measures corresponding to the interpolated spectrum obtained after applying the FFT procedure to the zero-padded frames of signal, or the tonality measures may be assigned to the bins corresponding to the spectrum obtained without zero-padding. Considering the latter case, the tonality measures yielded by the described procedure are assigned to the bins in the l-th step of analysis of numbers derived from

$$k_{\mathrm{s\,max}\,N}\left[j^{(l)}\right] = \left\lfloor f_{\mathrm{M}}\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right]\right]\frac{N}{f_s} + 0.5 \right\rfloor \tag{5.29}$$

In all of the experiments described in the next section the tonality measures are allocated in accordance with the above formula as it makes easier to compare the efficiency of the engineered algorithm with other tonality estimation methods.

All spectral components not detected as spectral maxima or not assigned any tonality measure by the described procedure are treated as a totally noise-like ones, and the tonality measure equal to 0 is associated with them ($\mathrm{FTM}[k,l]=0$ or $\mathrm{FTM}[k_N,l]=0$ where $k_N=0,\ 1,\ 2,\ \ldots,\ N-1$). The inverted tonality measure $1-\mathrm{FTM}_{\mathrm{trk}}\left(T_{\mathrm{trk}}\left[r^{(l)}\right]\right)$ assigned to the spectral maxima $k_{\max}\left[i^{(l)}\right]$ detected within l-th (current) frame is denoted further as ftm[$k_{\max}$].

6 EVALUATION OF ALGORITHM PERFORMANCE

The developed tonality estimator was compared with the selected existing methods which are mainly used in psychoacoustic models for the estimation of the basilar membrane stimuli characteristics. However, two of these algorithms express the tonality in a different way than the FTM method which makes a direct comparison impossible. Thus the additional assumptions were made for them in order to allow comparison with the FTM method. The algorithms used for comparison together with a short description and assumptions made are as follows:

- SFMs – were calculated for critical bands with frequency boundaries defined by Zwicker [174]. Because the SFMs express the tonality of the entire frequency band, and they get 0 for tonal subbands and 1 for noisy ones, it is assumed that 1–SFM are assigned only to those spectra maxima that have the power exceeding the arithmetic mean of the power spectrum subband. All remaining components are assumed to have zero tonality.
- UM – similarly to SFMs, here the 1–UM values instead of UM are used. The 1–UM values are assigned only to spectral bins corresponding to local spectra maxima. All remaining components are assumed to have the tonality of zero.
- Tonal components detector employed in MPEG psychoacoustic model 1 (further denoted as a M1). This is a simple discrete classifier assigning tonality equal to 1 to spectra maxima which tend to be tonal and 0 to all remaining bins [67]. In this case no additional conditioning was necessary.

The defining equations for the above-mentioned tonality estimators were given in Section 4.

6.1 EFFICIENCY OF TONAL COMPONENT DETECTION

The proposed algorithm may be viewed as a probabilistic (scoring) classifier, and thus its performance was evaluated using the technique of ROC (Receiver Operating Characteristic) graphs [46]. In order to calculate the ROC curves, a set of signals was generated according to the following formula:

$$\begin{aligned} x[n] = {} & A\sum_{h=1}^{20}\sin\left(\frac{2\pi f_{\mathrm{p}}hn}{F_{\mathrm{s}}}+\frac{d_{\mathrm{FM}}f_{\mathrm{p}}h}{f_{\mathrm{FM}}}\sin\left(\frac{2\pi f_{\mathrm{FM}}n}{F_{\mathrm{s}}}\right)+\varphi_0(h)\right)+ \\ & A\sum_{h=1}^{5}\sin\left(\frac{2\pi(20f_{\mathrm{p}}+1000h)n}{F_{\mathrm{s}}}+\frac{d_{\mathrm{FM}}(20f_{\mathrm{p}}+1000h)}{f_{\mathrm{FM}}}\sin\left(\frac{2\pi f_{\mathrm{FM}}n}{F_{\mathrm{s}}}\right)+\varphi_0(h)\right) \end{aligned} \tag{6.1}$$

where A is a sinusoid amplitude, $f_{\mathrm{p}} = 110 \times 2^{p/12}$ denotes the fundamental frequency related to pitch, n=0, 1, … N_{sig}–1, N_{sig}=64000, F_{s}=32 kHz, φ_0 is a randomly generated initial phase which is different for every sinusoid, and the parameters are p=0, 1, 2, …, 24, f_{FM} denotes modulation rate and d_{FM} stands for frequency modulation depth expressed in the percentage of the carrier frequency. Assuming a particular combination of p, f_{FM}, d_{FM} parameters, the $x[n]$ consists of 20 harmonics and 5 sinusoids occupying a higher frequency band spaced at intervals of 1 kHz. Since the analyzed signals were a mixture of $x[n]$ and white Gaussian noise, the power of white noise and the amplitude of the sinusoidal components were adjusted in order to achieve a desired SNR;

$$\mathrm{SNR}[\mathrm{dB}] = 10\log_{10}\frac{\sum_{n=0}^{N_{\mathrm{sig}}-1}x^2[n]}{\sum_{n=0}^{N_{\mathrm{sig}}-1}y^2[n]} \tag{6.2}$$

where $y[n]$ stands for white noise. For a particular SNR and p, f_{FM}, d_{FM} parameters the total number of 25 signals were generated (for p=1, 2, …, 24 the f_{p} ranges from 110 Hz to 440 Hz with one semitone step), and then further analyzed using the frame length equal to 32 ms (1024 samples), and the hop size adjusted to a quarter of the frame length (256 samples). Because the frequencies of sinusoidal components were known a priori the vector containing values equal to one for all indices related to sinusoids frequencies and zeros for all remaining spectra maxima was created. Further this vector was used to calculate the true positive rate (positives correctly classified divided by total positives) and false positive rate (negatives incorrectly classified divided by total negatives) of all examined classifiers. Assuming a particular combination of f_{p}, f_{FM}, d_{FM} and SNR, at least 3000 tonality estimation results were gathered for every examined algorithm. Consequently, for a group of 25 signals generated with particular SNR, f_{FM}, d_{FM}, and p=1, 2, …, 24 the vectors containing at least 3000×25 tonality estimates were used in order to generate ROC graphs employing the procedure described in [46]. The

ROC graphs for all investigated algorithms along with FTM are presented in Fig. 6.1 for SNR=10 dB and Fig. 6.2 for SNR= 0 dB. For each SNR, two sets of signals were analyzed – the first one without frequency modulation applied (f_{FM}=0, d_{FM}=0) and the second one with applied frequency modulation simulating a typical, instrumental vibrato effect (f_{FM}=6 Hz, d_{FM}=0.03 corresponding to ±0.5 semitone) [114][154].

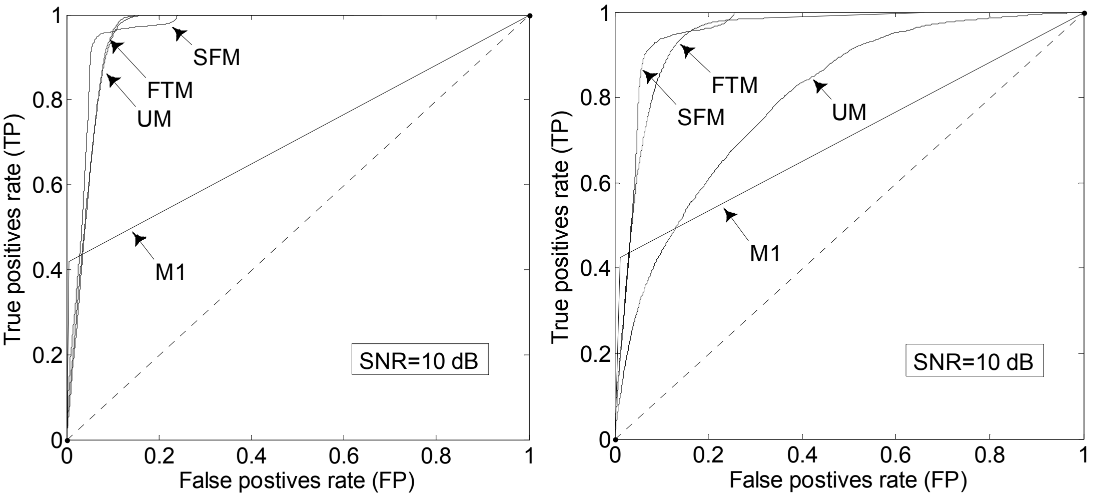


Fig. 6.1 ROC graphs for examined algorithms with SNR=10 dB; left – constant frequency sinusoids, right – frequency modulated sinusoids

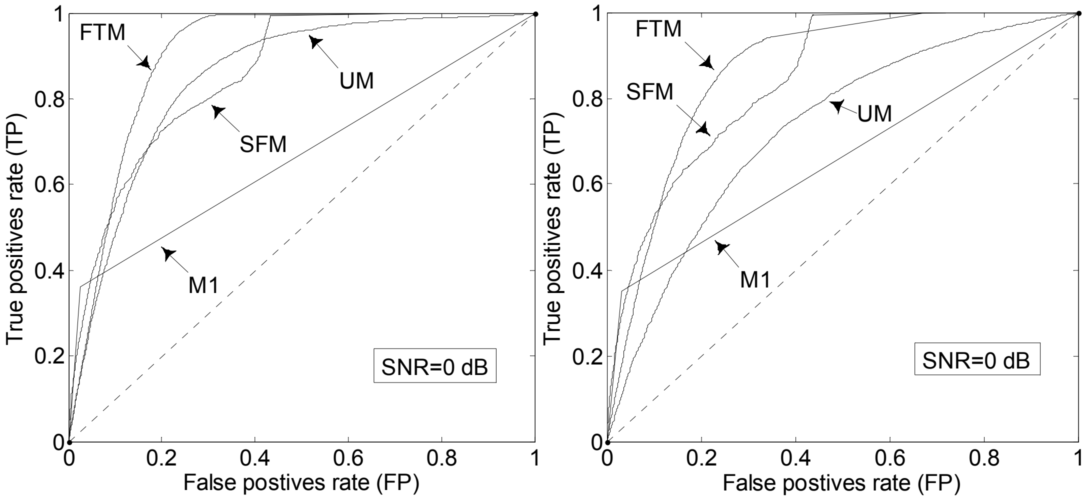


Fig. 6.2 ROC graphs for examined algorithms with SNR=0 dB; left – constant frequency sinusoids, right – frequency modulated sinusoids

It can be observed from Fig. 6.1 that FTM, UM and SFM operate similarly to each other and almost perfectly when detecting constant frequency tonal components and when the

SNR is equal to 10 dB. Considering the characteristics obtained for the UM method presented in Fig. 6.1 and 6.2, it is apparent that this method of tonal components detection is most sensitive to frequency modulation among others as its performance is significantly decreased in this case. When comparing the area contained under the curves (AUCs – Tab. 6.1) and the ROC graphs of the FTM and SFM methods it is noticeable that their performance is similar. However, FTM seems to operate slightly better than SFM (modified for the purpose of this comparison) when SNR is equal to zero because it gets closer to (1, 1) point in the ROC graphs. The AUC for ROC graphs of FTM algorithm are either equal or higher than AUCs obtained for the UM and M1 which directly proves its overall robustness.

Tab. 6.1 Areas under ROCs for examined methods

SNR [dB]	FM	AUC(FTM)	AUC(UM)	AUC(SFM)	AUC(M1)
10	no	0.96	0.95	0.96	0.71
	yes	0.94	0.8	0.95	0.71
0	no	0.91	0.85	0.86	0.67
	yes	0.86	0.72	0.85	0.67

Although a higher rate of true to false positives can be achieved with FTM than with the M1 method, the threshold turning the scoring FTM classifier into discrete one should be reasonably selected when considering its use in the MPEG psychoacoustic model 1. The M1 method is the most conservative among all examined, because it provides true positive rate around 0.4 while the false positive rate does not exceed 0.05 even if the SNR is equal to 0 dB and the frequency modulation is applied to the sinusoids being analyzed. Thus, it seems that the detection threshold for FTM should be selected so that the low false positive rate of tonal components detection would be quite low. Assuming that the false positive rate should not exceed 0.1 level for the MPEG psychoacoustic model 1 the true positive rate provided by FTM would be significantly higher than the one provided by the M1 method. The experiments allowing the determination of the optimum threshold level for the FTM in order to optimize the reliability of the MPEG psychoacoustic model 1 are out of the dissertation scope.

6.2 TONALITY ESTIMATION

6.2.1 Stationary sinusoids

In order to find out the relation between tonality measures yielded by the examined algorithms and the SNR of a tonal component, a set of signals consisting of the constant frequency sinusoid and noise with SNR varying from –20 dB to 30 dB with 2 dB step were generated and analyzed. The sampling rate was 16000 Sa/s and the frequency of the sinusoidal component was adjusted to 440 Hz. The generated signals of particular SNRs with duration time 4s consisted of 64000 samples (4 s) and the initial phase of the sinusoidal component was randomly selected. The length of the analysis frame was set to 32 ms (512 samples), and the hop size was adjusted to a half of the frame length. As the number of spectral bins corresponding to the tonal component was known a priori, a vector representing tonality estimated within successive steps of analysis was created for every generated sample. The mean and minimal values as well as the standard deviation were calculated for all 26 vectors gathered using FTM, UM and SFM methods. The statistics obtained are presented in Fig. 6.3.

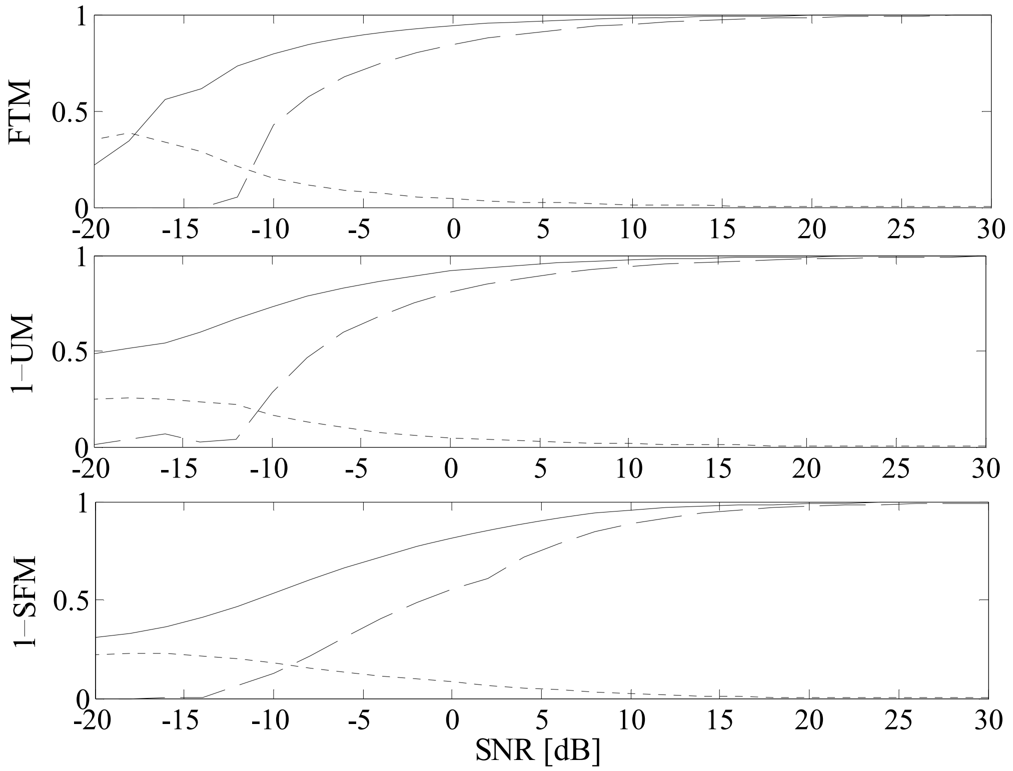


Fig. 6.3 The statistics obtained with the UM (top), FTM (middle) and SFM (bottom) methods – sample tonal component of constant frequency

Notwithstanding small differences in the FTM, UM and SFM characteristics, they seem to perform similarly when the tonal component of constant frequency is considered. The difference can be noticed for SNR falling below –10 dB. In this range the mean values of tonality obtained using SFM and UM descend more gradually comparing to FTM algorithm. However, this property seems to have negligible importance, as the difference of tonality in this SNR range is rather hardly perceived.

6.2.2 Effect of frequency modulation

Next, two groups of signals consisting of a single sinusoidal component of a nominal frequency equal to 120 Hz and 400 Hz (F_s=16000 Sa/s) modulated nonlinearly by the sinusoidal function of 6 Hz frequency were analyzed using the FTM, UM and SFM methods. The SNR of the sinusoid was equal to 20 dB. While the modulation rate was constant, the modulation depth altered from 0 (no modulation) to the depth corresponding to ±1 semitone with the step equal to 0.05 of semitone. The length of the analysis frame was set to 32 ms (512 samples), and the hop size was adjusted to a half of the frame length – identically to those used in experiment described in subsection 6.2.1. For every selected modulation depth the vectors containing tonality measures related to the tonal component, were gathered using three above-mentioned algorithms. In Fig. 6.4 a comparison of the mean values of tonality measures obtained using FTM, UM and SFM methods are presented for the carrier frequency of the modulated component equal to 120 Hz and 440 Hz, respectively.

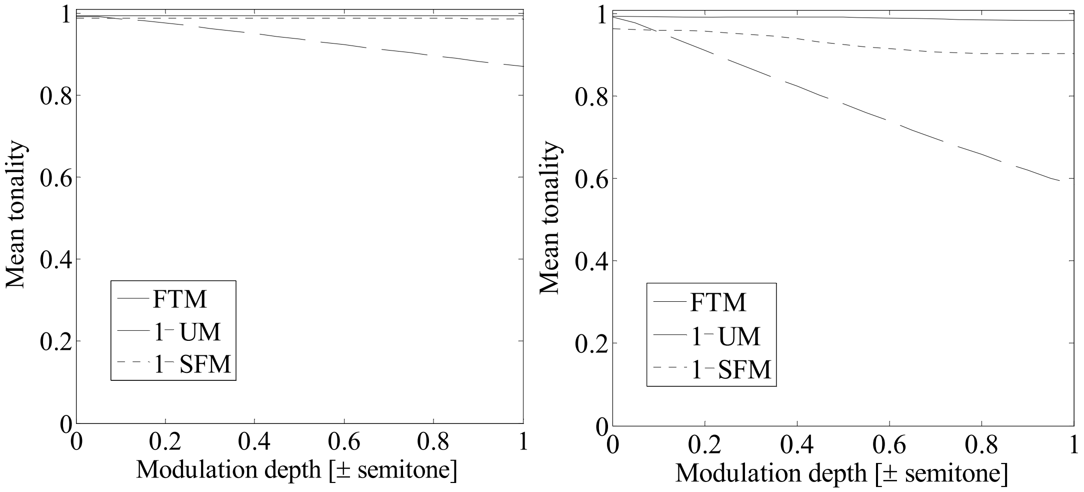


Fig. 6.4 Influence of frequency modulation depth on the mean tonality measurements with carrier frequency 120 Hz (left) and 440 Hz (right)

Considering the results presented in Fig. 6.4 for the 120 Hz carrier, it is apparent that the FTM and SFM methods are almost totally insensitive to the frequency modulation of ±1 semitone depth (approximately 113 to 127 Hz in this case). Although the tonality estimates provided by the UM algorithm in case of such a modulation depth are approximately 10 times smaller than for a constant frequency sinusoid, this degradation is acceptable. However, when comparing Fig. 6.4 (right) and Fig. 6.3 (middle) it can be noted that the tonality measure yielded by UM for the sinusoid of 20 dB SNR modulated within 415 Hz to 466 Hz range (440 Hz ±1 semitone) is equal to the tonality measure obtained for a constant frequency sinusoid of approximately –15 dB SNR. While the UM method yields inadequate tonality measures for deeply frequency modulated components occupying mid and high frequency bands, the FTM and SFM methods seem to be much more insensitive in this case. Consequently, using UM for the offset calculation accordingly to (3.16), would lead to the hearing threshold estimate corresponding to a stimuli similar to that obtained for noisy spectra components. Using the FTM instead of the UM in this case would lead to a more reliable estimate of the hearing threshold.

6.2.3 Effect of amplitude modulation

In order to evaluate an influence of the amplitude modulation (AM) on the tonality measures provided by the investigated algorithms, experiments similar to that described in previous subsection were carried out. In this case, a set consisting of 39 sinusoids of constant frequency equal to 440 Hz was generated and AM was applied to them. The AM depths were selected within the range from 0 to 0.95 with 0.025 steps in order to obtain the set of AM sinusoids modulated with various depths. The frequency of AM was constant for every sinusoid and equal to 5 Hz. The generated signals consisted of 64000 samples (4 s) and the initial phase of sinusoidal components was randomly selected. During the analysis, the frame length was adjusted to 32 ms (512 samples), and the hop size was equal to a half of it. The SNR of the generated sinusoids was selected to be equal to 40 dB in order to ensure that even when the sinusoid is deeply modulated its instantaneous SNR does not go below the particular level (tonality close to 1 should be assigned to it). The mean values of tonality yielded by the FTM, UM and SFM algorithms for various AM depths are presented in Fig. 6.5.

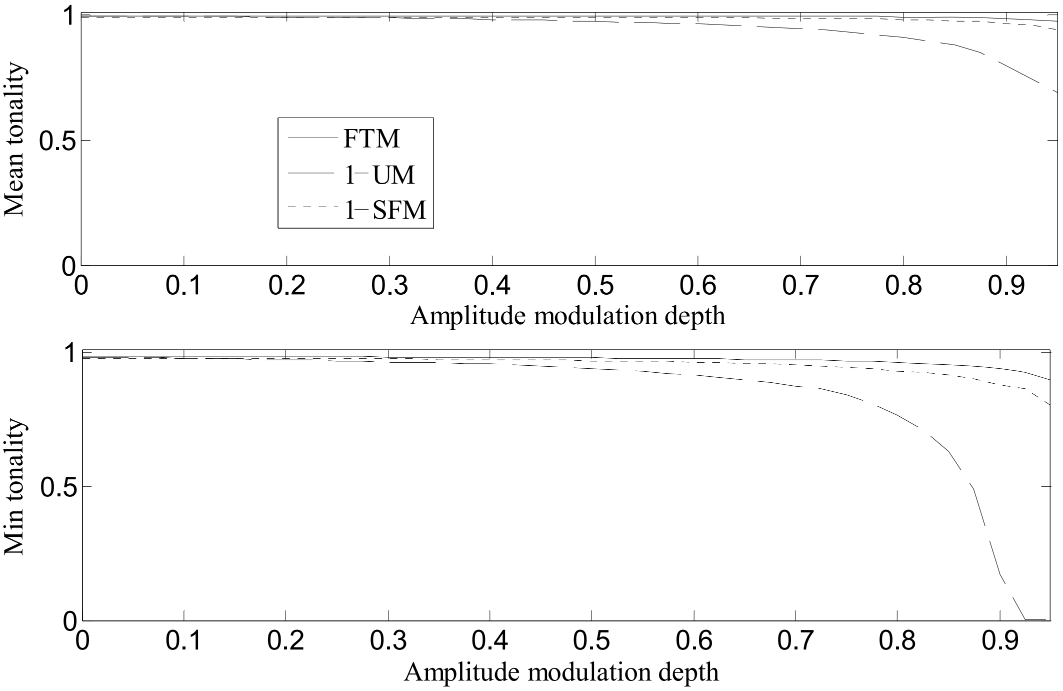


Fig. 6.5 Influence of an exemplary amplitude modulation of tonal components on the mean (upper part) and minimum (lower part) tonality estimates (UM, FTM, SFM)

The comparison of characteristics shown in Fig. 6.5 indicates that the amplitude modulation of tonal component affects the tonality measures obtained using the FTM and SFM algorithms just slightly. The robustness of the FTM and SFM methods in terms of insensitivity to AM is directly related to that they do not assume any particular inter-frame behavior of the magnitude of tonal components. Although the UM method is the most sensitive to AM among all examined algorithms, the influence of AM tends to be significant only for the modulation depths higher than approximately 0.8. Obviously, the cause of the UM limitations in this modulation depth range is a consequence of the assumption regarding the inter-frame linear changes of spectral bin magnitudes.

6.2.4 Experiments employing speech and music recordings

Contrarily to the experiments involving synthetic sound samples, in this subsection the experimental results for speech and music recordings are presented. The recording with male speech sampled at 32 kHz was analyzed using 32 ms (1024 samples) von Hann window and the hop size equal to the ¼ of the frame length (256 samples). In case of the recording containing a fragment of jazz music sampled with 44.1 kHz, the frame length was adjusted to 46.4 ms (2048 samples) and the hop size was set to the ¼ of the frame length (512 samples). In Fig. 6.6, the fragments of spectrograms of the recording containing speech and music are shown.

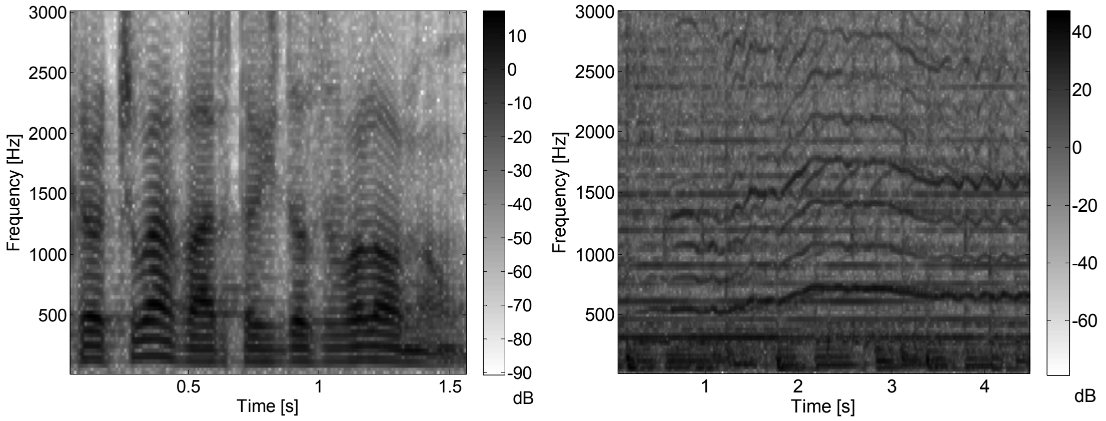


Fig. 6.6 Spectrograms of the analyzed speech (left part) and polyphonic (right part) recordings (up to 3000 kHz)

Instead of the term ‘spectrogram’ the ‘tonalogram’ was introduced as a three-dimensional intensity plot displaying the tonality of spectral components as a function of both time and frequency. In Figs. 6.7 and 6.8, the tonalograms obtained using the FTM, UM, SFM and M1 methods applied to speech and polyphonic recordings are shown. The tonality estimates provided by the UM and SFM algorithms were assigned only to the spectra maxima according to the statements made in the beginning of Section 5.

It can be observed from Figs. 6.7 and 6.8 that FTM yields high tonality measures for partials even if the instantaneous frequency of tonal components related to them varies significantly in time. This is not the case for the UM method, where the tonality close to 1 is only assigned to constant or slowly varying partials. Consequently, basing on the results presented in subsection 6.1 it can be expected that modulated partials or their fragments visible in the UM tonalogram would be discarded after applying a reasonable detection threshold (e.g. preserving false positives rate at the 0.5 level). Although the M1 method does not provide continuous tonality measures falling within the [0,1] range it is also included here for the purpose of the comparison.

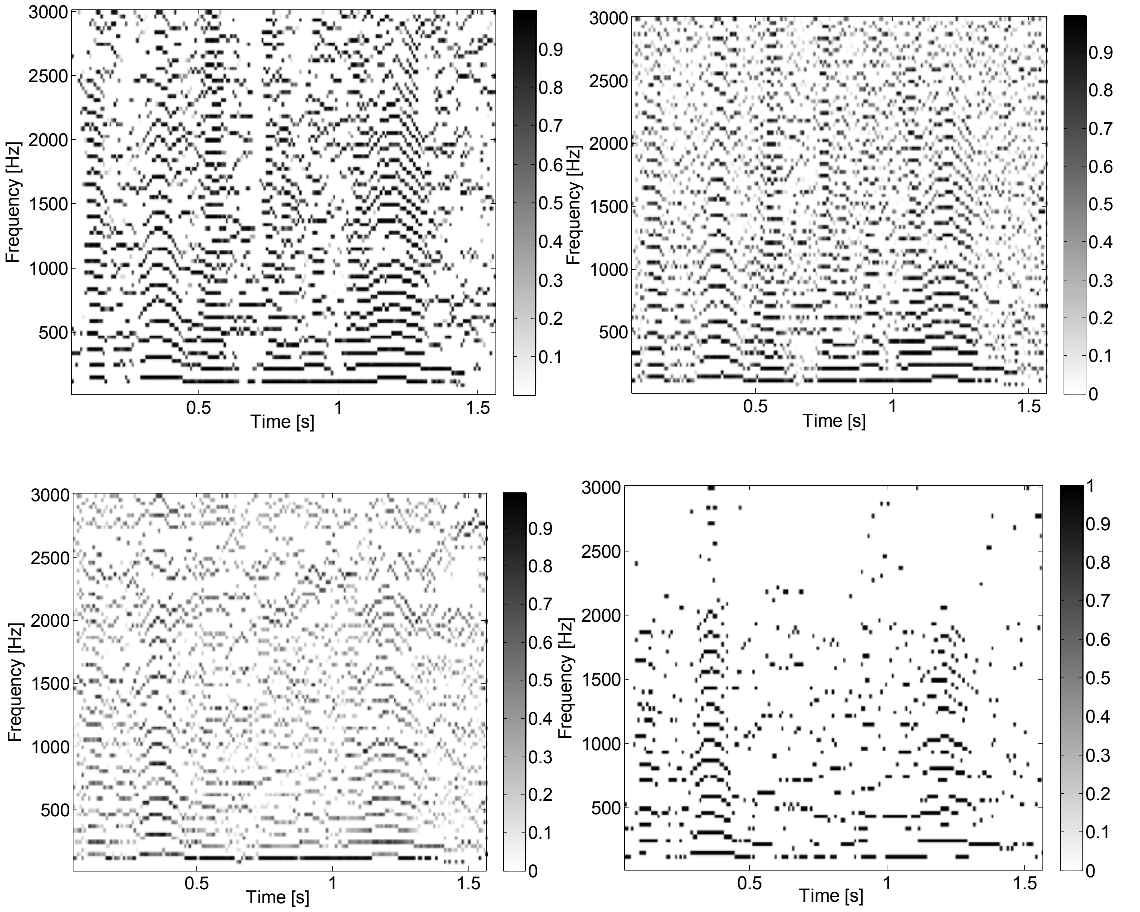


Fig. 6.7 Tonalograms of speech recording (clockwise from the top left: FTM, 1-UM, M1, 1-SFM)

The speech tonalogram for this discrete tonality detector contains only a few incomplete partials. It is interesting however, that this heuristic detector seems to operate significantly better in the case of the polyphonic recording which was analyzed using the frame length two times higher than the frame length which is intended to be used when the sampling of the signal is performed at 44100 Sa/s [67]. Considering the definition of the SFM (geometric to arithmetic mean of the power spectrum) it can be noticed that the tonality estimates provided by the SFM depends on the number of tonal components belonging to the particular band and the frequency distance between them. The 1–SFM value of the frequency band containing a large number of spectral peaks corresponding to the tonal components of a relatively low pitch would be lower than for a subband containing one or only a few strong tonal components and noise. This effect can be observed when comparing the tonalograms obtained using the SFM for speech (pitch is approximately equal to 100 Hz) and polyphonic recordings (there are strong tonal components related to the singer voice – see spectrogram in Fig. 6.6).

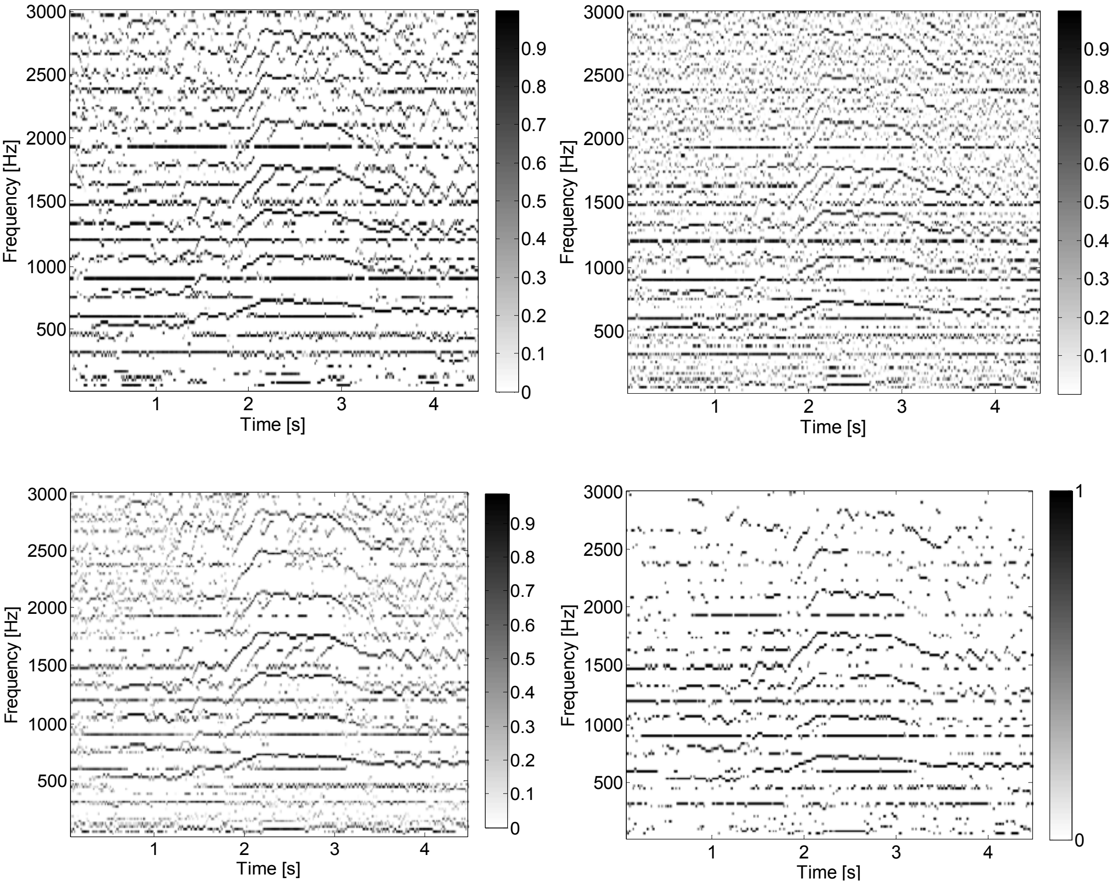


Fig. 6.8 Tonalograms of polyphonic recording (clockwise from the top left: FTM, 1-UM, M1, 1-SFM)

Although the graphical representation of tonality estimates yielded by the examined algorithms should be treated rather as an overview of the methods than a systematic comparison, it proves that the engineered algorithm operates as it was intended for the real-life sound sample recordings.

6.3 LIMITATIONS OF THE ALGORITHM

The frequency resolution of the DFT (Discrete Fourier Transform) analysis usually used in audio processing applications may not be high enough to produce a spectrum where low frequency tonal components are visible as local maxima. When the recordings containing very low pitch sounds (e.g. harmonics of bass guitar) are going to be analysed with the FTM algorithm, it may be necessary to employ the multi-resolution DFT approach [84]. Another solution is to use the UM method up to e.g. 300 Hz frequency instead of FTM, because it operates efficiently for low pitch sounds even if they are deeply modulated.

Considering implementation issues, it seems that computationally the FTM is slightly more complex than other investigated methods. However, with nowadays DSP (Digital Signal Processors) and other processors efficiency it should not be the problem. Because the probability that tonal components will occur in high frequency regions (i.e. over 10 kHz) is quite low, the hybrid approach may be used. In this case the FTM may be applied to signal frequencies up to some limit, and for higher frequencies the SFM or other method may be used.

7 RELIABILITY OF HEARING THRESHOLD ESTIMATION

7.1 PSYCHOACOUSTIC MODEL

As it was stated in the end of Section 5 the inverted tonality measure $1 - \mathrm{FTM}_{\mathrm{trk}}\left(T_{\mathrm{trk}}\left[r^{(l)}\right]\right)$ assigned to the spectral maxima $k_{\max}\left[i^{(l)}\right]$ detected within l-th frame is denoted further as ftm[$k_{\max}$]. The FTM algorithm is intended to provide a substitute for the UM method, used in the psychoacoustic model combined with the AAC or MP3 encoder (MPEG psychoacoustic model 2). However, there are two major differences between these algorithms, which make direct substitution impractical:

- for sinusoidal component of a particular SNR they provide slightly different values of tonality;
- UM indicates high tonality at spectral maxima and also neighbouring bins, when a strong sinusoidal component is present within the analyzed audio signal fragment [33]. The FTM assigns a high tonality value only at spectral peaks. Therefore, ftm[$k_{\max}$] values must be mapped into the UM space and appropriate spreading must be applied to the ftm[$k_{\max}$], before further experiments are carried out.

It is assumed that FTM should operate identically to the UM when audio signal contains stationary tonal components, because the UM operates perfectly in this case. The symbol notation used is identical to the symbol notation used within the ISO MPEG standard in all cases where it is reasonable [70].

7.1.1 FTM to UM mapping at spectral peaks

It is required to discover the relation behind the SNR of the analyzed sinusoidal components and the tonality estimates, produced by the UM and FTM algorithms at spectral peaks. Knowing such a characteristic, allows mapping the FTM into the UM at spectral peaks. In order to obtain the desired mapping characteristic, the set of constant frequency sinusoids of SNR varying from 50 to –30 dB with 1 dB step were generated and further analyzed. The sampling frequency was equal to 44100 Hz, the frame length was adjusted to 46.4 ms (2048 samples) and the hop size was equal to 23.2 ms (1024 samples). Every analyzed signal consisted of 88200 samples (2 s). The UM and FTM values corresponding to tonal component, of a priori known frequency, were stored in vectors for the signal of a particular SNR. Next, the mean values of these vectors were calculated in order to obtain the FTM to UM mapping function. The experiment was repeated for sinusoids of various frequencies, in order to reveal the influence of the tonal components frequency on the mapping characteristics. In Fig. 7.1 the mapping characteristics obtained for the sinusoids of instantaneous frequencies equal to 861.3 Hz and 870.8 Hz together with the estimated mapping function are presented (bin offset denoted as a k_{off} is derived from the quadratic interpolation procedure applied to the spectrum maximum and its neighboring bins as defined in (5.10) [1][15].

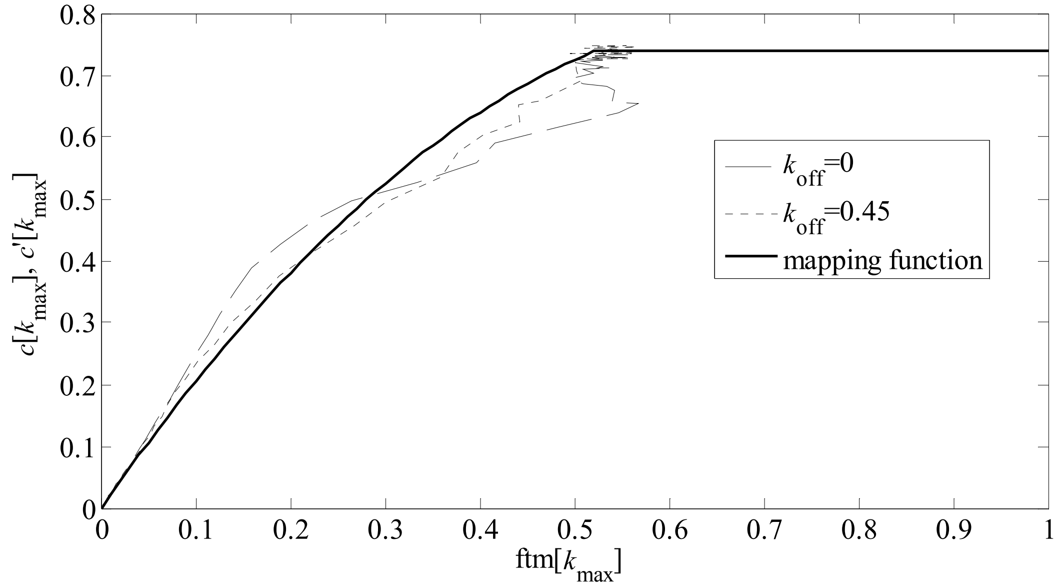


Fig. 7.1 FTM to UM mapping function

Since in the UM implementation given in the MPEG standard the $c[k_{\max}]$ symbol stands for UM values, here $c'[k_{\max}]$ symbol is used for values derived from FTM (frame index is omitted for clearance). The FTM to UM mapping function at spectral maxima is approximated as follows

$$c'[k_{\max}] = \begin{cases} -1.5\mathrm{ftm}^2[k_{\max}] + 2.2\mathrm{ftm}[k_{\max}], & \mathrm{ftm}[k_{\max}] \le 0.52 \\ 0.74, & \mathrm{ftm}[k_{\max}] > 0.52 \end{cases} \tag{7.1}$$

7.1.2 FTM spreading – method 1

The UM indicates high tonality not only at the spectral maxima corresponding to the particular tonal components, but also at bins neighboring these maxima. Obviously, this tonality spreading is directly related to the spectral characteristic of the window function used during STFT calculation. The UM indicates high tonality when the amplitude and phase of the spectral bins change linearly. When the strong stationary tonal component is present in the analyzed audio signal, the above-mentioned criterion is also met by the bins on the both sides of the spectral peak [93][94].

The psychoacoustic model combined with the AAC encoder generates hearing threshold in bands (partitions) corresponding to the approximately one third of the critical bands. Consequently, the $c[k]$ values produced by the UM algorithm are used in order to estimate the weighted unpredictability measure of threshold calculation partitions, according to the following formula (based on (3.7));

$$c[b] = \sum_{k=k_{\mathrm{low}}[b]}^{k_{\mathrm{high}}[b]} r^2[k]c[k] \tag{7.2}$$

where b is partition number, $k_{\mathrm{low}}[b]$ and $k_{\mathrm{high}}[b]$ are the partition boundaries and $r[k]$ is the magnitude spectrum in linear scale. Considering (7.2), it can be noted that $c[k]$ values are multiplied by the energy spectrum which is obviously smeared due to the spectral leakage. Therefore, the $c'[k_{\max}]$ estimates yielded by the FTM algorithm, must be spread over neighboring bins, before $c'[k]$ may be put into the (7.2) instead of $c[k]$. However, the spreading is applied only to the maxima meeting the following criterion

$$g[k_{\max}] > 2g_{\mathrm{thd}} \tag{7.3}$$

where g_{thd}=9 dB as it was stated in subsection 5.2.

In order to model the tonality spreading, the ratios

$$u_-(c[k_{\max}]) = \frac{c[k_{\max}-1]}{c[k_{\max}]} \tag{7.4}$$

$$u_+(c[k_{\max}]) = \frac{c[k_{\max}+1]}{c[k_{\max}]} \tag{7.5}$$

are first examined as a function of $c[k_{\max}]$. The sample characteristics obtained for stationary sinusoids of various SNRs, 865 Hz frequency and sampled at 44100 Hz are presented in Fig. 7.2.

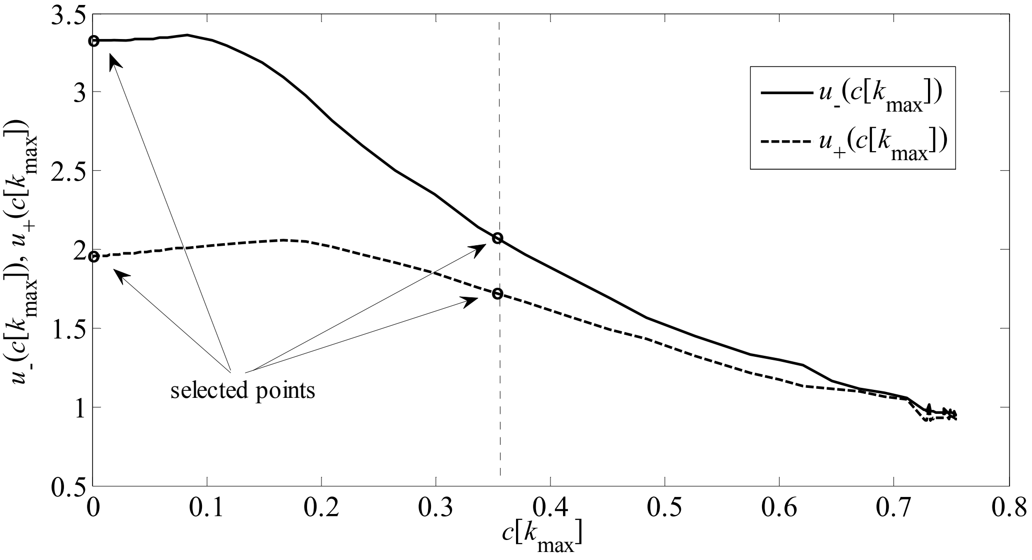


Fig. 7.2 Sample characteristics of ratios $u_-(c[k_{\max}])$ and $u_+(c[k_{\max}])$

The characteristics of $u_-(c[k_{\max}])$ and $u_+(c[k_{\max}])$ ratios presented in Fig. 7.2 differ from each other, because the sinusoid of frequency equal to 865 Hz has k_{off}≈0.18. But when k_{off}=0, both characteristics are identical. It was decided to model these characteristics using linear functions, passing the original curves for two selected arguments: $c[k_{\max}]\to 0$ and $c[k_{\max}]=0.37$ (the latter value was selected as a half of the mean $c[k]$ values when white Gaussisan noise is analyzed). Then, the FTM values for bins on the both sides of the spectral maximum are given by:

$$c'[k_{\max}-1]=a_-(k_{\text{off}})c'[k_{\max}]+b_-(k_{\text{off}}) \tag{7.6}$$

$$c'[k_{\max}+1]=a_+(k_{\text{off}})c'[k_{\max}]+b_+(k_{\text{off}}) \tag{7.7}$$

where $a_-(k_{\text{off}})$, $b_-(k_{\text{off}})$ and $a_+(k_{\text{off}})$, $b_+(k_{\text{off}})$ are the coefficients of the linear functions, which model characteristics of $u_-(c[k_{\max}])$ and $u_+(c[k_{\max}])$ ratios. In order to calculate these coefficients, it is necessary to know the characteristics values at points $c[k_{\max}]\to 0$ and $c[k_{\max}]=0.37$ for various values of k_{off}. These values denoted further as $u_{-0}(k_{\text{off}}), u_{-0.37}(k_{\text{off}})$ and $u_{+0}(k_{\text{off}}), u_{+0.37}(k_{\text{off}})$ were found experimentally, using sinusoids of various frequencies having k_{off} in the range between –0.48 and 0.48 (see Figs. 7.3 and 7.4).

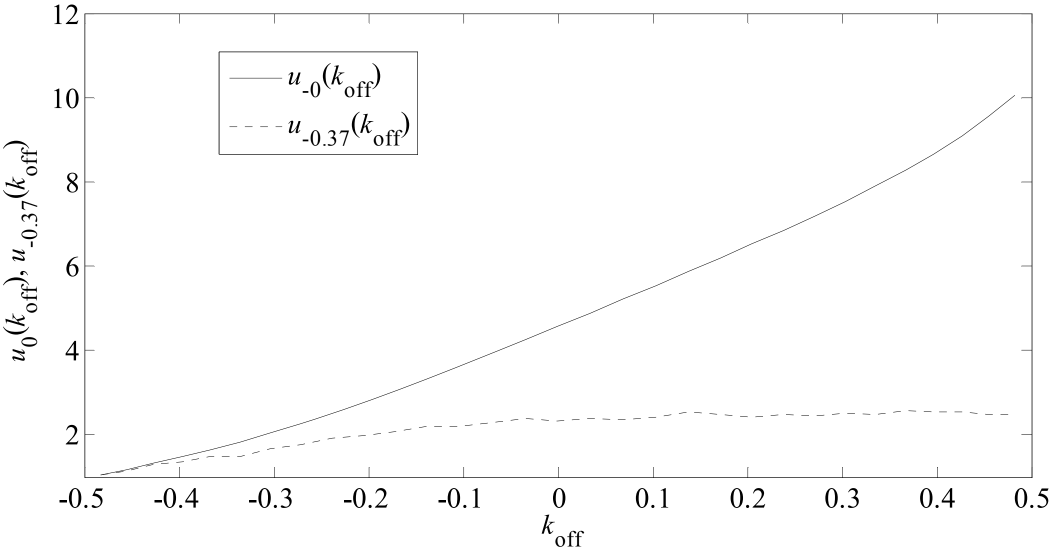


Fig. 7.3 Characteristics of ratios $u_{-}(c[k_{\max}])$ for $c[k_{\max}]\rightarrow 0$ and $c[k_{\max}]=0.37$ as a function of bin offset

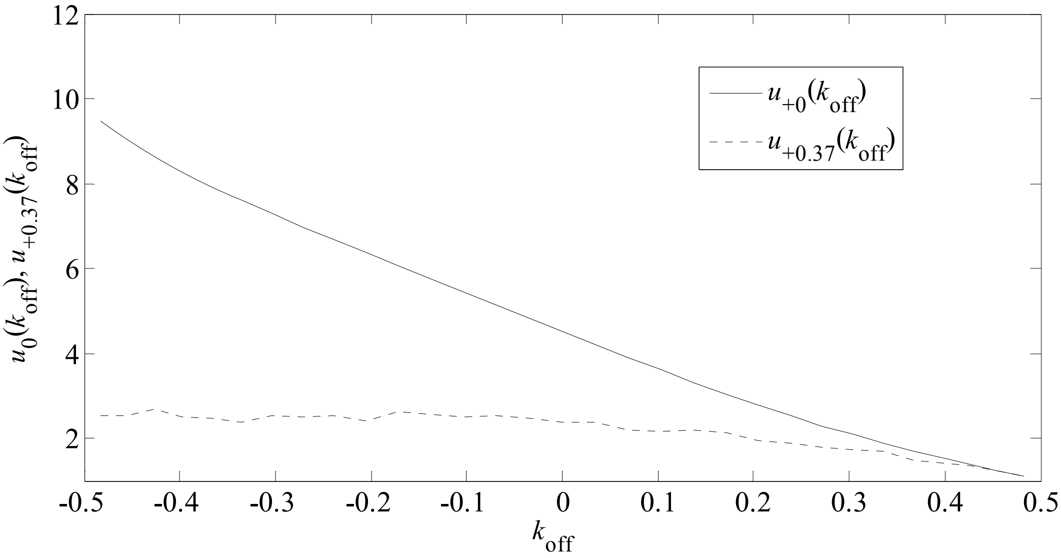


Fig. 7.4 Characteristics of ratios $u_{+}(c[k_{\max}])$ for $c[k_{\max}]\rightarrow 0$ and $c[k_{\max}]=0.37$ as a function of bin offset

As expected the characteristics presented in Figs. 7.3 and 7.4 are symmetrical each to the other. All curves obtained for $c[k_{\max}]\rightarrow 0$ and $c[k_{\max}]=0.37$ are modeled using quadratic functions as given by:

$$u_{-0}(k_{\text{off}}) = 2.4k_{\text{off}}^2 + 3.8k_{\text{off}} + 2.3 \tag{7.8}$$

$$u_{-0.37}(k_{\text{off}}) = -1.1k_{\text{off}}^2 + 1.1k_{\text{off}} + 1.8 \tag{7.9}$$

$$u_{+0}(k_{\text{off}}) = 2.4k_{\text{off}}^2 - 3.8k_{\text{off}} + 2.3 \tag{7.10}$$

$$u_{+0.37}(k_{\text{off}}) = -1.1k_{\text{off}}^2 - 1.1k_{\text{off}} + 1.8 \tag{7.11}$$

Next, the coefficients of linear functions given in (7.6) and (7.7) may be calculated

$$a_-(k_{\text{off}}) = \frac{u_{-0.37}(k_{\text{off}}) - u_{-0}(k_{\text{off}})}{0.37} \tag{7.12}$$

$$b_-(k_{\text{off}}) = u_{-0}(k_{\text{off}}) \tag{7.13}$$

$$a_+(k_{\text{off}}) = \frac{u_{+0.37}(k_{\text{off}}) - u_{+0}(k_{\text{off}})}{0.37} \tag{7.14}$$

$$b_+(k_{\text{off}}) = u_{+0}(k_{\text{off}}) \tag{7.15}$$

Having all necessary coefficients, it is possible to estimate the tonality values $c'[k_{\max} - 1]$ and $c'[k_{\max} + 1]$ surrounding spectral maximum using (7.6) and (7.7). Furthermore, if the bin offset $|k_{\text{off}}|>0.35$ then $c[k_{\max}]$ is spread also over the bin having the $k_{\max}$–2 or $k_{\max}$+2 index, then

$$c'(k_{\max} + 2) = [c'(k_{\max} + 1)]^2, \; k_{\text{off}} > 0.35 \tag{7.16}$$

$$c'(k_{\max} - 2) = [c'(k_{\max} - 1)]^2, \; k_{\text{off}} < -0.35 \tag{7.17}$$

When the tonal components have frequencies close each to the other, the spread $c[k]$ values may overlap. In this case the mean values are assigned to the spectral bins. For all remaining spectral bins, which were not detected as a tonal the $c'[k] = 0.74$ was assigned, because this is the mean $c[k]$ value obtained when white Gaussian noise is analyzed using the UM algorithm (see also Fig. 8.1 for ftm[$k_{\max}$]≥0.74).

Although the FTM spreading method described above allows exact modeling of the UM spreading, it is quite complex and requires high computational power [94]. Because increasing the computational complexity should be omitted in the coding applications, a simpler method for FTM spreading is proposed in the next subsection.

7.1.3 FTM spreading – method 2

In this subsection the simple, heuristic procedure for FTM spreading is proposed. The $c'[k]$ is spread over spectral bins laying on the both sides of the peak. The general rule is that the higher the peakiness of the spectral local maxima, the more tonal should be the spectral bins neighboring the peak. Thus, first the parameter related to the peakiness of tonal spectral peak is calculated according to:

$$m_g[k_{\max}] = \left(\frac{g[k_{\max}] - g_{\text{thd}}}{g_{\text{tnl}}[k] - g_{\text{thd}}} \right)^2 \tag{7.18}$$

where g_{thd}=9dB as it was indicated in subsection 6.3, and $g_{\text{tnl}}[k]$ is defined by:

$$g_{\text{tnl}}[k] = \begin{cases} 18 + \dfrac{12k}{\left(\left\lfloor \dfrac{Nf_{\text{c1}}}{F_{\text{s}}} \right\rfloor \right)^2}, & k < \left\lfloor \dfrac{Nf_{\text{c1}}}{F_{\text{s}}} \right\rfloor \\ 30, & k \geq \left\lfloor \dfrac{Nf_{\text{c1}}}{F_{\text{s}}} \right\rfloor \end{cases} \tag{7.19}$$

where f_{c1}=800 Hz was determined experimentally and $\lfloor\ \rfloor$ stands for rounding to the nearest integer towards minus infinity. The parameters expressing the energy relation between spectral bin corresponding to the peak and bins laying on the both sides of the peak are given by:

$$m_{e-}[k_{\max}] = \left(\frac{r[k_{\max} - 1]}{r[k_{\max}]} \right)^{0.5} \tag{7.20}$$

$$m_{e+}[k_{\max}] = \left(\frac{r[k_{\max}+1]}{r[k_{\max}]} \right)^{0.5} \tag{7.21}$$

where $r[k]$ denotes the magnitude spectrum in linear scale (as in the MPEG standard). Further, the parameters defined in (7.18), (7.20) and (7.21) are multiplied and bounded to the 1 value as in (7.21) and (7.23):

$$m_{ge-}[k_{\max}] = \begin{cases} m_g[k_{\max}]m_{e-}[k_{\max}], & m_g[k_{\max}]m_{e-}[k_{\max}] < 1 \\ 1, & \text{otherwise} \end{cases} \tag{7.22}$$

$$m_{ge+}[k_{\max}] = \begin{cases} m_g[k_{\max}]m_{e+}[k_{\max}], & m_g[k_{\max}]m_{e+}[k_{\max}] < 1 \\ 1, & \text{otherwise} \end{cases} \tag{7.23}$$

Finally, the spread $c'[k_{\max}-1]$ and $c'[k_{\max}+1]$ are calculated according to:

$$c'[k_{\max}-1] = c'[k_{\max}] + (c_{\text{ns}} - c'[k_{\max}])(1 - m_{ge-}[k_{\max}]) \tag{7.24}$$

$$c'[k_{\max}+1] = c'[k_{\max}] + (c_{\text{ns}} - c'[k_{\max}])(1 - m_{ge+}[k_{\max}]) \tag{7.25}$$

where c_{ns}=0.74 and is the mean $c[k]$ value obtained when white Gaussian noise is analyzed using UM algorithm. When the tonal components have frequencies close each to the other, the spread $c[k]$ values may overlap. In this case the lower values are assigned to the spectral bins. For all remaining spectral bins, which were not detected as a tonal, the $c'[k]$=c_{ns} is assigned.

7.1.4 Hybrid tonality estimator

The FTM algorithm requires that the tonal components are detectable as local spectra maxima. However, for low pitched sounds (e.g. a bass guitar), the frequency resolution of spectral analysis may be insufficient to meet this requirement – as it was stated in subsection 7.3. Therefore, the minimal value among the $c[k]$ and $c'[k]$ is chosen, for the frequencies up to the f_{c2}=300 Hz during hearing threshold estimation:

$$c''[k]=\begin{cases}\min\left(c[k],c'[k]\right), & k\le\left\lfloor\dfrac{2048 f_{\mathrm{c2}}}{F_{\mathrm{s}}}\right\rfloor\\ c'[k], & \text{otherwise}\end{cases} \tag{7.26}$$

The sample magnitude spectrum of the frame coming from polyphonic recording containing mainly stationary tonal components together with hearing thresholds estimated by the psychoacoustic model using $c[k]$ and $c''[k]$ is presented in Fig. 7.5 (spectral bins up to the quarter of the F_{s} are presented). The spreading procedure described in subsection 7.1.3 (method 2) was employed in order to calculate $c'[k]$ and $c''[k]$ basing on the $\mathrm{ftm}[k]$ values.

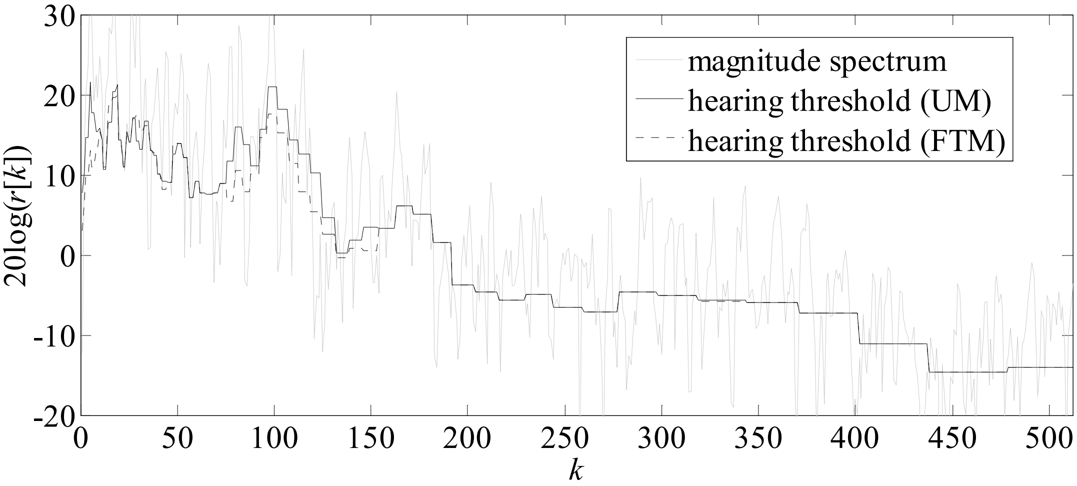


Fig. 7.5 Hearing thresholds estimated using UM and FTM methods for spectrum containing mainly stationary tonal components

It can be observed from Fig. 7.5 that the differences between hearing thresholds estimated using the FTM and UM methods are rather insignificant. This proves, that for recordings containing stationary tonal components the FTM operates similarly to the UM. Contrarily, in Fig. 7.6 the hearing thresholds obtained for recording containing male vocal vibrato is presented. Considering results shown in Fig. 7.6, it can be noted that the hearing threshold estimated using the FTM is approximately 10 dB lower than estimated using the UM, in partitions containing strong modulated tonal components. It is directly related to the limitations of the UM method, when modulated tonal components are present in analyzed audio material.

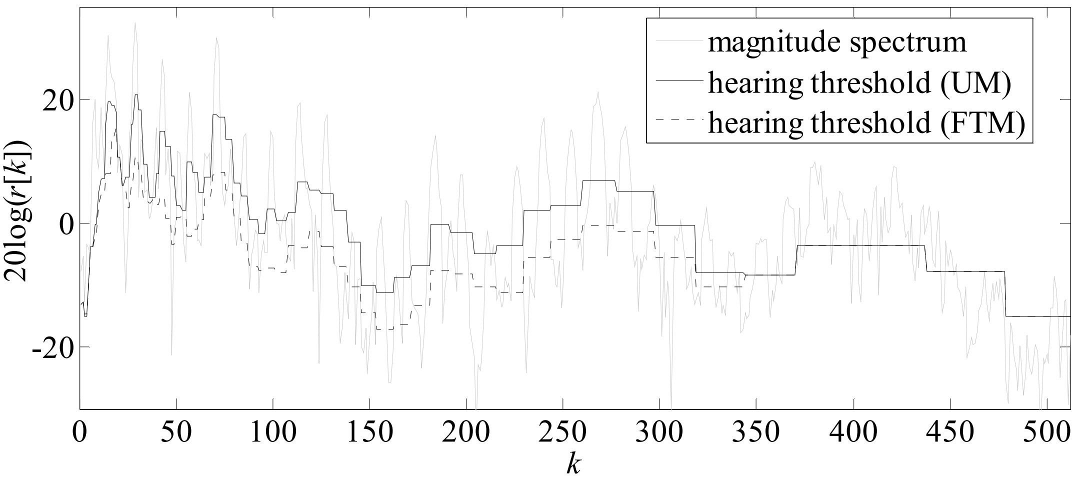


Fig. 7.6 Hearing thresholds estimated using UM and FTM methods for spectrum containing mainly modulated tonal components

The difference between presented hearing thresholds is similar to the difference between the TMN and NMT equal to the 12 dB in the AAC psychoacoustic model [17][70]. It can be then concluded that the FTM operates as it was intended, indicating high tonality values even if tonal components are deeply modulated.

7.2 PERCEPTUAL EVALUATION

7.2.1 Codec system

The reliability of hearing threshold estimates produced by the psychoacoustic model employing the UM and FTM methods, was evaluated employing the codec system based on the AAC structure presented in Fig. 7.7. The hearing threshold is generated using the MPEG psychoacoustic model 2 employing the UM or FTM algorithm, depending on the selection of the algorithm. The filter bank based on the MDCT and the sinusoidal window function (the switching between Kaiser-Bessel and sinusoidal window functions is not implemented), is used for calculating 1024 and 128 spectral coefficients for long and short blocks, respectively. However, for short MDCT blocks, the hearing threshold is always generated employing the UM method, because in this case the spectral resolution of the analysis is insufficient for a proper FTM operation.

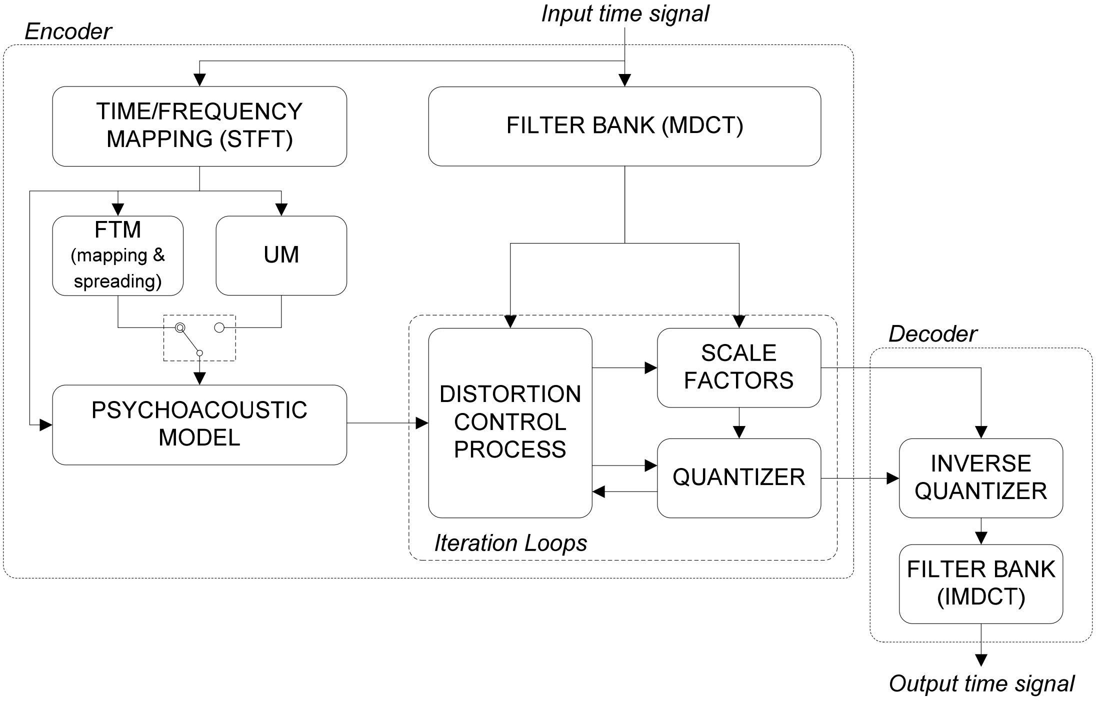


Fig. 7.7 The block diagram of codec system used for evaluation of reliability of hearing threshold estimation

The MDCT coefficients, grouped into the scale-factor bands are quantized according to the well known formula already given in (2.1) (it is repeated here for convenience) [17][70]:

$$x_q[k] = \left\lfloor \left| x^{0.75}[k] \right| 2^{0.1875\,(s_{\mathrm{cfc}} - s_{\mathrm{fc}}[k])} + 0.4054 \right\rfloor \tag{7.27}$$

where $x[k]$ are the MDCT coefficients, s_{cfc} is the common scale-factor and $s_{\mathrm{fc}}[k]$ are the scale-factors assigned to individual MDCT coefficient within particular scale-factor band:

$$s_{\mathrm{fc}}[k] = \{s_{\mathrm{fc}}[m]; \quad k_{\mathrm{low}}[m] < k < k_{\mathrm{high}}[m] \tag{7.28}$$

where $k_{\mathrm{low}}[m]$ and $k_{\mathrm{high}}[m]$ are the low and high boundaries of m-th scale-factor band.

The outer and inner iteration loops in the AAC encoder, determine the s_{cfc} and $s_{\mathrm{fc}}[m]$ based on the estimated hearing threshold and number of bits available for encoding of particular block. It is well known that the perceptual coding is based on the assumption that the transparent coding quality may be obtained when quantization noise is below hearing threshold [21][86]. Thus, the AAC procedure controlling the coding distortions was modified here so that the introduced quantization noise is always close to the estimated hearing threshold. In other words, it is assumed that the encoder always have the enough number of available bits, allowing keeping the quantization noise just on the hearing threshold level estimated using the UM or FTM method (Fig. 7.8).

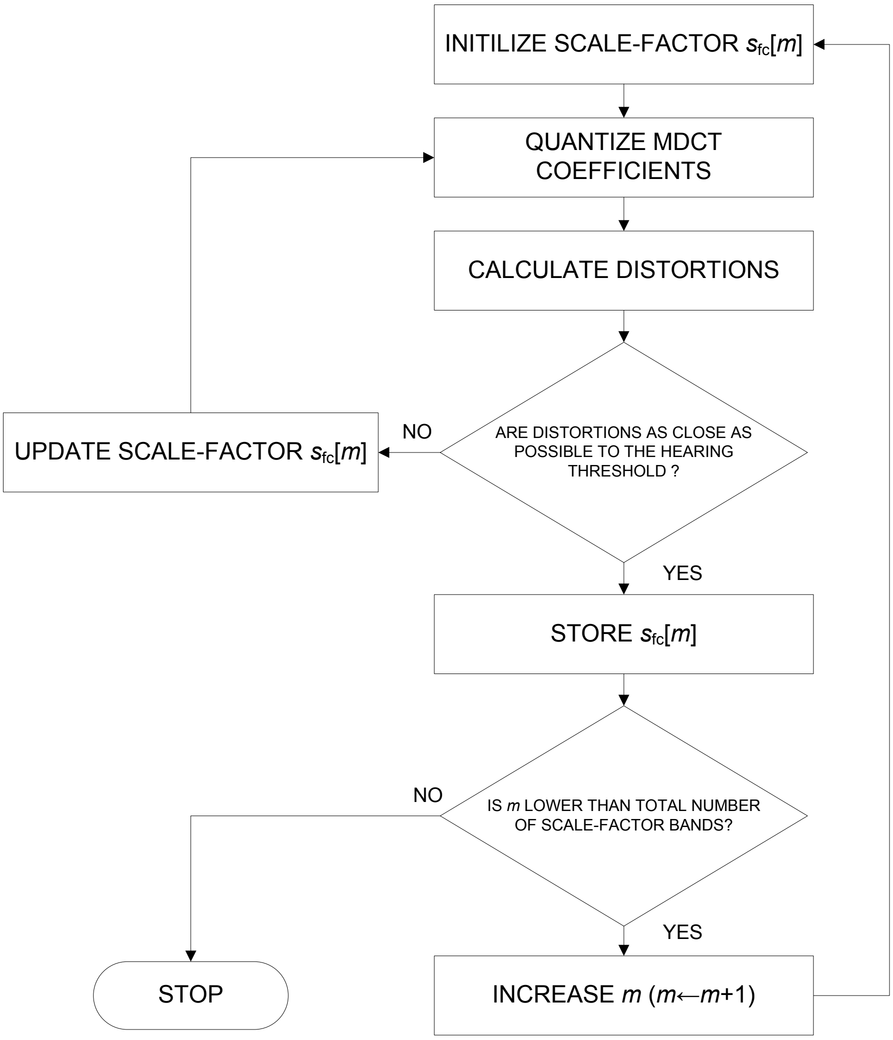

INITILIZE SCALE-FACTOR $s_{fc}[m]$
QUANTIZE MDCT COEFFICIENTS
CALCULATE DISTORTIONS
ARE DISTORTIONS AS CLOSE AS POSSIBLE TO THE HEARING THRESHOLD ?
NO
UPDATE SCALE-FACTOR $s_{fc}[m]$
YES
STORE $s_{fc}[m]$
IS m LOWER THAN TOTAL NUMBER OF SCALE-FACTOR BANDS?
NO
YES
STOP
INCREASE m ($m \leftarrow m$+1)

Fig. 7.8 The block diagram of distortion control

The MDCT coefficients belonging to scale-factor bands quantized according to (7.27) are inversely quantized basing on the s_{cfc} and $s_{fc}[m]$ in the decoder. The output time signal is produced after applying the inverse (IMDCT) transform to the decoded MDCT coefficients [63].

7.2.2 Testing procedure

The codec system presented in Fig. 7.7 was used during experiments. In order to reveal the influence of the tonality estimation method used in perceptual modeling on the reliability of the generated hearing threshold, two groups of mono sound samples were prepared. While the first group of sound recordings was taken from the twelve recordings containing mainly stationary tonal components, the second group comprises fourteen recordings containing vibratos or other modulated components produced by the instrumentalists or vocalists. The sound recordings were selected from the commercial recordings of various styles as well as from EBU-SQAM (European Broadcast Union-Sound Quality Assessment Material) Compact Disc containing a set of audio programme signals which are recommended by the EBU for subjective test purposes [156]. The F_s of all samples was 44100 Sa/s. Every sound recording was encoded so that the quantization noise was within ±1 dB range related to the hearing threshold estimated using the UM and FTM methods. Consequently, two quantized representations of every single sound recording were obtained.

The quality of coded sound samples were compared using the PEAQ (Perceptual Evaluation of Audio Quality) method combined within *Opera* system developed by the *Opticom* company [129]. The PEAQ algorithm may operate in basic or advanced mode. Since the precision of quality measurement was the key issue, the advanced mode of PEAQ algorithm was selected [73][157][160]. The sound sample encoded using psychoacoustic model with UM and FTM tonality estimators were compared to the reference signal resulting in two scores expressed in the Objective Difference Grade (ODG). The ODG is closely related to the Subjective Difference Grade (SDG), which is used in the listening tests performed according to ITU-R BS.1116 recommendation [72].

7.2.3 Analysis of results

The results of quality evaluation for groups of recordings containing stationary (denoted as “1S” through “12S”) and modulated (denoted as “1M” through “14M”) tonal components, are presented in Figs. 7.9 and 7.10, respectively.

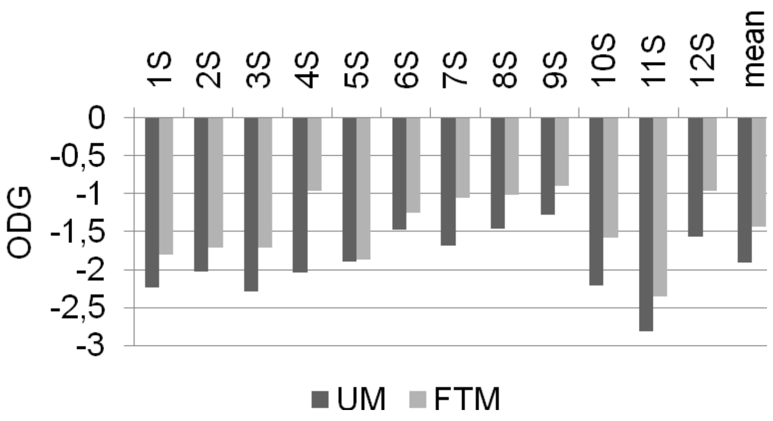


Fig. 7.9 The ODG scores obtained for sound recordings containing mainly stationary tonal components

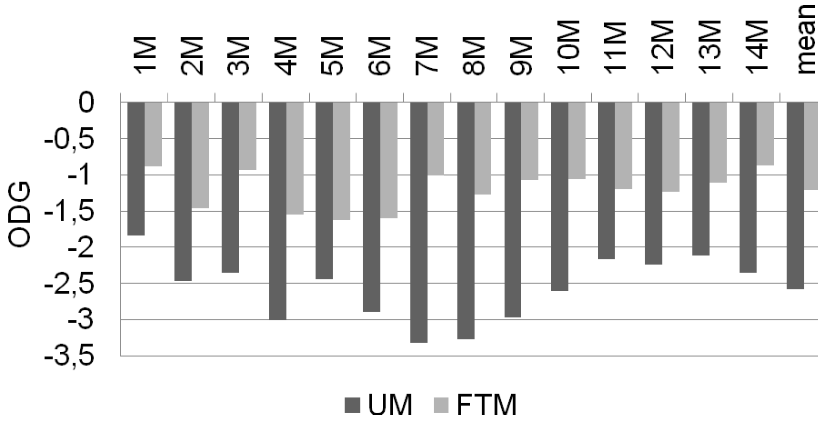


Fig. 7.10 The ODG scores obtained for sound recordings containing modulated tonal components

The mean difference between ODG scores obtained for sound recordings encoded, using psychoacoustic model with the FTM and UM methods is relatively low and equals 0.47 in case of the recordings containing mainly stationary tonal components. Furthermore, the mean ODG scores obtained using the UM and the FTM are around -1.5 for most sound recordings, which means that coding quality is “excellent” in both cases. However, when considering results presented in Fig. 7.10, it can be observed that the difference between mean ODG scores equals approximately –1.4. While the mean ODG score obtained for sound samples containing modulated tonal components and encoded using the UM tonality estimator is around –2.5, the mean ODG for psychoacoustic model employing the FTM method is higher than –1.5. Consequently, the quality provided by the encoder using the UM and FTM algorithms can be described as “good” and “excellent”, respectively. The presented results prove that using the FTM algorithm instead of the UM, increases the reliability of hearing threshold estimation [94].

It must be pointed out, that distortions introduced to the audio signals by the examined codec, providing that the quantization noise is just on the hearing level, are hardly perceived. This is due to two main reasons:

1. The MPEG psychoacoustic model 2 seems to estimate the hearing level slightly lower than it is [16];
2. When the quantization noise is just above the hearing threshold in a few subbands, the signal quality in these subbands is slightly affected. Introduced distortions results in negligible dynamic limitation rather than annoying coding artifacts.

Considering these observations, the further experiments, which tends to prove the advantages of the FTM method (see Section 9), are going to be carried out with more restricted constrains related to the number of bits available for signal encoding.

8 PERCEPTUAL NOISE SUBSTITUTION

The main reason for not recommending the UM method for noise-like bands detection is its limited efficiency in detecting sinusoids of varying frequency [148]. However, it was proved true that noise-like bands detection based on the tonality index derived from the psychoacoustic model employing the UM method, may be quite efficient for some types of music [171]. Because the FTM method is specially designed to deal with modulated components, it may be used as a basis for an efficient detector of noise-like bands. The experiments revealing the FTM usefulness for PNS technique are described in this Section.

8.1 DETECTION OF NOISE-LIKE SCALE-FACTOR BANDS

The AAC as well the MP3 algorithms, encode the MDCT spectral coefficients grouped within the scale-factor bands [17][20][66][68][70][71]. Thus, before PNS may be applied, it is necessary to detect scale-factor bands containing only pure noise-like components. These noise-like bands are detected here basing on three following parameters:

- tonality index derived from the $c''[k]$ – the scale-factor bands containing tonal components should be quantized and encoded according to the usual MP3/AAC procedure;
- flatness of the magnitude spectrum – although the FTM method is significantly more efficient in detecting tonal components than UM method, it may fail under some circumstances (e.g. when deeply modulated tonal components of low SNR are present within the analyzed audio signal). Additionally, when the noise in scale-factor band is coloured, the PNS should be omitted. This parameter is then used here in order to improve the reliability of noise detection and to prevent PNS being applied to the bands containing not white noise;
- energy variations – if noise energy varies significantly within the analyzed frame, filling particular scale-factor band with noise of constant energy would lead to audible distortions. Therefore, the energy variations are analysed in order to select only the bands containing stationary noise.

The methods for calculation of above-mentioned parameters, are presented in three following subsections. The further considerations concern the AAC algorithm operating with F_S=44100 Sa/s or 48000 Sa sampling rate. The scale-factor band boundaries are the same for both sampling rates. The PNS is applied here only to the long (N=2048) blocks.

8.1.1 Tonality index

In the procedure of hearing threshold estimation employed in the AAC encoder, the $c_b''[b]$ values calculated using (7.2) are spread across the partitions, according to the functions approximating masking patterns. Further, the tonality of every partition is calculated and clipped so that it is kept within [0,1] range, as it is determined by the term – similar to (3.12)

$$t_b[b] = -0.299 - 0.43\log(c_b''[b]) \qquad (8.1)$$

where $c_b''[b]$ is the $c''[b]$ spread across partitions according to the masking patterns being normalized. Next, the tonality of scale-factor bands is determined basing on:

$$t[m] = \max\left\{t_b\left[b_{\text{low}}[m]\right], \ldots, t_b\left[b_{\text{high}}[m]\right]\right\} \qquad (8.2)$$

where m=1, 2, …, 49 (assuming F_S=44100 or 48000 Sa/s) is the number of scale-factor band, b_{low}[m] and b_{high}[m] are the numbers of the lower and higher partitions occupying frequency range common with particular scale-factor band.

8.1.2 Flatness of scale-factor bands

The method used here for flatness estimation of scale-factor bands, is similar to the SFM (Spectral Flatness Measure) method, which was already introduced in subsection 4.1.2 [86][94]. Instead of calculating a group of SFMs parameters corresponding directly to the scale-factor bands, two representations of smoothed spectrum are estimated first. While the first representation is calculated using a moving arithmetic-average filter, the second one is obtained using a moving geometric-average filter. The filters are defined by:

$$r_a[k] = \frac{1}{M}\sum_{j=k-M/2}^{k+M/2-1} r[j] \qquad (8.3)$$

$$r_g[k] = \exp\left(\frac{1}{M}\sum_{j=k-M/2}^{k+M/2-1}\ln(r[j])\right) \tag{8.4}$$

where M is the filter order (the same for each filter), and $r_a[k]$, and $r_g[k]$ are the magnitude spectra smoothed using the moving arithmetic-average and geometric-average filters, respectively. The order of the filters was adjusted in relation to the widths of the scale-factor bands in the frequency range above 3.1 kHz (starting from 22nd scale-factor band). The order of the filters for scale-factor bands occupying frequencies from 3.1 to 5 kHz (22nd to 26th scale-factor band) was adjusted to M=8 and M=24 was used for the frequencies above 5 kHz (27th to 49th scale-factor bands). The flatness of spectral bins is given by:

$$z_{\text{bin}}[k] = 20\log\left(\frac{r_a[k]}{r_g[k]}\right) \tag{8.5}$$

The smoothed spectra calculated for a fragment of sample spectrum (from 6 to 12 kHz) using filters defined by (8.3) and (8.4) along with the flatness calculated according to (8.5), are presented in Fig. 8.1.

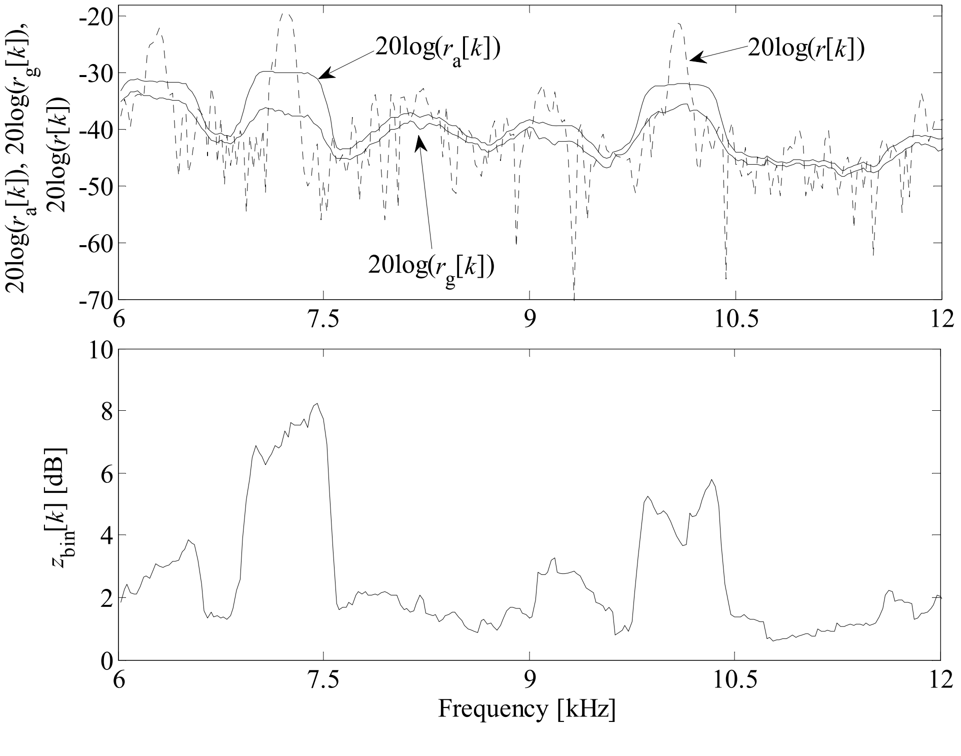


Fig. 8.1 Sample recording spectrum with its smoothed representations using (8.3) and (8.4) – upper plot; spectral flatness calculated according to the (8.5) – lower plot.

The spectrum presented in Fig. 8.1 is related to a sound sample recording containing violin vibrato. It can be noted that $z_{\mathrm{bin}}[k]$ is high for the regions of the spectrum containing evidently tonal components. Furthermore, also relatively high $z_{\mathrm{bin}}[k]$ is assigned to the spectrum parts having non-flat energy distribution over frequency region 9–9.5 kHz. Basing on the $z_{\mathrm{bin}}(k)$ the flatness of every scale factor-band is determined according to the formula:

$$z_{\mathrm{sfb}}[m] = \frac{1}{k_{\mathrm{high}}[m] - k_{\mathrm{low}}[m] + 1} \sum_{k=k_{\mathrm{low}}[m]}^{k_{\mathrm{high}}[m]} z_{\mathrm{bin}}[k] \tag{8.6}$$

where $k_{\mathrm{low}}[m]$ and $k_{\mathrm{high}}[m]$ are the spectral bin indices corresponding to the scale-factor band boundaries.

8.1.3 Energy variations

The PNS technique is applied here only to the long codec blocks (N=2048). Therefore, the scale-factor bands of the blocks with high energy variation, are excluded from the PNS analysis by the codec transient detector. The noise stationary within long blocks is checked, basing on the short AAC codec blocks (N=256) corresponding to 5.8 ms assuming F_{s}=44100 Hz. Although the human auditory system analyses the signal with approximately 2 ms resolution, for most cases the analysis with 5.8 ms step is sufficient [148]. Since short spectra are calculated by the codec for hearing threshold estimation of the short blocks, using them for energy variation analysis would not increase the computation load. The scale-factor band boundaries defined for long spectra, are mapped to the short spectra and then their energies are calculated. Subsequently, the standard deviation of the log-energies in every scale-factor band denoted as $\sigma[m]$ is determined by:

$$\sigma[m] = \sqrt{\frac{1}{8} \sum_{l_{\mathrm{s}}=1}^{8} \left(e_{nl}[m, l_{\mathrm{s}}] - \overline{e_{nl}}[m, l_{\mathrm{s}}] \right)} \tag{8.7}$$

where $e_{nl}[m,l_{\mathrm{s}}]$ is the log- energy of m-th scale factor band in l_{s}=0, 1, …, 7 short block, and $\overline{e_{nl}}[m,l_s]$ is the mean log-energy in m-th scale factor band for 8 following short spectra.

8.1.4 Parameter thresholds

The recordings of various styles were analyzed, in order to determine the appropriate thresholds for $t[m]$, $z_{\text{sfb}}[m]$ and $\sigma[m]$ allowing efficient detection of noise-like scale-factor bands. Finally, the criteria were formulated as:

$$\mathrm{PNS}[\mathrm{m}] = \begin{cases} 1, & t[m] \le t_{\text{thd}} \wedge z_{\text{sfb}}[m] < 2 \wedge \sigma[m] < 6 \\ 0, & \text{otherwise} \end{cases} . \tag{8.8}$$

where $t_{\text{thd}} = 0.01$ for high bit-rate mode of codec operation, and $t_{\text{thd}} = 0.05$ for low bit-rate mode, and $m = 22, 23, \ldots, 49$. The experiments revealed, that in very low bit-rate mode of codec operation it is even better to substitute the scale-factor bands of weak tonality with noise.

8.2 EVALUATION OF IMPLEMENTED PNS MODULE

The codec system used for evaluation of implemented PNS module is based on the AAC structure and is similar to the one presented in Fig. 7.7 (Fig. 8.2).

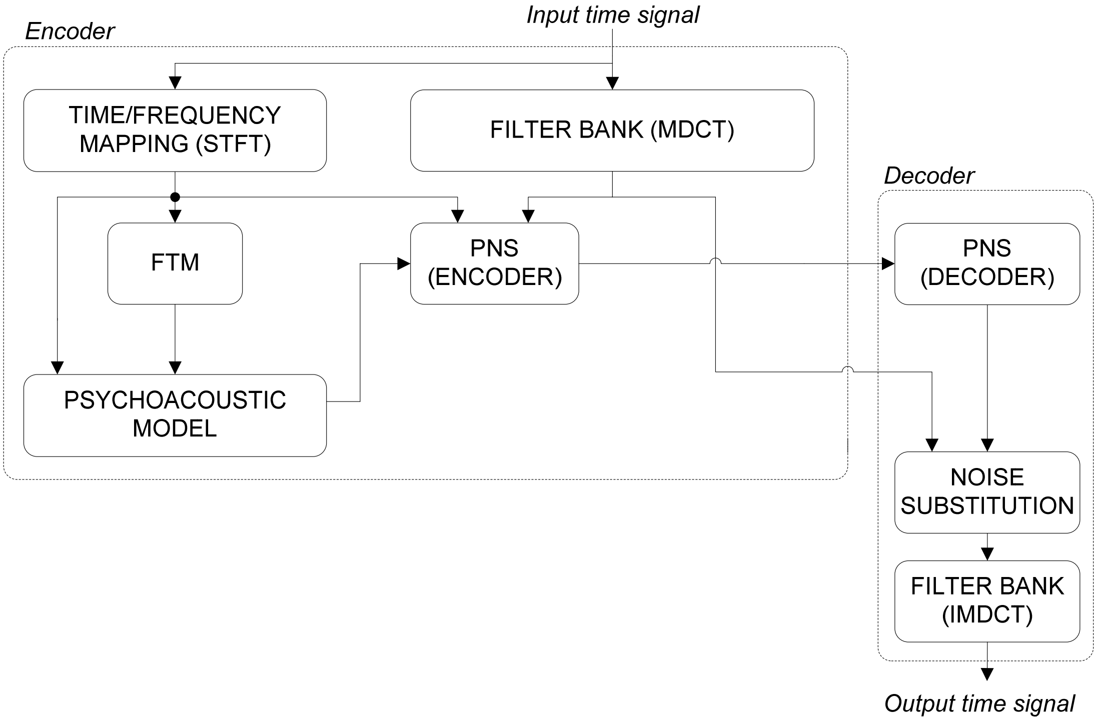


Fig. 8.2 The block diagram of codec system used for evaluation of implemented PNS module

The description of the AAC system was presented in subsection 2.2.2. The codec system employed for assessment of reliability of hearing threshold estimation, was described in details in subsection 7.2.1. Therefore, here only the codec system details important for experiments related to the PNS module are characterized.

The general idea behind the experiments carried out, was to replace the noise-like scale-factor bands by noise, while keeping all tone-like scale-factor bands non-quantized. The scale-factor bands detected by the PNS module and classified as noise-like ones, are filled by the decoder, with white Gaussian noise of energy equal to the energy of scale-factor bands of original signal. In fact, the noise is generated in the time domain and further transformed into the frequency domain. Then it is weighed adequately, and put into the noise-like scale-factor band. While the tone-like bands are just the copy of the bands from the original signal, the noise-like bands contain MDCT coefficients, corresponding to the locally generated noise. Finally, the output time signal is produced after applying the inverse (IMDCT) transform to the all MDCT coefficients.

8.2.1 *Efficiency of noise-like bands detection*

The particular scale-factor band is classified as noisy, if three conditions given in (8.8) are met simultaneously. In order to evaluate the influence of every single condition on the final decision regarding the PNS, the statistics can be calculated as in:

$$q_* = \frac{n_{\mathrm{sfb}*}}{n_{\mathrm{asfb}}} 100\,\% \tag{8.9}$$

where n_{asfb} represents the total number of scale-factor bands that were considered to be substituted by noise in the particular sound recording sample and $n_{\mathrm{sfb}*}$ is the number of scale-factor bands fulfilling one of the conditions: $t[m]<t_{\mathrm{thd}}$, $z_{\mathrm{sfb}}[m]<2$, $\sigma[m]<6$ or all of them simultaneously. The asterisk should be then replaced with t, z_{sfb}, σ or PNS, respectively.

8.2.2 *Subjective evaluation*

In order to evaluate the perceptual efficiency of the proposed PNS implementation, the quality of sound samples encoded using the PNS were compared to the quality of sample recordings encoded by the ordinary perceptual coding algorithm. Therefore, every excerpt: M1–M5 and S1–S3 was encoded twice, and further two obtained signal representations were compared to each other, in order to reveal whether the PNS module affects the signal quality or not. The listening tests were performed according to the ITU-R BS.1116 recommendation [26][72]. Three stimuli were presented to the experts in every listening test: the reference, stimulus A and stimulus B. The reference signals were original , not coded sound sample recordings. Either stimulus A or B was identical to the reference. The other remaining stimulus in every triple was encoded using the codec system presented in Fig. 8.2 with the PNS module enabled. The listeners had to appoint whether stimulus A or B is identical to the reference, and score the remaining one (see Appendix 1). The scoring range is from 0 (no quality difference) to 5 (totally degraded quality). The sound samples were presented to the users through high quality headphones, using high quality sound card inside a quiet room. Every stimulus could be repeated as requested by the experts as many times as it was required to make a final decision.

Since the performed subjective tests should be treated as pilot-tests, allowing to validate preliminary the implemented PNS module, the simple method for subject post-screening was used [172]. If the expert incorrectly scores the stimuli A or B being indeed the copy of the reference signal, this result was excluded from the analysis. When this situations occurred more than twice for a particular listener, he/she was treated as an unreliable expert and all his/her test results were neglected. Finally, ten experts among twelve, involved in the testing procedure met the above-mentioned criteria. The mean scores expressed in Subjective Difference Grade (SDG) for all evaluated sound sample recordings, along with 95% confidence intervals, are presented in Fig. 8.3. In Fig. 8.4 the q_{PNS} ratios are shown.

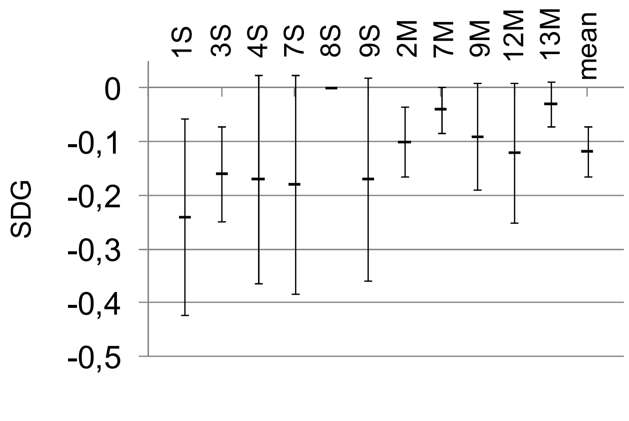


Fig. 8.3 The SDG scores obtained for the PNS module employing the FTM algorithm

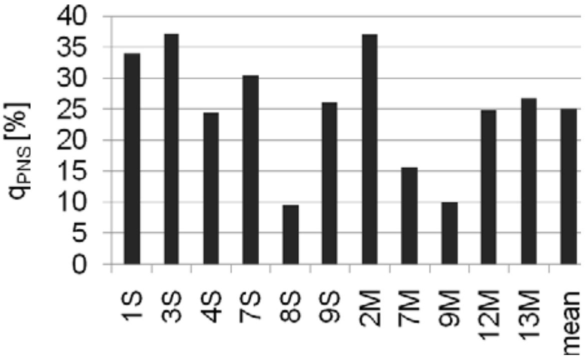


Fig. 8.4 Ratio of number of scale-factor bands substituted by noise to total number of bands considered for applying the PNS.

It can be noted, that the SDG does not fall below the value of –0.25 for any sound sample recording, encoded using the PNS technique. Consequently, substituting scale-factor bands by locally generated noise does not introduce perceptible coding artifacts to the coded recordings. This proves, that PNS module employing the FTM based detector of noise-like bands operates efficiently. The q_{PNS} varies depending on the content of the analyzed recording from approximately 9 to 37 % and is equal to 25%, in average. The amount of saved bits may be then used, to increase the coding accuracy of more demanding scale-factor bands occupying lower frequency regions [148].

9 DETERMINING SIGNIFICANCE OF TONALITY INDEX ESTIMATION TO AUDIO CODING QUALITY

9.1 EXPERIMENTAL CODEC STRUCTURE

All the experiments described in this Section were carried out using the coding system being a combination of systems presented in Fig. 7.7 and Fig. 8.3. The system presented in Fig. 9.1 may operate in one of the five modes described in Tab. 9.1.

Table 9.1 Operation modes of experimental coding system

Mode	PNS	Tonality estimation	Description
1	Inactive	UM	All scale-factor bands are quantized basing on the hearing threshold generated by psychoacoustic model employing the UM tonality estimator.
2	Inactive	FTM	All scale-factor bands are quantized basing on the hearing threshold generated by psychoacoustic model employing the FTM tonality estimator.
3	Inactive	None	All scale-factor bands are quantized basing on the hearing threshold generated by psychoacoustic model without any tonality estimator (all spectral components are treated as noise-like ones).
4	Active	UM	Noise-like scale factor bands are encoded employing the PNS module. Tone-like scale-factor bands are encoded basing on the hearing threshold generated by psychoacoustic model employing the UM tonality estimator. Bits saved by the PNS module are reused during tone-like bands encoding.
5	Active	FTM	Noise-like scale factor bands are encoded employing the PNS module. Tone-like scale-factor bands are encoded basing on the hearing threshold generated by psychoacoustic model employing the FTM tonality estimator. Bits saved by the PNS module are reused during tone-like bands encoding.

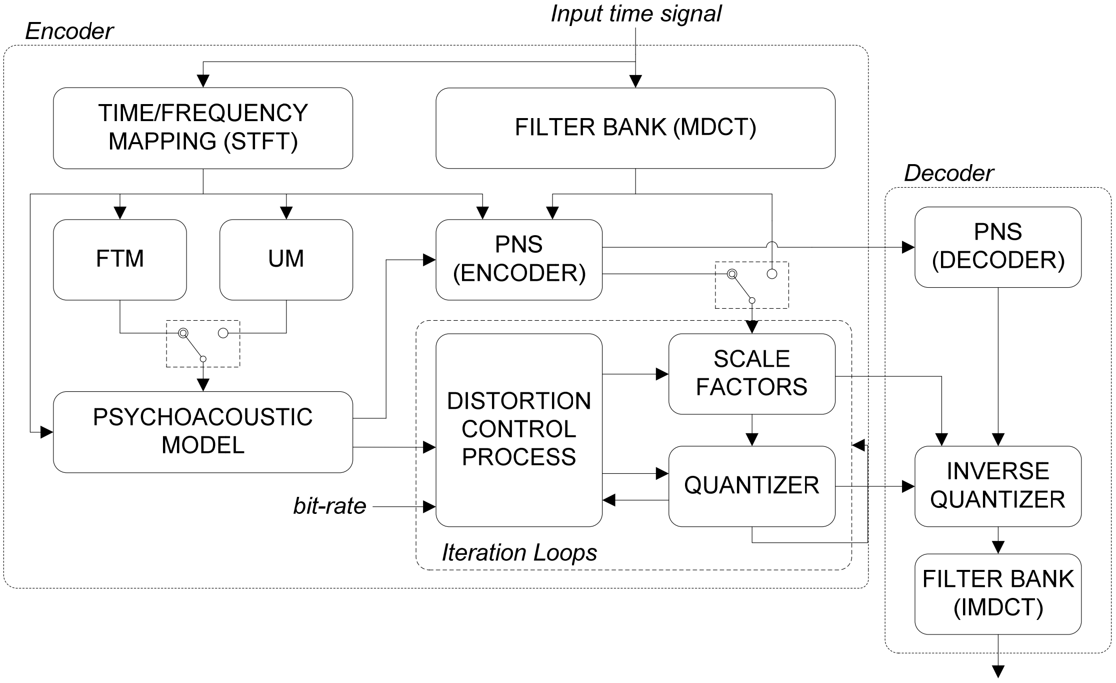


Fig. 9.1 The block diagram of experimental codec system

The AAC bit-stream is not formed within the system presented in the Fig. 9.1, because the researched methods were focused only on the lossy part of the coder. The codec bit-rate was scaled here using modified formula of perceptual entropy provided by the MPEG psychoacoustic model. Although the MPEG standard defines the number of bits required for encoding of a particular block basing on the perceptual entropy, it is not adequate to codec bit-rate. Therefore, the bit amount calculated according to the formula given in the MPEG standard is multiplied by the constant term found experimentally as [70]:

$$b_{\mathrm{alc}} = w_{\mathrm{PE3}} b_{\mathrm{alcAAC}} \tag{9.1}$$

where w_{PE3}=1.8. The formula for perceptual entropy alternative to (3.17) is given by:

$$\mathrm{PE} = \sum_{n=1}^{49} \left[\left(k_{\mathrm{high}}[m] - k_{\mathrm{low}}[m] \right) \log 10 \left(\frac{1}{\mathrm{ISMR}[m]} \right) \right] \tag{9.2}$$

where $k_{\mathrm{low}}[m]$ and $k_{\mathrm{high}}[m]$ are the spectral bin indices corresponding to the scale-factor band boundaries, and ISMR[m] stands for inverted SMR[m] given by (3.18).

The b_{alc} was treated as a number of bits required to encode a particular signal block. It was found that 80 kbps is enough to quantize the MDCT coefficients related to mono audio channel, while keeping the introduced distortions below hearing threshold or just at its level. In the experiments, the audio samples were encoded with 64 kbps and 48 kbps rates. The encoding process described in details in two following subsections, comprises of two major stages:

1. Bit allocation a – find the SNR[m] for all scale-factor bands, so that the $b_{\mathrm{alc}}\leq b_{\mathrm{avb}}$ (b_{avb} stands for number of bits available for encoding of the particular block and is equal to the bit-rate divided by the number of blocks per second).
2. MDCT quantization – quantize the MDCT coefficients so that the introduced distortions are as close as possible to the SNR[m] calculated in the previous step.

9.1.1 Bit allocation procedure

There are various coding quality optimization methods, when the bit-rate is too low for keeping the introduced distortions below hearing threshold. These methods vary depending on the codec implementations, and are not described within the MPEG standards. A few approaches to this issue have been evaluated in the initial state of the research, and finally a reasonable combination of them was selected. It is a common practice to limit the encoded signal bandwidth according to the bit-rate constrains [45]. If this is the case, more bits may be distributed over scale-factor bands occupying lower frequency regions. Therefore, while in the 64 kbps mode of codec operation the entire signal spectrum is encoded, in the low bit-rate mode the encoded signal band is limited to the 16 kHz. However, even if the bandwidth is limited, the available bit amount may be insufficient for keeping the quantization noise below hearing threshold. Consequently, the codec must efficiently assign bits to the scale-factor bands, basing on the SNRs provided by the psychoacoustic model. Decreasing the SNR in the scale-factor bands, which is relatively low (e.g. 6 dB), leads to some very annoying artifacts. On the other hand, the most bit-requiring scale-factor bands are the bands containing tonal components – the psychoacoustic model may indicate that the SNR should be up to 18 dB for assuring their transparent coding. Decreasing the SNR in these bands affects the coding quality, but does not lead to the annoying artifacts providing that the SNR is higher than approximately 6 dB. Therefore, if the bit-rate available for encoding of the particular block is lower than the bit-rate required for transparent coding, the distortions are introduced to the scale-factor bands proportionally to the SNR, indicated by the psychoacoustic model. The scale factors s_{cfc} and $s_{\mathrm{fc}}(m)$ are adjusted here, so that the distortion introduced by the quantizer are as close as possible to the distortions determined by the iterative process defined by:

$$\mathrm{SNR}\left[m, s+1\right] \leftarrow \left(1-\eta\right)\mathrm{SNR}\left[m, s\right] \tag{9.3}$$

and illustrated in Fig. 9.2. In (9.3) m=1, 2, …, 49 is the index of scale-factor band as it was defined previously, s is the iteration number, and η=0.015%. The SNR[m,1] is equal to the Signal-to-Mask-Ratio provided by the psychoacoustic model.

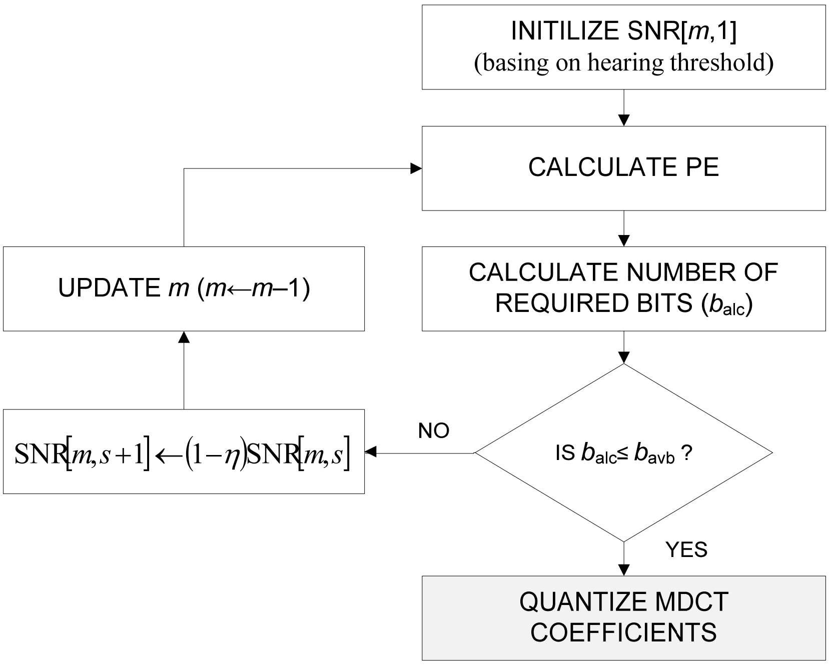

INITILIZE SNR[m,1]
(basing on hearing threshold)
CALCULATE PE
CALCULATE NUMBER OF REQUIRED BITS (b_{alc})
UPDATE m (m←m–1)
IS $b_{alc} \le b_{avb}$?
NO
$\mathrm{SNR}[m,s+1] \leftarrow (1-\eta)\mathrm{SNR}[m,s]$
YES
QUANTIZE MDCT COEFFICIENTS

Fig. 9.2 Bit allocation procedure

Since the lower frequency region is of primary importance for audio quality, the distortions are introduced starting from highest scale-factor band towards the lower bands. Considering (9.3) it can be noted that the higher the SNR required for particular band, the more distortions are introduced to this band in the single iteration step. The iterative process terminates when the bit-rate requirements are fulfilled.

9.1.2 MDCT coefficients quantization

The procedure for quantization of MDCT coefficients is almost identical to the one described in subsection 7.2.1. The only difference is related to the level of distortions introduced by the quantizer. In the experiments described in Section 7 the level of introduced distortions was as close as possible to the estimated hearing threshold regardless of the available number of bits. Here, the MDCT coefficients are quantized so that the introduced distortions are as close as possible to the SNR[m] generated by the bit allocation procedure, described in subsection 9.1.1. The diagram illustrating the iterative process of MDCT coefficient quantization is presented in Fig. 9.3.

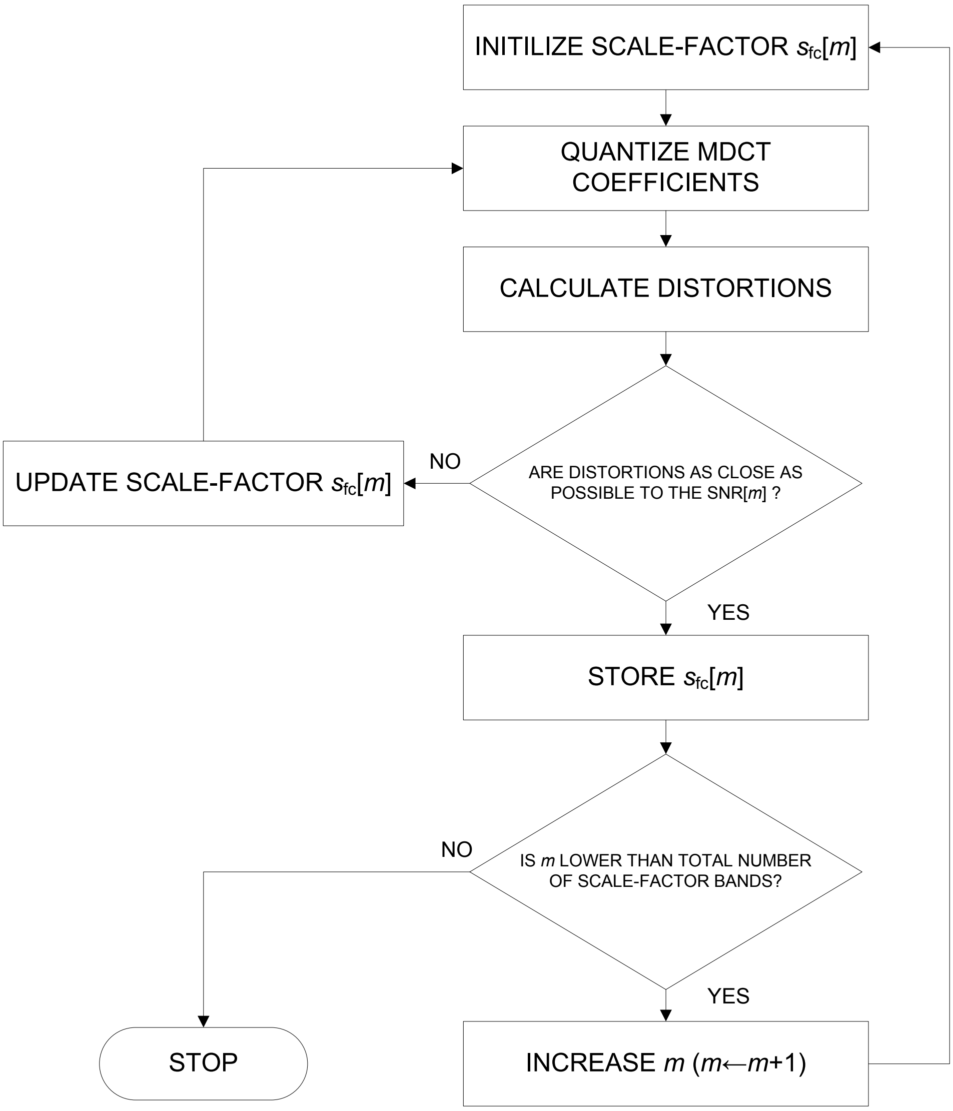

INITILIZE SCALE-FACTOR $s_{fc}[m]$
QUANTIZE MDCT COEFFICIENTS
CALCULATE DISTORTIONS
ARE DISTORTIONS AS CLOSE AS POSSIBLE TO THE SNR[m] ?
NO
UPDATE SCALE-FACTOR $s_{fc}[m]$
YES
STORE $s_{fc}[m]$
IS m LOWER THAN TOTAL NUMBER OF SCALE-FACTOR BANDS?
NO
YES
STOP
INCREASE m ($m \leftarrow m$+1)

Fig. 9.3 Block diagram of MDCT coefficients quantization process

The description of the MDCT coefficients quantization was already described in subsection 9.1.1. Therefore it would be redundant to repeat it here.

9.1.3 Codec operation with active PNS module

When the PNS module is active, the scale-factor bands detected as the noise-like ones are not quantized according to the (7.27). Instead of this, the energy of these bands is mapped into the procedure which fills the frequency bands with locally generated white Gaussian noise. It is assumed here that encoding of the noise energy and PNS[m] mask consumes 12 bits for single scale-factor band. The PNS is applied to the particular scale-factor band, provided that the number of saved bits is higher than 12. Obviously, the number of saved bits may vary from block to block significantly. Therefore, the additional number of bits available for encoding of tonal scale-factor bands within particular block represent the average value of bits saved by the PNS module among 12 blocks (>250 ms). The procedure allowing calculation of the bits saved by the PNS module which can be further used for tonal scale-factor bands encoding, is presented in Fig. 9.4.

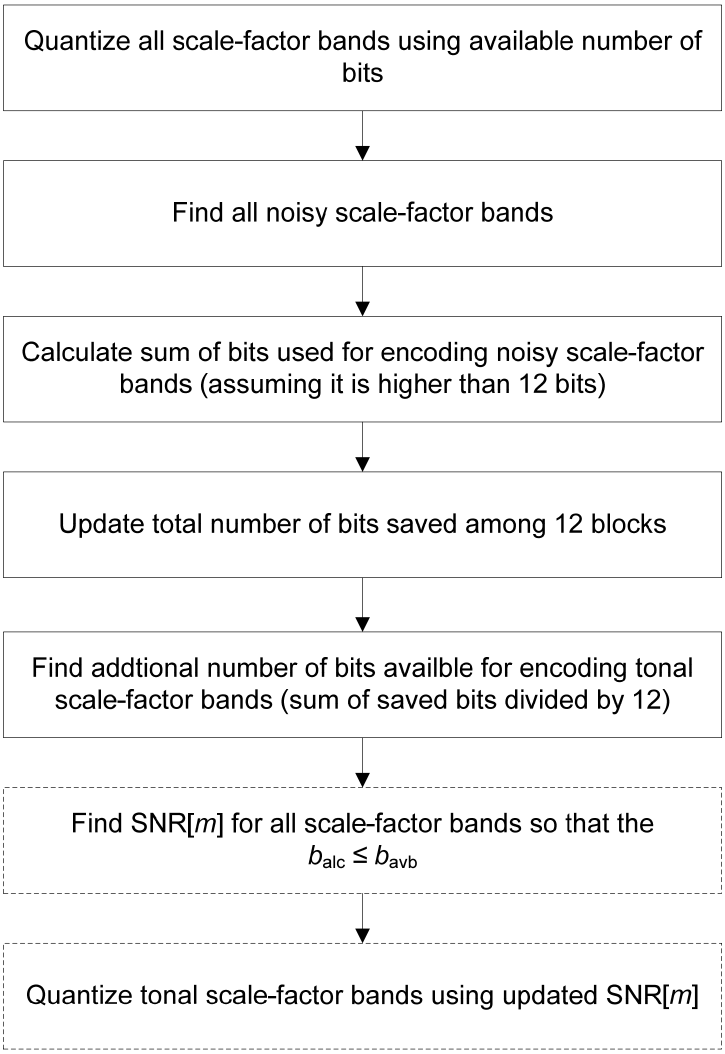

Quantize all scale-factor bands using available number of bits
Find all noisy scale-factor bands
Calculate sum of bits used for encoding noisy scale-factor bands (assuming it is higher than 12 bits)
Update total number of bits saved among 12 blocks
Find addtional number of bits availble for encoding tonal scale-factor bands (sum of saved bits divided by 12)
Find SNR[*m*] for all scale-factor bands so that the $b_{\mathrm{alc}} \leq b_{\mathrm{avb}}$
Quantize tonal scale-factor bands using updated SNR[*m*]

Fig. 9.4 Procedure for calculation the number of bits available for tonal scale-factor bands encoding when PNS module is enabled

Surprisingly, when the codec operates in very low bit-rate mode, filling the noisy bands during decoding with non-quantized noise may lead to some unpleasant audible effects related to the lack of coherence of signal brightness. In some circumstances, the scale-factor band substituted by non-quantized noise may be perceived as more bright than the other band quantized according to (7.27). Therefore, in the experiments carried out, the noisy scale-factor bands were filled with quantized noise (SNR was 5.5 and 5 dB for 64 and 48 kbps, respectively).

9.1.4 Decoder

The MDCT coefficients belonging to either tonal or all scale-factor bands, quantized according to the (7.27) are inversely quantized in the decoder basing on the s_{cfc} and $s_{\mathrm{fc}}[m]$. If the PNS module is enabled, the noisy bands are filled with scaled, locally generated noise as it was described in Section 8. Finally, the output time signal is produced after applying the inverse transform to the decoded MDCT coefficients.

9.2 LISTENING TESTS

9.2.1 Excerpts selection

The sound excerpts were selected among recordings used for evaluation of reliability of hearing threshold estimation. The notation concept is the same as previously ("M" stands for excerpts containing modulated tonal components and "S" for recordings containing mostly stationary tonal components). However, the indices of excerpts do not correspond directly with the indices of excerpts used in experiments described in subsection 7.2.

The significance of tonality index estimation to the audio coding efficiency was evaluated using a group of 8 sound excerpts selected as a critical material. While 5 recordings (denoted as M1–M5) contain modulated tonal components (guitar vibrato, singer vibrato, etc.), 3 of them (denoted as S1–S3) contain mainly stationary tonal components. Every sound recording was processed by the codec operating at 64 kbps and 48 kbps rate and in the following 4 modes (described in subsection 9.1):

- Mode 1: the tonality index $c_b[b]$ was calculated basing on the $c[k]$values (the UM method).
- Mode 2: the tonality index $c_b[b]$ was calculated basing on the $c'[k]$ values (the FTM method).
- Mode 3: the tonality index $c_b[b]$ was set to 1 for all partitions used by the psychoacoustic model. In this mode all scale-factor bands were treated as noise-like.

- Mode 5: the tonality index $c_b[b]$ was calculated basing on the $c'[k]$ values (the FTM method) and the PNS module was active.

Mode number 4 was omitted in order to reduce the time required for sample recordings evaluation and effort of the experts. Additionally, the statistics related to the PNS module were calculated for all selected sound sample recordings according to the (8.9). The results obtained for 8 excerpts assuming t_{thd}=0.01 are presented in Fig. 9.6.

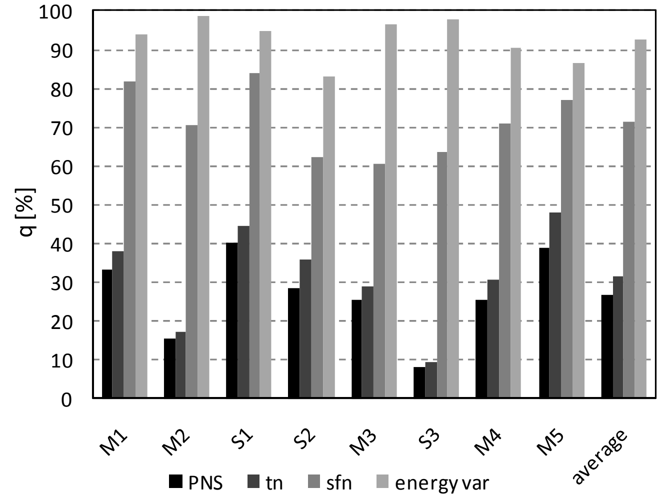


Fig. 9.5 Efficiency of PNS module for t_{thd}=0.01

It can be noted from Fig. 9.5, that in average, 27% of all scale-factor bands were substituted with noise by the PNS module. Obviously, the number of substituted bands depends strictly on the content of the recordings. While within S3 sample recording less than 9% were substituted, within M5 up to 40% of considered bands were encoded by the PNS module. The tonality of scale-factor bands derived from the FTM is of primary importance for detecting noisy bands, because q_t is just slightly higher than q_{PNS} for all cases. Although the criterion based on energy variance is usually met (>92% in average), it plays an important role when the recordings contain very expressive articulation effects as it occurred in the S2 sound sample.

9.2.2 Test method

The listening tests were performed according to the ITU-R BS.1534 recommendation defining MUSHRA (MUlti Stimulus with Hidden Reference and Anchor) procedure [74][153]. Beside the eight representations of the sound sample recordings produced by the codec operating at two bit-rate modes and four configurations, two additional excerpts were presented to the experts during listening tests: a hidden reference and 3.5 kHz anchor (sound recording degrade in predefined way e.g. band-limited version of the original sound recording). Furthermore, the reference signals were available to the subjects on request during the test. The MUSHRAM software designed for MUSHRA tests was used for excerpts playback and grading [165]. According to the ITU-R BS.1534 recommendation the procedure was comprised of the following two major phase:

1. Training phase – the subjects were able to become familiar with all the sound excerpts under test and their quality level ranges. The graphic user interface implemented in MATLAB used during training phase is presented in Fig. 9.6. The subjects were also instructed how to use the test equipment and the grading scale.
2. Blind grading phase – the sound sample recordings are scored between 0 (bad quality) and 100 (excellent quality) by the experts using sliders presented on the computer screen. At least one excerpt must be given a grade of 100 because the unprocessed reference signal is included as one of the excerpts to be graded. The graphic user interface implemented in MATLAB used during the evaluation phase is presented in Fig. 9.7.

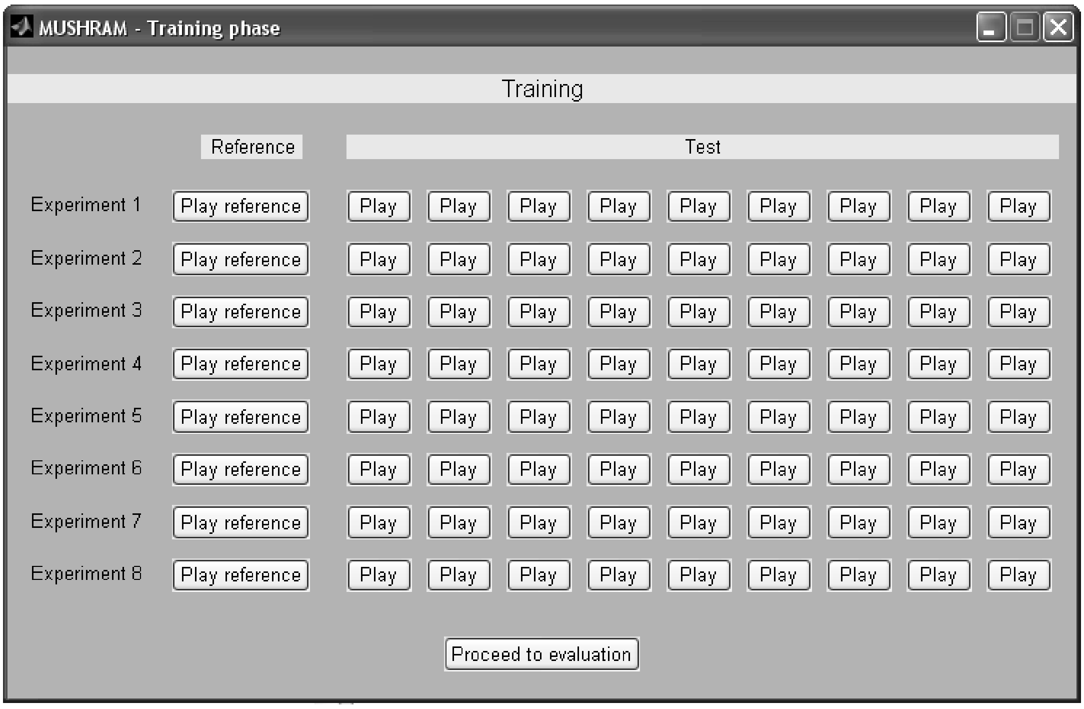
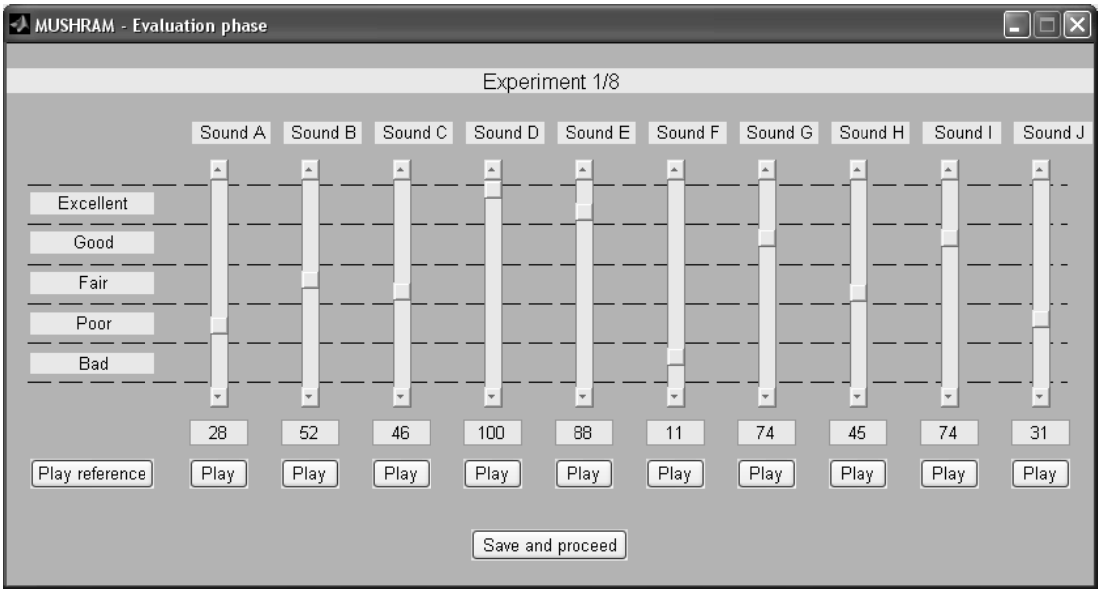


Fig. 9.6 Graphic user interface used during training phase



Sound A	Sound B	Sound C	Sound D	Sound E	Sound F	Sound G	Sound H	Sound I	Sound J
28	52	46	100	88	11	74	45	74	31

Fig. 9.7 Graphic user interface used during blind grading phase

Subjects were asked to score the excerpts roughly first and further to compare the excerpts having similar quality in order to assign to them precise scores. The external sound card with more than 100 dB SNR of digital-to-analog conversion and the AKG K240 Studio MKII headphones were used for excerpts playback.

9.2.3 *Subject pre-screening*

The group of 18 subjects of about 30 years old was considered as a listening panel. One subject was rejected, as not meeting the following requirements:

- have at least 5 years experience in sound listening in critical way [74][153],
- have normal hearing.

Consequently, the listening panel was composed of 17 experienced listeners.

9.2.4 *Subject post-screening*

In the paired comparison test, the experts reliability may be simply verified by comparing the scores given by them in two following listening test series [110]. However, in the MUSHRA test all of the sound sample recordings are scored by the subject only once. Therefore, the other method for subject post-screening must be employed. Although the BS.1534 recommendation does not define precisely the methods allowing to discard unreliable subjects, it suggests to verify their reliability basing on a comparison of scores given by the particular subject to the mean grading of all subjects [74][153]. Thus, the variance of the scores given to the evaluated sound sample recordings was calculated for all subjects as given by:

$$\sigma^2_{\mathrm{ALL}}[r_t] = \frac{1}{80}\sum_{w_t=1}^{10}\sum_{u_t=1}^{8}\left(s_{\mathrm{MSH}}[u_t, r_t, w_t] - \frac{1}{17}\sum_{r_t=1}^{17} s_{\mathrm{MSH}}[u_t, r_t, w_t]\right)^2 \qquad (9.4)$$

where w_t=1, 2, …, 10 is the index of sound sample recording being evaluated (w_t=1 for hidden reference signal), u_t=1, 2, …, 8 is the experiment index, r_t=1, 2, …, 17 is the subject index and $s_{\mathrm{MSH}}[u_t, r_t, w_t]$ is the score given to the w_t-th sound sample recording in the u_t-th experiment by the r_t-th subject.

Furthermore, it is assumed here that the reliable subjects should be able to identify the reference signal hidden between sample recordings processed by the systems being evaluated. Thus, the mean square error of the scores given to the hidden reference signal was calculated for all subjects as:

$$e_{\mathrm{HR}}[r_t] = \frac{1}{8}\sum_{u_t=1}^{8}\left(s_{\mathrm{MSH}}[u_t, r_t, 1] - 100\right)^2 \qquad (9.5)$$

The $\sigma^2_{\mathrm{ALL}}[r_t]$ and $e_{\mathrm{HR}}[r_t]$ for all subjects are presented in Fig. 9.8.

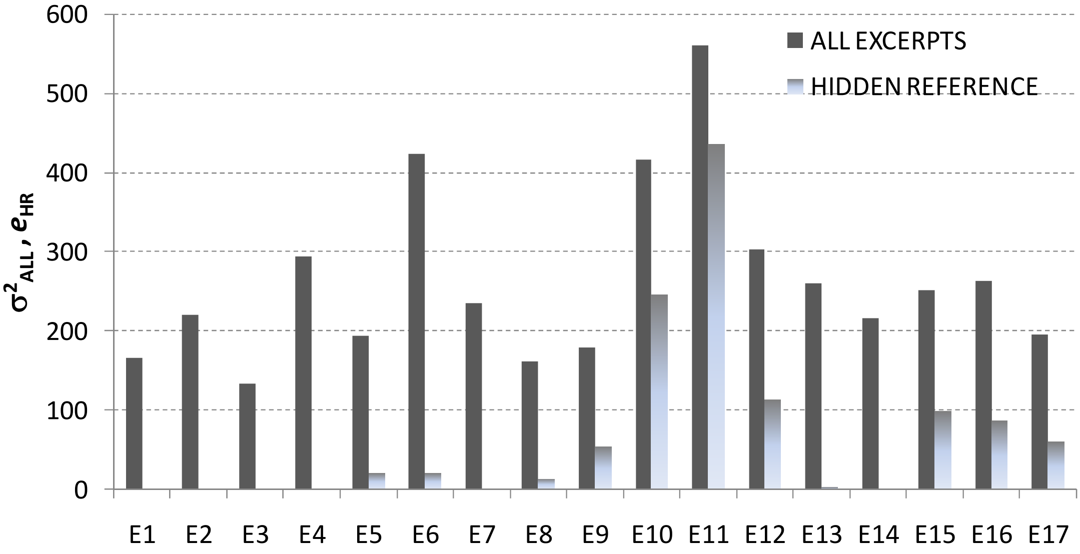


Fig. 9.8 Statistics of the subject scores

It can be noticed from Fig. 9.8 that the mean square error of the scores given to the hidden reference signal by experts E10 and E11 is relatively high comparing to the rest of the subjects. Regarding the variance, these subjects seem to be also too critical or less critical than other. Therefore, grades provided by these two experts were rejected and the final quality assessment was made based on the scores provided by the 15 subjects.

9.2.5 Analysis of results

The mean scores for all evaluated excerpts together with 95% confidence intervals are presented in Figs. 9.9 and 9.10.

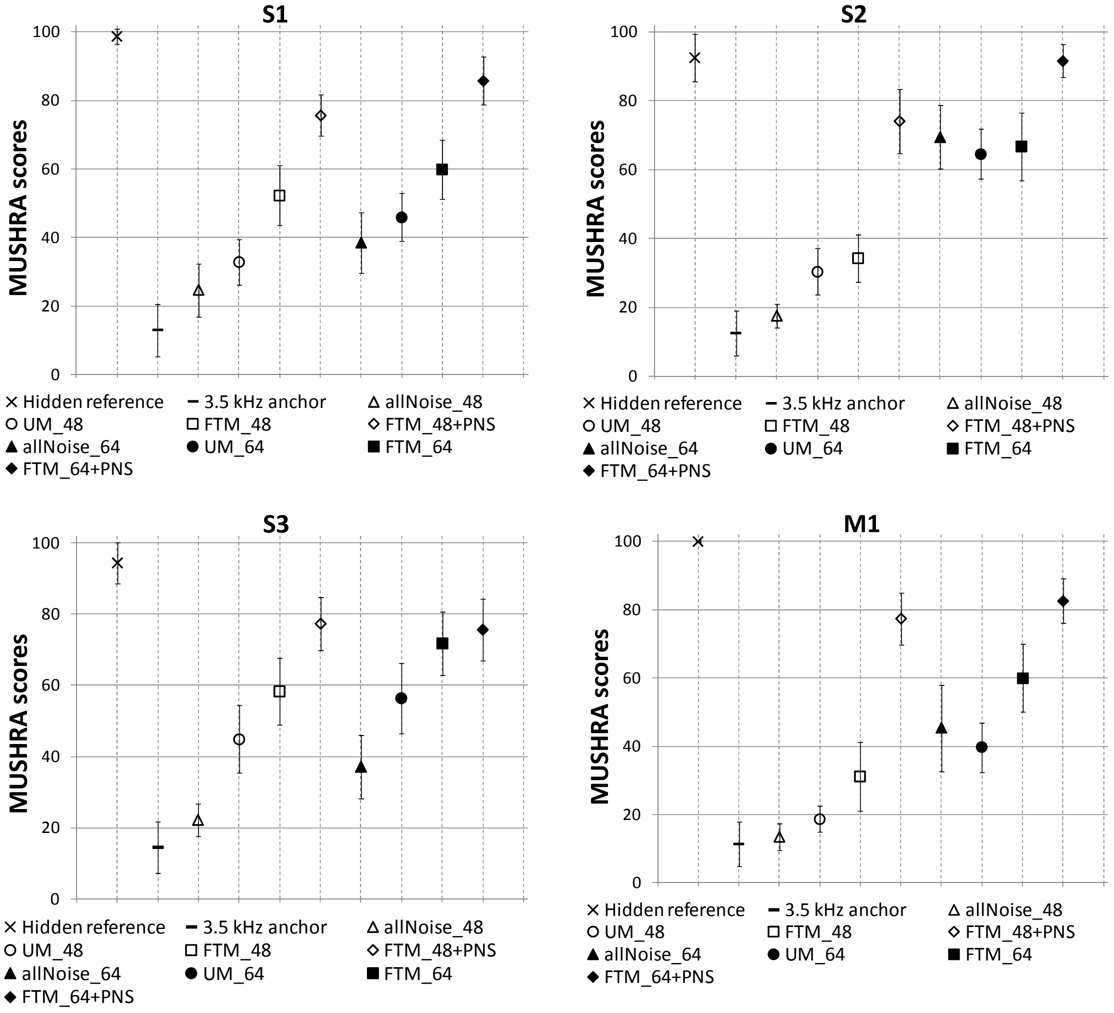


Fig. 9.9 Results of MUSHRA listening tests (S1 – S3 and M1 excerpts)

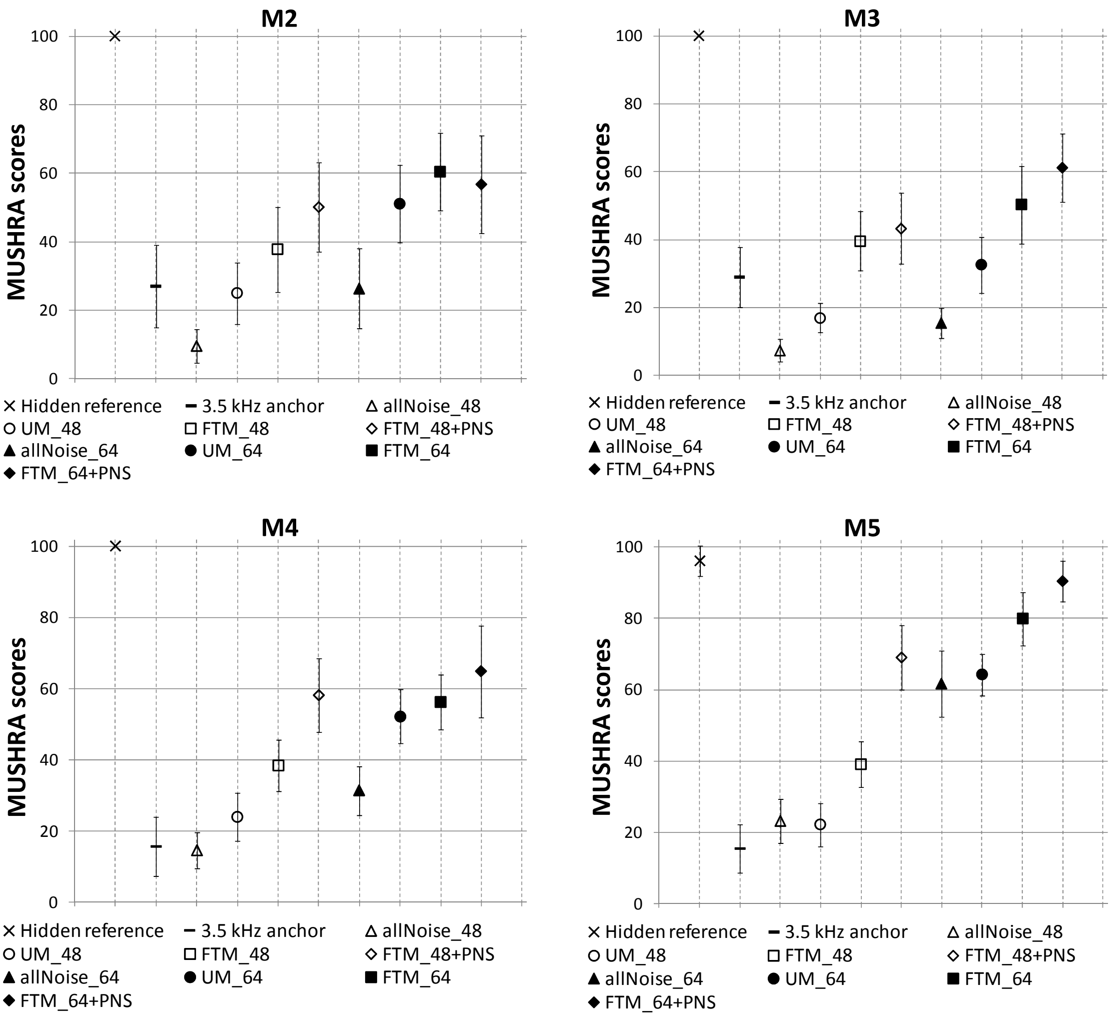


Fig. 9.10 Results of MUSHRA listening tests (M2 – M5 excerpts)

In Figs. 9.11 and 9.12 the following MUSHRA score differences for all sound samples encoded with 48 and 64 kbps rates are shown:

- difference between grades given to excerpts encoded employing psychoacoustic model with the UM and no tonality estimator (all components treated as noise),
- difference between grades given to excerpts encoded employing psychoacoustic model with the FTM and no tonality estimator (all components treated as noise),
- difference between excerpts encoded employing psychoacoustic model with the FTM and UM tonality estimators,
- difference between grades given to excerpts encoded employing psychoacoustic model and the PNS module employing the FTM tonality estimator and experts encoded with the PNS module inactive.

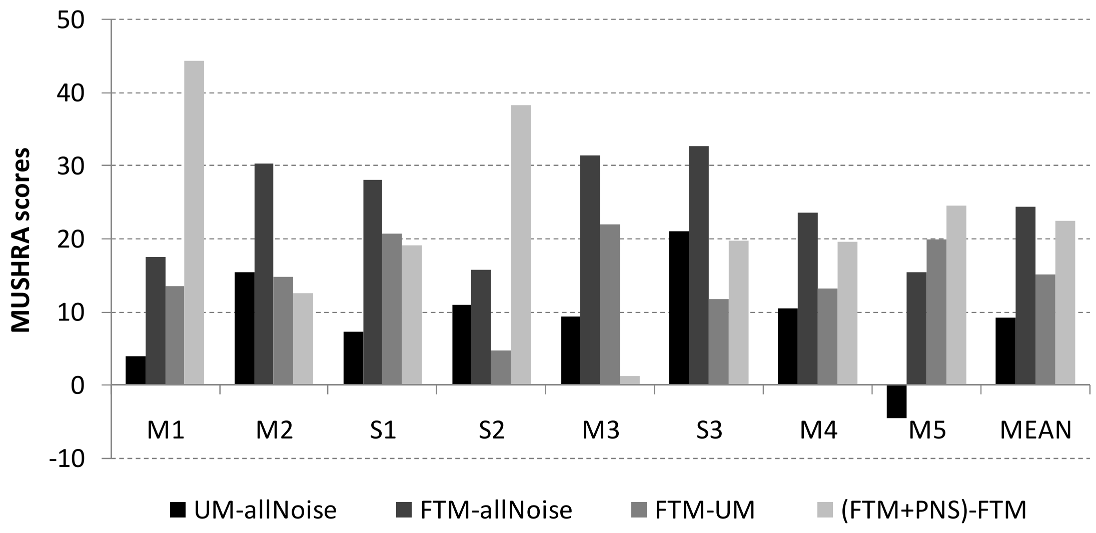


Fig. 9.11 Mushra score differences (48 kbps rate)

Considering results presented in Figs. 9.9 to 9.12 it can be noted that excerpts containing modulated tonal components, encoded using psychoacoustic model with the FTM algorithm are generally scored 10 to 20 points higher than the same sound sample recordings encoded employing tonality index derived from the UM. This tendency is viewable regardless of whether the bit-rate is on low or mid level. The difference between mean MUSHRA scores given for the sample recordings encoded with 48 kbps rate and either the FTM or UM tonality estimator is approximately equal to 15. When codec operates with 64 kbps rate the quality gain is slightly lower (about 12 MUSHRA scores). Surprisingly, also two excerpts containing mainly stationary tonal components encoded employing the UM were scored lower than when using the FTM. That was not a case for S2 excerpt, where the quality is similar to each other regardless of the method used for tonality estimation. It can be noted, that psychoacoustic model with the UM operates totally inefficiently for M1, and M5 sound sample recordings regardless of used coding bit-rate. In these cases, the scores obtained using the UM and model without any tonality estimation algorithm (all components treated as noise-like), are similar to each other. This is due to the limited ability of the UM to reliably estimate tonality of modulated components. Reversely, the quality of M1 and M5 excerpts encoded employing the FTM is significantly higher than the quality obtained with psychoacoustic model devoided of tonality estimator.

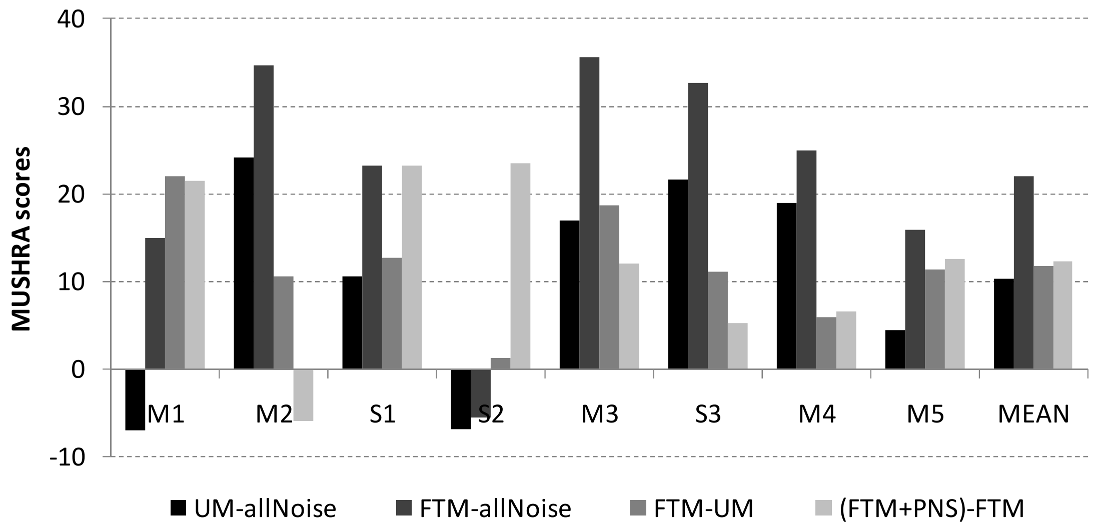


Fig. 9.12 MUSHRA score differences (64 kbps rate)

It can be noted from Figs. 9.9 and 9.10 that the encoder operating at 48 kbps rate with the FTM and PNS module enabled, provides similar or even higher MUSHRA scores comparing to the grades obtained with encoder operating at higher bit-rate. The coding quality of S2 and M5 sound sample recordings, at 64 kbps rate with the PNS module enabled, was found transparent by many experts, as they indicated this excerpt as a reference signal. Unfortunately, the quality of the M2 excerpt encoded with 64 kbps rate was found slightly lower after activating the PNS module. The suspected reason is related to the limited efficiency of detection of the noisy scale-factor bands, occupying high frequency region when the signal contains a lots of modulated tonal components. Since these components are spread over many spectra bins, they can be classified improperly. While encoding scale-factor bands containing weak tonal components with 64 kbps rate using the PNS technique can deteriorate the coding quality, it does not affect the coding quality in 48 kbps mode.

The mean quality gain obtained after activating the PNS module at 48 kbps rate is around 22 MUSHRA scores. However, it is ten scores lower for codec operating at 64 kbps rate (Figs. 9.11 and 9.12). Consequently, the PNS technique was found to be more efficient rather in low than mid bit-rate mode of codec operation.

10 CONCLUSIONS

In this dissertation the novel method for tonality estimation of spectra components for audio coding applications has been proposed and evaluated. The method is well suited to assign uncorrupted tonality estimates to spectral components corresponding to sinusoids of constant and varying frequency and/or amplitude. The FTM algorithm may be also a useful tool for extending sinusoidal modeling framework, as it incorporates a module responsible for matching the spectra components coming from contiguous frame spectra into the tonal tracks. The properties of the engineered algorithm have been investigated and compared to other methods used in the audio coding applications. Furthermore, the proposed method has been successfully combined with the psychoacoustic model and the PNS module in order to reveal its usefulness for perceptual codecs like MP3 and AAC.

Although the FTM algorithm provides continuous tonality measures falling within [0,1] range, it can operate also as a discrete classifier allowing detection of tone-like spectral components. In order to evaluate its performance as a binary classifier, the ROC curves have been generated for the FTM and other selected tonality detectors. Experimental results have indicated that the FTM outperforms other tonality detectors when the analyzed signal contains sinusoids of varying frequency. In case of audio signal containing stationary sinusoidal components, the performance of the FTM method is similar to the other ones. The characteristics depicting an influence of the amplitude and frequency modulation depth on the tonality measures yielded by the FTM and selected tonality estimation methods have been presented. Considering the results of above-mentioned experiments it can be concluded that the first thesis of the dissertation: "It is possible to estimate tonality of unmodulated or frequency-modulated sinusoidal components of audio signals through the comparison of their instantaneous frequency variations determined employing both: an estimator processing spectral amplitude samples and estimator processing spectral phase samples" has been proven true.

The FTM algorithm has been developed in order to become the substitute of tonality estimator (UM) used in the MPEG psychoacoustic model 2. However, due to the limited resolution of spectral analysis the FTM algorithm may fail when audio material contains harmonic signals of very low pitch. Therefore, it was decided to use hybrid tonality estimator in all experiments related to the audio coding. Up to experimentally chosen cut-off frequency (300 Hz) the tonality of spectral components is determined as a maximum of tonality provided by the UM and FTM algorithms. In the upper band, the tonality is estimated using only the FTM method. Furthermore, the direct substitution of the UM with FTM is impractical due to the difference in the way they assign tonality to spectral bins. While the FTM assigns high tonality only to the bins being spectra maxima, the high tonality values indicated by the UM are usually spread also over bins laying on the both sides of the peak. Therefore, two methods for the FTM spreading has been proposed. The first method is based on the spreading model built basing on the measured characteristics of the UM spreading. Because this method is computationally inefficient, the second method has been developed. This method is simpler than the first one and has been found to yield similar results to the first method.

In order to verify an influence of the FTM method on reliability of hearing threshold estimation, the modified AAC coding scheme was used. The audio signals were quantized so that the introduced quantization noise was just on the hearing threshold. In such coding scenario the introduced quantization noise should be just perceived. Two groups of sound sample recordings were used during this experiment. While the first group was composed of the recording containing mainly stationary sinusoidal components, the second one comprised of recordings containing modulated tonal components. All sound sample recordings were encoded twice. At the first attempt, the psychoacoustic model with the UM tonality estimator was used. Next, all samples were encoded using the psychoacoustic model with the FTM algorithm. The quality degradation of all sample recordings was determined basing on the scores provided by the PEAQ algorithm. It was shown that quality degradation is lower when the psychoacoustic model with the FTM algorithm is used. The quality gain obtained with the FTM is significant for recordings containing modulated tonal components.

The FTM was also used as a basis of detector for choosing signal subbands containing pure noise-like components. These bands were omitted during usual AAC quantization and filled with locally generated, weighted noise in the decoder. Although the detector is based on the FTM algorithm it examines also the flatness of subband spectrum and variation of its energy. It was shown that tonality estimates yielded by the FTM are of primary importance for detecting totally noise-like subbands. The efficiency of above-mentioned detector was examined using codec architecture based on the AAC. While all tonal subbands were quantized so that the introduced distortions were just on the hearing threshold, remaining subbands were substituted with noise by the decoder. The group of selected recordings were also encoded with the PNS module switched off. These two encoded representations of recordings were used during listening tests, performed according to the ITU-T BS. 1116 recommendation. Test results revealed that the implemented PNS module operates efficiently. The quality degradation was negligible whenever the recording contained or not modulated tonal components.

In order to fully explore the benefits of the FTM method for coding applications, the listening tests were performed in accordance with the ITU-T BS.1534 recommendation (MUSHRA). The group of selected recordings was encoded with bit-rate constrained to 48 kbps and 64 kbps by the codec operating in various coding modes. Every sound sample recording was quantized by the AAC encoder combined with the psychoacoustic model employing either the FTM or UM tonality estimator. Furthermore, in one of the codec modes, all spectral components were assumed to be noisy. The selected recordings were also encoded by the AAC-based encoder operating with the PNS module enabled. The bits saved by the PNS module were injected to the subbands containing tone-like components in order to optimize the coding quality. The experts were asked to score 10 excerpts including hidden reference and 3.5 kHz anchor signals. The results of listening tests proved that substituting UM with FTM leads to the quality increase up to 20 scores in MUSHRA scale for recordings containing modulated tonal components. The implemented PNS module based on the FTM algorithm allows lifting up the coding quality towards higher quality categories (e.g. from fair to good). Considering the results of the listening tests and other experiments carried out it can be noticed that the second thesis of dissertation: "The distortions introduced during perceptual audio coding may be effectively limited employing tonality estimation algorithm proposed in this dissertation" has been proven true. Furthermore, it should be concluded that all four research aims described in subsection 1.2 have been achieved.

The personal contributions of the author to the field of audio coding applications are:

1. the novel algorithm for tonality estimation of spectral components and the method of its integration with the MPEG psychoacoustic model,
2. the novel detector of noise-like signal bands which can be combined with the PNS module,

3. the results of analysis revealing the significance of tonality index estimation to audio coding quality.

11 ACKNOWLEDGMENTS

The author of this dissertation would like to thank his supervisor – Prof. Andrzej Czyżewski for his support and guidance in scientific issues. Without him, it would be impossible to complete this dissertation. The author would like to thank also his colleagues from Multimedia Systems Department for they valuable suggestions related to the research, and all experts being involved into the subjective tests. The research was partially funded by the Polish Ministry of Science and Higher Education within the grants No. PBZ-MNiSW-02/II/2007 and N N517 378736.

12 REFERENCES

[1] M. Abe and J. O. Smith III, “Design Criteria for Simple Sinusoidal Parameter Estimation based on Quadratic Interpolation of FFT Magnitude Peaks”, *in Proc. of 117th AES Conf.* , San Francisco, USA (October 2004).

[2] AES Staff Writer, “New Developments in Low Bit-Rate Coding”, *J. Audio Eng. Soc.*, vol. 53, No. 3, pp. 235–241 (March 2005).

[3] S. Ahmadi, M. Jelinek, “On the architecture, operation, and applications of VMR WB: The new cdma2000 wideband speech coding standard”, *IEEE Communication Magazine*, vol. 44, No. 5, pp. 74–81 (May 2006).

[4] E. Alexandre, A. Pena, M. Sobreira, “Low-complexity bit-allocation algorithm for MPEG AAC audio coders”, *IEEE Sig. Proc. Letters*, vol. 12, No. 12 (December 2005).

[5] ATSC, US. Advanced Television Systems Committee, “Digital Audio Compression (AC-3) Standard”, Doc. A/52/10 (December 1995).

[6] V. S. Babu, A.K Malot, V. M. Vijayachandran, V.M., Vinay M. K., “Transient Detection for Transform Domain Coders”, *in Proc. of 116th AES Conf.*, Berlin (May 2004).

[7] M. Bartkowiak, T. Żernicki, "Harmonic Sinusoidal+Noise Modeling of Audio Based on Multiple F0 Estimation”, *in Proc. of 125th AES Conf.*, San Francisco, USA (October 2008)

[8] K. Bauer, M. Vinton, “The choice of MPEG-4 AAC encoding parameters as a direct function of the perceptual entropy of the audio signal”, *in Proc. of 13th IEEE International Conference on Networks*, vol. 1, pp. 394–398 (2005)

[9] F. Baumgarte, “Improved audio coding using psychoacoustic model based on a cochlear filter bank”, *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 7, pp. 495–503, (October 2002).

[10] F. Baumgarte, C. Ferekidis, H. Fuchs, “A Nonlinear Psychoacoustic Model Applied to ISO/MPEG Layer 3 Coder”, *in Proc. of 99th AES Conf.,* (October 1995).

[11] F. Baumgarte and C. Faller, “Binaural Cue Coding - Part I: Psychoacoustic Fundamentals and Design Principles,” *IEEE Trans. SAP*, vol. 11, pp. 509–519 (2003).

[12] F. Baumgarte and C. Faller, "Binaural Cue Coding - Part II: Schemes and Applications," *IEEE Trans. SAP*, vol. 11, pp. 520–531 (2003).

[13] B. Bessette, R. Lefebvre, R. Salami, "Universal Speech/Audio Coding Using Hybrid ACELP/TCX Techniques", *in Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP, (March 2005).

[14] M. Betser, P. Collen, G. Richard, B. David, "Estimation of frequency for AM/FM models using the phase vocoder framework", *IEEE Trans. On Signal Processing*, vol. 56, No. 2, pp. 505–517 (2008).

[15] M. Betser, P. Collen, G. Richard, B. David, "Review and discussion on classical STFT-based frequency estimators", *in Proc. 120th AES Conf.*, Paris, France (May 2006).

[16] A. Biswas, A.C. den Brinker, "Perceptually Biased Linear Prediction", *J. Audio Eng. Soc.,* vol. 54, No. 12, pp. 1179–1188 (December 2006).

[17] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, Y. Oikawa, "ISO/IEC MPEG-2 Advanced Audio Coding", *J. Audio Eng. Soc.*, vol. 45, No. 10, pp. 789–814 (1997 October).

[18] K. Brandenburg and G. Stoll, "ISO-MPEG-1 Audio: A generic standard for coding of high quality digital audio", *J. Audio Eng. Soc.,* vol. 42, No. 10, pp. 780–792 (October 1994).

[19] K. Brandenburg and J.D. Johnston, "Second Generation Perceptual Coding: The Hybrid Coder", *in Proc. of 88th AES Conf.*, Montreaux, Switzerland (March 1990), *J. Audio Eng. Soc. (Abstract)*, vol. 38, pp. 383 (May 1990).

[20] K. Brandenburg and M. Bosi., "Overview of MPEG audio: current and future standards for low-bit-rate audio coding", *J. Audio Eng. Soc.,* vol. 45, No. 1/2, pp. 4–21 (January/February 1997).

[21] K. Brandenburg, "MP3 and AAC Explained", *in Proc. of 17th AES Int. Conf. on High Quality Audio Coding* (August 1999).

[22] K. Brandenburg, O. Kunza, A. Sugiyamab, „MPEG-4 natural audio coding", *IEEE Signal Processing: Image Communication*, vol. 15, issue 4–5, pp. 423–444 (January 2000).

[23] J. Breebaart, G. Hotho, J. Koppens, E. Schuijers, W. Oomen, S. van de Par, "Background, Concept, and Architecture for the Recent MPEG Surround Standard on Multichannel Audio Compression", *J. Audio Eng. Soc.*, vol. 55, No. 5, pp. 331–351 (May 2007).

[24] J. Breebaart, S. van de Par, A. Kohlrausch, E. Schuijers, “Parametric Coding of Stereo Audio,” *EURASIP J. Appl. Signal Process*, vol. 9, pp. 1305–1322 (2004).

[25] J.C. Brown and M.S. Puckette, “A high resolution fundamental frequency determination based on phase changes of the Fourier transform*”, J. Acoust. Soc. Am.*, vol. 94, No. 2, pp.662-667 (August 1993).

[26] D. Campbell, E. Jones, M, Glavin, “Audio quality assessment techniques – A review, and recent developments” *J. Sig. Proc.*, Elsevier, vol. 89, No. 8, pp. 1489–1500 (March 2009)

[27] F.J. Charpentier, “Pitch detection using the short-term Fourier transform”, *in Proc. of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 11, pp. 113–116, Tokyo (1986).

[28] S.B. Chon, M. Lee, H-K. Pang, K-M. Sung, “A Bit Reduction Algorithm for Spectral Band Replication Using the Masking Effect”, *in Proc. of 34th Int. Conf.: New Trends in Audio for Mobile and Handheld Devices* (August 2008).

[29] M.G. Christensen, A. Jakobsson, “Multi-Pitch Estimation”, Synthesis Lectures on Speech and Audio Processing, vol. 5, No. 1, pp. 1–160, (2009).

[30] W.C Chu, “Speech Coding Algorithms. Foundation and Evolution of Standardized Coders”, John Wiley & Sons, Hoboken (2003).

[31] C. Colomes, M. Lever, J. Rault, Y. Dehery, G. Faucon, ”A Perceptual Model Applied to Audio Bit-Rate Reduction”, *J. Audio Eng. Soc.*, vol. 43, No. 4, pp. 233–240 (April 1995).

[32] A. Czyzewski, A. Ciarkowski, A. Kaczmarek, J. Kotus, M. Kulesza, P. Maziewski, „DSP techniques for determining ‘wow’ distortion”, *J. Audio Eng. Soc.,* vol. 55, No. 4, pp. 266–284 (April 2007).

[33] A. Czyzewski, M. Dziubinski, Ł. Litwic, P. Maziewski, „Intelligent Algorithms for Movie Sound Tracks Restoration”, *Transactions on Rough Sets V*, Springer Link, vol. 4100/2006, pp. 123–145 (November 2006).

[34] A. Czyzewski, ”Dźwięk cyfrowy”, Akademicka Oficyna Wydawnicza EXIT, Warszawa (1998).

[35] Daudet L., “A Review on Techniques for the Extraction of Transients in Musical Signals”, *in proc. of CMMR'05*, Pisa (2005).

[36] M. Desainte-Catherine, S. Marchand, “High-Precision Fourier Analysis of Sounds Using Signal Derivatives”, *J. Audio Eng. Soc*, vol. 48, No. 7/8, pp. 654–667 (July/August 2000)

[37] M. Dietz, L. Liljeryd, K. Kjorling, O. Kunz, "Spectral Band Replication, a Novel Approach in Audio Coding", *in Proc. of 112th AES Conf.* (April 2002)

[38] I Dimkoviae, D. Milovanovaiae, Z. Bojkoviae, "Fast software implementation of MPEG advanced audio encoder", *in Proc. of 14th Int. Conf. on Dig. Sig. Proc.*, vol. 2, pp. 839–843 (2002).

[39] I. Dimkovic, "Improved ISO AAC encoder", PsyTEL Research, Belgrade, Yugoslavia, http://www.mp3-tech.org/programmer/docs/di042001.pdf

[40] Dolby Laboratories, http://www.dolby.com

[41] K. Dressler, "Sinusoidal extraction using an efficient implementation of a multi-resolution FFT", *in Proc. of Digital Audio Effects (DAFx) Conf.*, Montreal, Canada (September 2006).

[42] EBU Project Group B/AIM, "Possible use of Spectral Band Replication in DAB", technical report No. 3310, Geneva (December 2005)

[43] B. Edwarts and N. Viemeister, "Masking of a brief probe by sinusoidal frequency modulation" *J. Acoust. Soc. Am*, vol. 101, No. 2, pp. 1010–1018, (February 1997).

[44] A. Ehret, M. Dietz, K. Kjorling, "State-of-the-Art Audio Coding for Broadcasting and Mobile Applications", *in Proc. of 114th AES Conf.* (March 2003).

[45] "FAAC – Freeware Advanced Audio Coder", www.audiocoding.com

[46] T. Fawcett "An introduction to ROC analysis", Elsevier Pattern Recognition Letter, vol. 27, pp. 861–874, (December 2006).

[47] C. Feller, "Method and Apparatus for Detecting Noise-like Signal Components", US Patent 6,647,365 (2003 November).

[48] A. J. S. Ferreira, "Tonality Detection in Perceptual Coding of Audio", in *Proc. of 98th AES Conf.*, Paris, France (February 1995).

[49] J. A. Feummeler, R.C. Hardie, W. R. Gardner, "Techniques for the regeneration of wideband speech from narrowband speech", *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 266–274 (2001)

[50] Flac – Free Lossless Audio Codec, http://flac.sourceforge.net/format.html

[51] J. Flanagan and R. Golden, "Phase vocoder", *Bell Syst. Tech. J.*, vol. 45, pp.1493-1509 (1966).

[52] G. Fuchs, R. Lefebvre, „A Scalable CELP/Transform Coder for Low Bit Rate Speech and Audio Coding", *in Proc. of 120th AES Conf.*, Paris, France, (May 2006).

[53] R. Geiger, R. Yu, J. Herre, S. Rahardja, S-W. Kim, X. Lin, M. Schmidt, "ISO/IEC MPEG-4 High Definition Scalable Advanced Audio Coding", *J. Audio Eng. Soc.*, vol. 55, No.1/2, pp. 27–43 (January/February 2007).

[54] S. A. Gelfand, "Hearing – an introduction to psychological and physiological acoustics", Marcel Dekker, New York (1998).

[55] M.A. Gerzon, P.G. Craven, J.R. Stuart, M.J. Law, R.J. Wilson, "The MLP Lossless Compression System for PCM Audio", *J. Audio Eng. Soc.*, vol. 52, No. 3, pp. 243–260 (March 2004).

[56] Goldberg R., Riek L., "A Practical Handbook of Speech Coders", CRC Press, Boca Raton (2000).

[57] GPSYCHO - A LGPL'd Psycho-Acoustic Model, http://lame.sourceforge.net/gpsycho.php

[58] R. P. Hellman, "Asymmetry of masking between tone and noise", Perception and Psy-choacoustics, vol. 11, pp. 241–246 (1972).

[59] O. Hellmuth, E. Allamanche, J. Herre, T. Kastner, M. Cermer, W. Hirsch, "Advanced audio identification using MPEG-7 content description", *in Proc. of 111th AES Conf.*, New York, USA (2001).

[60] J. Herre, "Temporal noise shaping, quantization and coding methods in perceptual audio coding: a tutorial introduction", *in Proc. of 17th AES Conf.* (August 1999)

[61] J. Herre and D. Schulz, "Extending MPEG-4 AAC Codec by Perceptual Noise Substitution", *in Proc.* of *104th AES Conf.* Amsterdam, Netherlands (May 1998).

[62] J. Herre, E. Eberline, B. Grill, K. Brandenburg, H. Gerhauser, "Method and Device for Determining the Tonality of Audio Signal", US Patent 5,918,203 (June 1999).

[63] J. Herre, E. Eberline, K. Brandenburg, "A Real Time Perceptual Threshold Simulator", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (October 1991).

[64] J. Herre, K. Kjorgling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Roden, W. Oomen, K. Linzmeier, K. S. Chong, "MPEG Surround—The ISO/MPEG Standard for Efficient and Compatible Multichannel Audio Coding", *J. Audio Eng. Soc.*, vol. 56, No.11, pp. 932–955 (November 2008).

[65] L. Humes, W. Jesteadt, "Models of the Additivity of Masking", *J. Acoust. Soc. Am.*, vol. 85, No. 3, pp. 1285-1294 (March 1989).

[66] ISO/IEC 11172-3, Coding of Moving Pictures and Associated Audio for Digital Storage Media up to 1.5 Mbit/s, Part 3: Audio (1992).

[67] ISO/IEC IS11172-3, Coding of Moving Pictures and Associated Audio for Digital Storage Media up to 1.5 Mbit/s, Part 3: Audio, Annex D (1992).

[68] ISO/IEC 13818-3 Information technology — Generic coding of moving pictures and associated audio information, Part 3: Audio, 2nd edition (April 1998).

[69] ISO/IEC FDIS 15938-4:2001, Information Technology – Multimedia Content Description Interface – Part 4: Audio, (June 2001)

[70] ISO/IEC 13818-7 Information technology — Generic coding of moving pictures and associated audio information, Part 7: Advanced Audio Coding (AAC), 4th edition (January 2006).

[71] ISO/IEC 14496-3 Information technology — Coding of audio-visual objects, Part 3: Audio, 3rd edition (December 2005).

[72] ITU-R Recommendation BS.1116-1: Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, ITU-R (October 1997).

[73] ITU-R Recommendation BS.1387-1: Methods for objective measurement of perceived audio quality, ITU-R (2001).

[74] ITU-R Recommendation BS.1534-1: Method for subjective assessment of intermediate quality level of coding systems, ITU-R (2003).

[75] ITU-T Recommendation G.711: Pulse code modulation (PCM) of voice frequencies, ITU-T (1988).

[76] ITU-T Recommendation G.722: 7 kHz audio-coding within 64 kbit/s, ITU-T (1988).

[77] ITU-T Recommendation G.722.2: Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB), ITU-T (June 2003).

[78] ITU-T Recommendation G.723.1: Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s, ITU-T (1996).

[79] ITU-T Recommendation G.726: 40, 32, 24, 16 kbit/s adaptive differential pulse code modulation (ADPCM), ITU-T (1990).

[80] ITU-T Recommendation G.727: 5-, 4-, 3- and 2-bits/sample embedded adaptive differential pulse code modulation (ADPCM), ITU-T (1990).

[81] ITU-T Recommendation G.728: Coding of speech at 16 kbit/s using low-delay code excited linear prediction (LD-CELP), ITU-T (1992).

[82] ITU-T Recommendation G.729: Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP), ITU-T (1996).

[83] P. J. B Jackson and C. H Shadle, “Pitch-Scaled Estimation of Simultaneously Voiced and Turbulence-Noise Components in Speech”, *IEEE Trans. on Speech and Audio Processing*, vol.9, pp. 713–726 (October 2001).

[84] H. Jeong and J. Ih, “Implementation of a New Algorithm using the STFT with Variable Frequency Resolution for the Time-Frequency Auditory Model”, *J. Audio Eng. Soc.*, vol. 47, pp. 240–251 (April 1999).

[85] W. Jesteadt, S. Bacon, J. Lehman, ”Forward Masking as a Function of Frequency, Masker Level, and Signal Delay”, *J. Acoust. Soc. Am.*, vol. 71, No. 4, pp. 950–962 (April 1982).

[86] J. D. Johnston, “Transform Coding of Audio Signals Using Perceptual Noise Criteria”, *IEEE J. on Selected Areas in Comm.*, vol. 6, pp. 314–323 (February 1988).

[87] J. D. Johnston, “Estimation of perceptual entropy using noise masking criteria”, *in Proc. of International Conference on Acoustics, Speech, and Signal Processing –* ICASSP, vol. 5, pp. 2524–2527 (1988).

[88] F. Keiler, S. Marchand, “Survey on extraction of sinusoids in stationary sound”, *in Proc. of the 5th Int. Conf. on Digital Audio Effects (DAFx-02)*, Hamburg, Germany (2002).

[89] W.B. Kleijn, K.K. Paliwal, Speech Coding and Synthesis, Elsevier, Netherlands (1995).

[90] E. Knapen, D. Reefman, E. Janssen, F. Bruekers, “Lossless Compression of 1-Bit Audio”, *J. Audio Eng. Soc.*, vol. 52, No. 3, pp.190–199 (March 2004).

[91] B. Kostek, “Perception-Based Data Processing in Acoustics. Applications to Music Information Retrieval and Psychophysiology of Hearing”, Springer Verlag, Series on Cognitive Technologies, Berlin, Heidelberg, New York (2005).

[92] G. Kubin, W. Bastiaan Kleijn, „On speech coding in perceptual domain”, *in Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1 (1999).

[93] M. Kulesza and A. Czyzewski, “Speech codec enhancements utilizing time compression and perceptual coding”, *in Proc. of 122nd AES Conf.*, Vienna, Austria (May 2007).

[94] M. Kulesza, A. Czyzewski, “Audio codec employing frequency-derived tonality measure”, *in Proc. of 127th AES Conf.*, New York, USA (October 2009).

[95] M. Kulesza, A. Czyzewski, “Novel approaches to wideband speech coding”. *GESTS Int. Trans. On Computer Science and Engineering*, vol. 44, No. 1, pp. 154-165 (2008).

[96] M. Kulesza, G. Szwoch, A. Czyzewski, „A hybrid speech codec employing parametric and perceptual coding techniques”, *in Proc. 121st AES Conf.*, San Francisco, USA (October 2006).

[97] M. Kulesza, A. Czyzewski, “Tonality Estimation and Frequency Tracking of Modulated Tonal Components”, *J. Audio Eng. Soc.*, vol. 57, No. 4, pp. 221–236 (April 2009).

[98] M. Kulesza, A. Czyzewski, „Frequency based criterion for distinguishing tonal and noisy spectral components”, *Int. J. of Computer Science and Security*, vol. 4, No. 1, pp. 1-16 (March 2010).

[99] M. Kulesza, L. Litwic, G. Szwoch, A. Czyzewski, “High quality speech codec employing sines+noise+transients model”, *Archives of Acoustics*, vol. 31, No. 4 (Supplement), pp. 183-188 (2006).

[100] E. Kurniawati, C. T. Lau, B. Premkumar, J. Absar, S. George, “New Implementation Techniques of an Efficient MPEG Advanced Audio Coder”, *IEEE Transactions on Computer Electronics*, vol. 50, No. 2, pages 655–665 (2004).

[101] E. Kurniawati, J. Absar, S. George, L.T Lau, B. Premkumar, “An investigation into different masking behaviors resulting from estimation of tonality index”, *in Proc. of 14th IEEE Int. Conf. on Dig. Sig. Proc.*, vol. 2, pp. 1035–1038 (2002).

[102] M.D. Kwong, R. Lefebvre, “Transient detection of audio signals based on an adaptive comb filter in the frequency domain” *in Proc. 37th Asilomar Conference on Signals, Systems and Computers*, Asilomar, Pacific Grove (November 2003).

[103] M. Lagrange, S. Marchand, J.B. Rault, “Sinusoidal Parameter Extraction and Component Selection in Non Stationary Model”, *in Proc. of Digital Audio Effects (DAFx) Conf.*, Hamburg, Germany (September 2002).

[104] M. Lagrange and S. Marchand, “Estimating the Instantaneous Frequency of Sinusoidal Components Using Phase-Based Methods”, *J. Audio Eng. Soc.*, vol. 55, pp. 385–399 (May 2007).

[105] M. Lahdekorpi, J. Nurminen, A. Heikkinen, J. Saarinen, “Perceptual irrelevance removal in narrowband speech coding”, *in Proc. of 8th European Conference on Speech Communication and Technology,* Geneva, Switzerland (September 2003).

[106] S.W. Lang, B.R Musicus, "Frequency estimation from phase difference", *in Proc. of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. pp. 2140–2143, United Kingdom (1989).

[107] M.C. Lavoie, G.A. Soulodre, "Subjective Evaluation of MPEG Layer II with Spectral Band Replication", *in Proc. of 117th AES Conf.*, San Francisco, USA (October 2004).

[108] R.R. Lawrence, R.W. Schafer, "Introduction to digital speech processing", *Fundations and Trends in Signal Processing*, No. 1-2, pp. 1-194 (2007).

[109] K. Lee, K. Yeon, Y. Park, D. Youn, "Effective Tonality Detection Algorithm Based on Spectrum Energy in Perceptual Audio Coder", *in Proc. of 117th AES Conf.*, San Francisco, USA (October 2004).

[110] T. Letowski, "Słuchowa ocena sygnałów i urządzeń", Akademia Muzyczna im. Fryderyka Chopina w Warszawie, Warsaw (1984).

[111] S. Levine and J. O. Smith III., "Improvements to the Switched Parametric & Transform Audio Coder", *in Proc. IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, New York, USA (October 1999).

[112] Z-N. Li, M.S Drew, "Fundamentals of multimedia", Prentice Hall (2004).

[113] S.P. Lipshitz and J. Vanderkooy, „Pulse-Code Modulation – An Overview", *J. Audio Eng. Soc.*, vol. 52, No. 3, pp. 200–215 (March 2004)

[114] R.C. Maher, "Control of synthesized vibrato during portamento musical pitch transitions", *J. Audio Eng. Soc.*, vol. 56, pp. 18–27 (January/February 2008).

[115] R. Martin, U. Heute, C. Antweiler, "Advances in Digital Speech Transmission", John Wiley & Sons (2008).

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

[117] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation", *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 34, pp. 744–754 (August 1986).

[118] R.E. Millman, C. Lorenzi, F. Apoux, C. Fullgrabe, G.R. Green, S.P. Beacon, "Effect of duration on amplitude-modulation masking", *J. Acoust. Soc. Am*, vol. 111, No. 6, pp. 2551–2554 (June 2002).

[119] B. C. J. Moore and J. I. Alcantara, "Masking patterns for sinusoidal and narrow-band noise maskers", *J. Acoust. Soc. Am*, vol. 104, No. 2, pp. 1023–1038 (August 1998).

[120] H. Najaf-Zadeh, H. Lahdili, L. Thibault, "Incorporation of inharmonicity effects into auditory masking model", *in Proc. of 113th AES Conf.*, Los Angeles, USA (October 2002).

[121] H. Najaf-Zadeh, H. Lahdili, M. Lavoie, L. Thibault, "Use of auditory temporal masking in the MPEG psychoacoustic model 2", *in Proc. of 114th AES Conf.*, Amsterdam, The Netherlands (March 2003).

[122] M. Neuendorf, P. Gournay, M. Multrus, J. Lecomte, B. Bessette, R. Geiger, S. Bayer, G. Fuchs, J. Hilpert, N. Rettelbach, F. Nagel, J. Robilliard, R. Salami, G. Schuller, R. Lefebvre, B. Grill, "A Novel Scheme for Low Bitrate Unified Speech and Audio Coding - MPEG RM0", *in Proc. of 126th AES Conf.*, Munich, Germany (May 2009)

[123] P. Noll, "MPEG Digital Audio Coding", *IEEE Signal Processing Magazine*, pp. 59–81 (September 1997).

[124] P. Noll, "Wideband Speech and Audio Coding," *IEEE Comm. Mag.*, vol. 31, pp. 34-44 (November 1993)

[125] OGG Vorbis Specification: http://xiph.org/vorbis/

[126] P. Ojala, A. Lakaniemi, H. Lephanaho, M. Joakimies, "The adaptive multirate wideband speech codec: system characteristics, quality advances, and deployment strategies", *IEEE Communication Magazine*, vol. 44, No. 5, pp. 59–65 (May 2006).

[127] J. Ojanpera, "Noise Detection for Audio Encoding by Mean and Variance Energy Ratio", US Patent 7,457,747 (2008 November).

[128] A. V. Oppenheim, R. W. Schafer and J. R. Buck, "Discrete-Time Signal Processing", 2nd ed., Prentice-Hall International, Upper Saddle River, New Jersey (1999).

[129] Opticom, Opera your digital ear, User manual, version 3.5 (2002).

[130] T. Painter and A. Spanias, "Perceptual Coding of Digital Audio", *in Proc. of IEEE*, vol. 88, pp. 451–513 (April 2000).

[131] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, S. H. Jensen, "A Perceptual Model for Sinusoidal Audio Coding Based on Spectral Integration," *EURASIP J. Appl. Signal Process.*, No. 9, pp. 1292–1304 (June 2005).

[132] S. van de Par, A. Kohlrausch, G. Charestan, R. Heusdens, "A New Psychoacoustical Masking Model for Audio Coding Applications", *in Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1805–1808, Orlando, USA (May 2002).

[133] G. Peeters, X. Rodet, “Signal Characterization in Terms of Sinusoidal and Non-Sinusoidal Components”, *in proc. of Digital Audio Effects (DAFx) Conf.*, Barcelona, Spain (November 1998).

[134] G. Peeters, X. Rodet, “SINOLA: A New Analysis/Synthesis Method using Spectrum Peak Shape Distortion, Phase and Reassigned Spectrum”, *in Proc. of the Int. Computer Music Conf.*, Beijing, China (October 1999).

[135] Public MP3 listening tests, http://www.listening-tests.info/mp3-128-1/

[136] M.S. Puckette and J.C. Brown, “Accuracy of frequency estimates using the phase vocoder”, *IEEE Trans. On Speech and Audio Processing*, vol. 6, No. 2, pp. 166–176 (1998).

[137] S. Ragot, B. Kovesi, D. Virette, R. Trilling, D. Massaloux, “A 8-32 kbps scalable wideband speech and audio coding candidate for ITU-T G729EV standardization”, *IEEE International Conference on Acoustic, Speech, and Signal Processing - ICASSP*, Toulouse (May 2006).

[138] D. C. Rife and R. R. Boorstyn “Single-Tone Parameter Estimation from Discrete-Time Observations”, *IEEE Trans. Info. Theory*, vol. 20, No. 5, pp. 591-598 (September 1974).

[139] C. Ritz, K. Adistambha, J. Lukasiak, I. Burnett, “A coodebook-based cascade coder for embedded lossless audio coding”, *in Proc. of 120th AES Conf.*, Paris, France (May 2006).

[140] C. Ritz., “Lossless wideband speech coding”, *in Proc. of 10th Int. Conf. on Speech Science and Technology*, Sydney, Australia (December 2004).

[141] X. Rodet, “Musical sound signal analysis/synthesis: sinusoidal+residual and elementary waveform models”, *in Proc. of IEEE Time-Frequency and Time-Scale Workshop*, Coventry, Grande Bretagne (1997).

[142] S-U. Ryu and K. Rose, “Enhanced Accuracy of the Tonality Measure and Control Parameter Extraction Modules in MPEG-4 HE-AAC”, *in Proc. of 119th AES Conf.*, New York, USA (October 2005).

[143] R. Salami, C. Laflamme, J-P. Adoul, A. Kataoka, S. Hayashi, T. Moriya, C. Lamblin, D. Massaloux, S. Proust, P. Kroon, Y. Shoham, “Design and description of CS-CELP: a toll quality 8 kb/s speech coder”, *IEEE Transactions on Speech and Audio Processing*, vol. 6, No. 2, pp. 116–130 (March 1998).

[144] R. Salami, R. Lefebvre, A. Lakaniemi, K. Kontola, A. Bruhn, A. Taleb, “Extended AMR-WB for high quality audio on mobile devices”, *IEEE Communication Magazine*, vol. 44, No.5, pp. 90–97 (May 2006).

[145] Samsudin, Boon P., Kurniawati E., Sattar F., Sapna G., “A Unified Transient Detector for Enhanced aacPlus Encoder”, *in Proc. of 120th AES Conf.*, Paris, France (May 2006).

[146] M. Schug, A. Groschel, M. Beer, F. Henn, “Enhancing Audio Coding Efficiency of MPEG Layer-2 with Spectral Band Replication (SBR) for DigitalRadio (EUREKA 147/DAB) in a Backwards Compatible Way”, *in Proc. of 114th AES Conf.*, Amsterdam, The Netherlands (March 2003).

[147] G. D. Schuller, B. Yu, D. Huang, B. Edler “Perceptual audio coding using adaptive pre- and post-filters and lossless compression”, *IEEE Transactions on Speech and Audio Processing*, vol. 10, No. 6 (September 2002).

[148] D. Schulz, “Improving Audio Codecs by Noise Substitution”, *J. Audio Eng. Soc.*, vol. 44, pp. 593–598 (July/August 1996).

[149] X. Serra and J. O. Smith III, “Spectral modelling synthesis: A sound analysis/ synthesis system based upon a deterministic plus stochastic decomposition”, *Computer Music Journal*, vol. 14, No. 4, pp. 12–24 (1990).

[150] X. Serra, “A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition”, Ph.D. Thesis, Stanford University, USA (1989).

[151] J. O. Smith III and X. Serra, “PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation”, *in Proc. of the Int. Comp. Music Conf.*, Tokyo, Japan (1987).

[152] M.J. Smithers and M.C. Feller, “Increased efficiency MPEG-2 AAC encoding”, *in Proc. of 111th AES Conf.* (September 2001).

[153] G. Stoll, F. Kozamernik, “EBU listening test on internet audio codecs”, EBU technical review (2002).

[154] J. Sunberg, “The Science of the Singing Voice”, Northern Illinois University Press (1987).

[155] E. Terhardt, G. Stoll, M. Seewann, ”Algorithm for Extraction of Pitch and Pitch Salience from Complex Tonal Signals”, *J. Acoust. Soc. Am*, vol. 71, No. 3, pp. 679-688 (March 1982).

[156] "Sound Quality Assessment Material for Subjective Tests", Technical Center of the European Broadcast Union, Tech. 3253-E (April 1988)

[157] T. Thiede, W.C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J.G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, B. Feiten, "PEAQ – The ITU Standard for Objective Measurement of Audio Quality", *J. Audio Eng. Soc.*, vol. 48, No. 1/2, pp. 3–29 (January/February 2000).

[158] Third Generation Partnership Project 2 Group, "Source-Controlled Variable-Rate Multimode Wideband Speech Codec (VMR-WB)", 3GPP2 C.S0052-0, ver. 1.0 (June 2004).

[159] Third Generation Partnership Project Group, "Extended Adaptive Multi-Rate - Wideband (AMR-WB+) codec", 3GPP TS 26.290 ver. 6.3.0, release 6 (June 2005).

[160] W.C. Treurniet and G.A. Soulodre, "Evaluation of the ITU-R Objective Audio Quality Measurement Method", *J. Audio Eng. Soc.*, vol. 48, No. 3, pp. 164–173 (March 2000).

[161] J. M. Valin, C. Montgomery, "Improved noise weighting in CELP coding of speech – applying the Vorbis psychoacoustic model to Speex", *in Proc. of 120th AES Conf.*, Paris, France (May 2006).

[162] T. Verma and T. Meng, "Time scale modification using a sines+transients+noise signal model", *in Proc. of Digital Audio Effects (DAFx) Conf.*, Barcelona, Spain (November 1998).

[163] T. Verma, T. Meng, "A flexible analysis / synthesis tool for transient signals", *J. Acoust. Soc. Am.*, vol. 103, No. 5, pp. 2756-2756 (May 1998).

[164] M. Vilermo, S. Streich, M. Vaananen, K. Linzmeier, B. Grill, Y. Wang, "Perceptual optimization of the frequency selective switch in scalable audio coding", *in Proc. of 114th AES Conf.* (March 2003).

[165] E. Vincent, "MUSHRAM 1.0 – User Guide", Center for Digital Music, Queen Mary, University of London (November 2005).

[166] Y. Wang and R. Kumaresan, "Real time decomposition of speech into modulated components", *J. Acoust. Soc. Am.*, vol. 119, No. 6, pp. EL68–EL73 (June 2006).

[167] M. Werner, G. Schuller, "An Enhanced SBR Tool for Low-Delay Applications", *in Proc. of 127th AES Conf.*, New York, USA (October 2009).

[168] M. Wilde, "Audio for Games", *J. Audio Eng. Soc.*, vol. 50, No. 5, pp. 392–396 (May 2002).

[169] M. Yang, “Low bit rate speech coding”, *IEEE Potentials*, vol. 23, No. 4, pp. 32-36 (2004).

[170] B. Yegnanarayana, C. Alessandro, V. Darisonos, “An Iterative Algorithm for Decomposition of Speech Signals into Periodic and Aperiodic Components”, *IEEE Trans. on Speech and Audio Proc.*, vol. 6, pp. 1–11 (January 1998).

[171] C-H. Yu and S. D. You, “Comparison of Two Different Approaches to Detect Perceptual Noise for MPEG-4 AAC”, Lecture Notes in Computer Science, vol. 3332/2005, pp. 505–512, Springer (2005).

[172] N. Zacharov, “A rapid listening test environment – helping managers make a better decision”, *in Proc. of 122nd AES Conf.*, Vienna, Austria (May 2007).

[173] U. Zolzer, “DAFX Digital Audio Effects”, John Wiley & Sons, United Kingdom (2002).

[174] E. Zwicker and H. Fastl, “Psychoacoustics: Facts and Models”, 3rd edition, Springer (2007).

[175] T. Żernicki and M. Bartkowiak ,"Audio bandwidth extension by frequency scaling of sinusoidal components, *in Proc. of 125th AES Conf.*, San Francisco, USA (October 2008).

13 APPENDICES

13.1 APPENDIX 1 – FORM USED DURING BS.1116 LISTENING TEST

There are eleven triples of sound sample recordings. In every triple, one sound recording is a reference and is denoted as "REF". Two more recordings are denoted as "A" and "B". Either excerpt "A" or "B" in every triple is the hidden reference signal. The excerpts are scored using continuous scale (for instance the 4.7 score may be given):

Grade	Impairment
5.0	Imperceptible
4.0	Perceptible, but not annoying
3.0	Slightly annoying
2.0	Annoying
1.0	Very annoying

An expert is asked to follow the instructions for every triple of sound recordings:

1. Among "A" and "B" excerpts select one which is the reference signal and assign 5.0 score to it;
2. Assign appropriate score to the remaining excerpt ("A" or "B" - not being the reference signal)

Test number	A	B
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		

Caution! *Scoring scale is continuous – you may assign 4.8 score to sample recording. The excerpts may be played back as many times as required to make a decision regarding quality.*

13.2 APPENDIX 2 – CD WITH SOUND SAMPLE RECORDINGS USED DURING PERCEPTUAL EVALUATION

13.2.1 PEAQ tests (subsection 7.2)

EXCERPT	REFERENCE (FILE NAME)	UM (FILE NAME)	FTM (FILE NAME)
1M	1M_REF.wav	1M_UM.wav	1M_FTM.wav
2M	2M_REF.wav	2M_UM.wav	2M_FTM.wav
3M	3M_REF.wav	3M_UM.wav	3M_FTM.wav
4M	4M_REF.wav	4M_UM.wav	4M_FTM.wav
5M	5M_REF.wav	5M_UM.wav	5M_FTM.wav
6M	6M_REF.wav	6M_UM.wav	6M_FTM.wav
7M	7M_REF.wav	7M_UM.wav	7M_FTM.wav
8M	8M_REF.wav	8M_UM.wav	8M_FTM.wav
9M	9M_REF.wav	9M_UM.wav	9M_FTM.wav
10M	10M_REF.wav	10M_UM.wav	10M_FTM.wav
11M	11M_REF.wav	11M_UM.wav	11M_FTM.wav
12M	12M_REF.wav	12M_UM.wav	12M_FTM.wav
13M	13M_REF.wav	13M_UM.wav	13M_FTM.wav
14M	14M_REF.wav	14M_UM.wav	14M_FTM.wav
1S	1S_REF.wav	1S_UM.wav	1S_FTM.wav
2S	2S_REF.wav	2S_UM.wav	2S_FTM.wav
3S	3S_REF.wav	3S_UM.wav	3S_FTM.wav
4S	4S_REF.wav	4S_UM.wav	4S_FTM.wav
5S	5S_REF.wav	5S_UM.wav	5S_FTM.wav
6S	6S_REF.wav	6S_UM.wav	6S_FTM.wav
7S	7S_REF.wav	7S_UM.wav	7S_FTM.wav
8S	8S_REF.wav	8S_UM.wav	8S_FTM.wav
9S	9S_REF.wav	9S_UM.wav	9S_FTM.wav
10S	10S_REF.wav	10S_UM.wav	10S_FTM.wav
11S	11S_REF.wav	11S_UM.wav	11S_FTM.wav
12S	12S_REF.wav	12S_UM.wav	12S_FTM.wav

13.2.2 BS.1116 subjective tests (subsection 8.2.2)

In every test either excerpt A or B was the hidden reference signal. The hidden reference signal is highlighted using grey background in the table below.

TEST NUM.	**REFERENCE (FILE NAME)**	**A (FILE NAME)**	**B (FILE NAME)**
1	12M_REF.wav	12M_A.wav	12M_B.wav
2	13M_REF.wav	13M_A.wav	13M_B.wav
3	2M_REF.wav	2M_A.wav	2M_B.wav
4	9S_REF.wav	9S_A.wav	9S_B.wav
5	7M_REF.wav	7M_A.wav	7M_B.wav
6	7S_REF.wav	7S_A.wav	7S_B.wav
7	3S_REF.wav	3S_A.wav	3S_B.wav
8	9M_REF.wav	9M_A.wav	9M_B.wav
9	8S_REF.wav	8S_A.wav	8S_B.wav
10	1S_REF.wav	1S_A.wav	1S_B.wav
11	4S_REF.wav	4S_A.wav	4S_B.wav

13.2.3 BS. 1534 subjective tests (subsection 9.2)

M1	
EXCERPT	**FILE NAME**
Reference	M1_REF.wav
3.5 kHz anchor	M1_3.5kHz_ANCHOR.wav
allNoise_48	M1_ALL_NOISE_48.wav
UM_48	M1_UM_48.wav
FTM_48	M1_FTM_48.wav
FTM_48+PNS	M1_FTM_48+PNS.wav
allNoise_64	M1_ALL_NOISE_64.wav
UM_64	M1_UM_64.wav
FTM_64	M1_FTM_64.wav
FTM_64+PNS	M1_FTM_64+PNS.wav
M2	
EXCERPT	**FILE NAME**
Reference	M2_REF.wav
3.5 kHz anchor	M2_3.5kHz_ANCHOR.wav
allNoise_48	M2_ALL_NOISE_48.wav
UM_48	M2_UM_48.wav
FTM_48	M2_FTM_48.wav
FTM_48+PNS	M2_FTM_48+PNS.wav
allNoise_64	M2_ALL_NOISE_64.wav
UM_64	M2_UM_64.wav
FTM_64	M2_FTM_64.wav
FTM_64+PNS	M2_FTM_64+PNS.wav
M3	
EXCERPT	**FILE NAME**
Reference	M3_REF.wav
3.5 kHz anchor	M3_3.5kHz_ANCHOR.wav
allNoise_48	M3_ALL_NOISE_48.wav
UM_48	M3_UM_48.wav
FTM_48	M3_FTM_48.wav
FTM_48+PNS	M3_FTM_48+PNS.wav
allNoise_64	M3_ALL_NOISE_64.wav
UM_64	M3_UM_64.wav
FTM_64	M3_FTM_64.wav
FTM_64+PNS	M3_FTM_64+PNS.wav

M4	
EXCERPT	**FILE NAME**
Reference	M4_REF.wav
3.5 kHz anchor	M4_3.5kHz_ANCHOR.wav
allNoise_48	M4_ALL_NOISE_48.wav
UM_48	M4_UM_48.wav
FTM_48	M4_FTM_48.wav
FTM_48+PNS	M4_FTM_48+PNS.wav
allNoise_64	M4_ALL_NOISE_64.wav
UM_64	M4_UM_64.wav
FTM_64	M4_FTM_64.wav
FTM_64+PNS	M4_FTM_64+PNS.wav
M5	
EXCERPT	**FILE NAME**
Reference	M5_REF.wav
3.5 kHz anchor	M5_3.5kHz_ANCHOR.wav
allNoise_48	M5_ALL_NOISE_48.wav
UM_48	M5_UM_48.wav
FTM_48	M5_FTM_48.wav
FTM_48+PNS	M5_FTM_48+PNS.wav
allNoise_64	M5_ALL_NOISE_64.wav
UM_64	M5_UM_64.wav
FTM_64	M5_FTM_64.wav
FTM_64+PNS	M5_FTM_64+PNS.wav
S1	
EXCERPT	**FILE NAME**
Reference	S1_REF.wav
3.5 kHz anchor	S1_3.5kHz_ANCHOR.wav
allNoise_48	S1_ALL_NOISE_48.wav
UM_48	S1_UM_48.wav
FTM_48	S1_FTM_48.wav
FTM_48+PNS	S1_FTM_48+PNS.wav
allNoise_64	S1_ALL_NOISE_64.wav
UM_64	S1_UM_64.wav
FTM_64	S1_FTM_64.wav
FTM_64+PNS	S1_FTM_64+PNS.wav

S2	
EXCERPT	**FILE NAME**
Reference	S2_REF.wav
3.5 kHz anchor	S2_3.5kHz_ANCHOR.wav
allNoise_48	S2_ALL_NOISE_48.wav
UM_48	S2_UM_48.wav
FTM_48	S2_FTM_48.wav
FTM_48+PNS	S2_FTM_48+PNS.wav
allNoise_64	S2_ALL_NOISE_64.wav
UM_64	S2_UM_64.wav
FTM_64	S2_FTM_64.wav
FTM_64+PNS	S2_FTM_64+PNS.wav
S3	
EXCERPT	**FILE NAME**
Reference	S3_REF.wav
3.5 kHz anchor	S3_3.5kHz_ANCHOR.wav
allNoise_48	S3_ALL_NOISE_48.wav
UM_48	S3_UM_48.wav
FTM_48	S3_FTM_48.wav
FTM_48+PNS	S3_FTM_48+PNS.wav
allNoise_64	S3_ALL_NOISE_64.wav
UM_64	S3_UM_64.wav
FTM_64	S3_FTM_64.wav
FTM_64+PNS	S3_FTM_64+PNS.wav

POLITECHNIKA GDAŃSKA
Wydział Elektroniki, Telekomunikacji
i Informatyki

Maciej Kulesza

Nowa Metoda Estymacji Tonalności Widmowej dla Potrzeb Kodowania Sygnałów Fonicznych

Rozprawa doktorska

STRESZCZENIE

Promotor:
prof. dr hab inż. Andrzej Czyżewski
Wydział Elektroniki, Telekomunikacji
i informatyki
Politechnika Gdańska

Gdansk, 2010

1 WPROWADZENIE

Przedstawiona rozprawa doktorska dotyczy efektywności algorytmów perceptualnego kodowania sygnałów fonicznych takich jak MP3 (MPEG 1 Layer 3) oraz AAC (Advanced Audio Coding). Kodeki tego typu dokonują kwantyzacji widma sygnału (współczynników MDCT) w taki sposób by wprowadzany szum kwantyzacji pozostawał, o ile to możliwe, poniżej chwilowego progu słyszenia. Z tego względu jednym z kluczowych elementów każdego kodeka perceptualnego jest model psychoakustyczny symulujący zjawiska zachodzące w systemie słuchowym człowieka, a w szczególności zjawisko maskowania jednoczesnego. Aby wiarygodnie określić chwilowy próg słyszenia, konieczna jest klasyfikacja rodzaju pobudzenia błony podstawnej na pobudzenie o charakterze szumowym lub tonalnym. W modelach psychoakustycznych zwykle stosowane są proste metody pozwalające na efektywne rozróżnienie sygnałów szumowych oraz tonów prostych. Niemniej nagrania muzyczne zawierają zwykle oprócz składowych tonalnych o niezmiennej częstotliwości chwilowej również składniki tonalne o modulowanej częstotliwości chwilowej. Modulacja wprowadzana jest przez instrumentalistów lub wokalistów w wyniku stosowania technik artykulacyjnych takich jak vibrato. Popularne metody rozróżniania charakteru pobudzenia zawodzą w przypadku gdy w sygnale występują modulowane składowe tonalne. Niewłaściwa klasyfikacja rodzaju pobudzenia błony podstawnej powoduje, iż estymata progu słyszenia nie odpowiada rzeczywistemu progowi słyszenia. W rezultacie może to prowadzić do degradacji jakości kodowania, gdyż widmo szumu kwantyzacji wprowadzane przez koder nie jest kształtowane zgodnie z rzeczywistym progiem słyszenia.

Głównym celem badań przedstawionych w ramach niniejszej rozprawy było opracowanie nowego algorytmu pozwalającego na rozróżnienie składowych szumowych oraz tonalnych o modulowanej i niemodulowanej częstotliwości chwilowej. Metoda ta jest oryginalnym wkładem autora rozprawy. Opisany algorytm dokonuje przetwarzania widma amplitudowego oraz fazowego sygnału fonicznego. W kolejnych krokach analizy wyodrębniane są maksima lokalne widma amplitudowego. Następnie tworzone są trójelementowe ciągi indeksów widma. Indeksy te odpowiadają maksimom lokalnym wykrytym w trzech kolejnych widmach sygnału. Komponenty widma o indeksach należących do danego ciągu stanowią kandydatów na komponenty tonalne. Utworzone ciągi są traktowane jako kandydaci do trójelementowych ścieżek tonalnych.

Weryfikacja kandydatów następuje na podstawie porównania zmian ich częstotliwości chwilowej określanych jednocześnie z wykorzystaniem estymatora bazującego na przetwarzaniu próbek widma amplitudowego oraz estymatora bazującego na przetwarzaniu widma fazowego sygnału. Zaproponowana metoda określania rodzaju pobudzenia błony podstawnej została zintegrowana z modelem psychoakustcznym zdefiniowanym w ramach standardu MPEG. Standard definiujący system kodowania AAC zawiera opis techniki kodowania w której podpasma sygnału zawierające wyłącznie komponenty szumowe kodowane są w inny sposób niż pasma zawierające komponenty tonalne. Technika ta określana jest jako Perceptual Noise Substitution (PNS). Autor rozprawy opracował nowy detektor podpasm sygnału zawierających wyłącznie komponenty szumowe. Metoda ta wykorzystuje wspomniany wcześniej algorytm klasyfikacji komponentów widma na szumowe i tonalne. Jej przewaga w stosunku do innych implementacji polega na mniejszej złożoności obliczeniowej, która wynika ze stosowania tej samej metody klasyfikacji komponentów widma zarówno w modelu psychoakustcznym jak i w module PNS.

Celem dodatkowym rozprawy było określenie w jakim stopniu efektywność metody klasyfikacji rodzaju pobudzenia błony podstawnej wpływa na wynikową jakość kodowania perceptualnego. Realizacja tego celu wymagała implementacji mechanizmu alokacji bitów stosowanych w kodekach perceptualnych oraz przeprowadzenia serii testów odsłuchowych zgodnych z procedurą ITU-T BS.1534.

Miara określana w niniejszej rozprawie doktorskiej jako *tonalność* jest wykorzystywana do rozróżnienia charakteru szumowego lub tonalnego bądź to pojedynczych próbek widma, bądź zdefiniowanych podpasm sygnału. Miara ta odzwierciedla stosunek energii komponentów tonalnych oraz szumowych należących do danego podpasma sygnału. Tonalność przyjmuje wartości z zakresu [0,1], gdzie wartość 0 wskazuje, iż dany komponent widma lub podpasmo sygnału jest całkowicie szumowe. Gdy komponent widma odpowiada składnikowi sinusoidalnemu o stałej lub modulowanej częstotliwości albo podpasmo sygnału zawiera jeden lub więcej takich komponentów, miara tonalności przyjmuje wartość równą 1. Algorytm zwracający ciągłą miarę tonalności z zakresu [0,1] nazywany jest *estymatorem tonalności* w przeciwieństwie do *detektora komponentów tonalnych*, który dostarcza jedynie binarnej informacji dotyczącej tonalności. W zależności od wymagań aplikacji konieczne może być zastosowanie estymatora tonalności lub detektora komponentów tonalnych. Każdy estymator tonalności może być wykorzystany jako detektor komponentów tonalnych po przyjęciu odpowiedniego progu.

Rozprawa doktorska zawiera opis badań mających na celu udowodnić dwie poniższe tezy:

1. **Estymacja miary tonalności komponentów sinusoidalnych sygnałów fonicznych o stałej lub modulowanej częstotliwości chwilowej jest możliwa do przeprowadzenia poprzez porównanie zmian ich częstotliwości chwilowych określanych z wykorzystaniem estymatora bazującego na analizie widma amplitudowego oraz estymatora bazującego na analizie widma fazowego.**
2. **Zaproponowana w rozprawie metoda estymacji miary tonalności składowych widma pozwala na efektywne ograniczanie zniekształceń wprowadzanych w procesie perceptualnego kodowania sygnałów fonicznych.**

Streszczenie rozprawy doktorskiej zawiera opis badań związanych z opracowaniem nowego estymatora tonalności komponentów widma oraz jego wykorzystania w aplikacjach kodowania perceptualnego sygnałów fonicznych. Opis wybranych metod kodowania, funkcji jakie pełnią algorytmy estymacji tonalności w aplikacjach przetwarzania sygnałów fonicznych oraz porównanie wybranych metod estymacji tonalności zawierają rozdziały od 1 do 4 rozprawy doktorskiej.

2 NOWA METODA ESTYMACJI TONALNOŚCI KOMPONENTÓW WIDMA

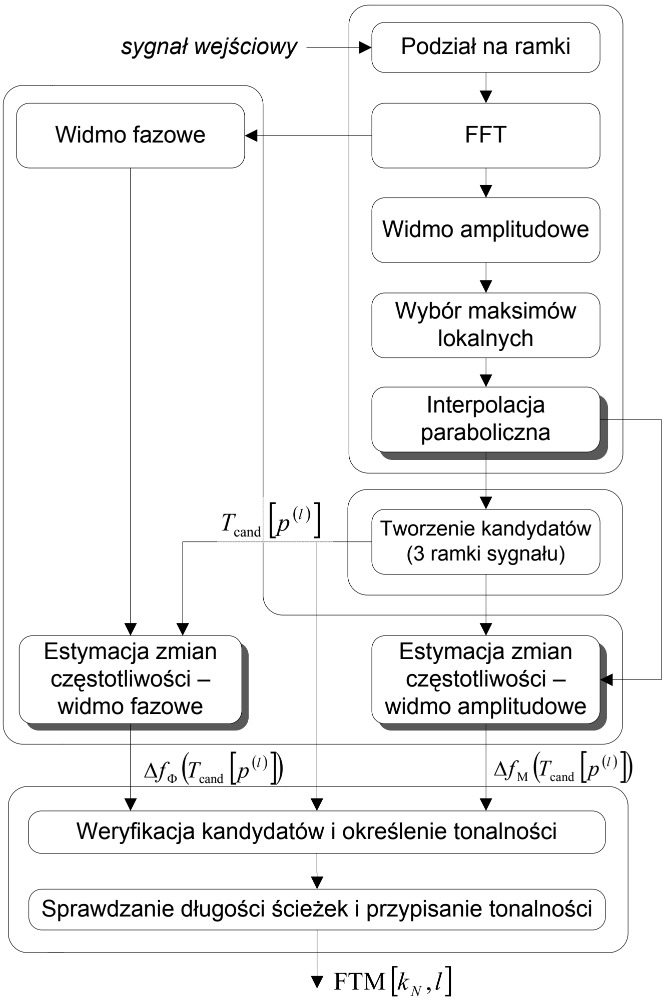
Główne założenia dla nowego algorytmu estymacji tonalności komponentów widma były następujące:

- Algorytm powinien dokonywać przetwarzania próbek pojedynczego widma (tak zwane przetwarzanie wewnątrz-ramkowe) oraz ciągów próbek należących do następujących po sobie w czasie widm (przetwarzanie między-ramkowe) w celu tworzenia ścieżek komponentów tonalnych będących odzwierciedleniem ich zmian częstotliwości.
- Algorytm powinien przypisywać wartości tonalności z zakresu [0,1] maksimom lokalnym widma. Pozostałym komponentom widma powinna być przypisana

wartość tonalności równa 0 (rozmywanie tonalności przypisanej maksimom widma na prążki widma sąsiadujące z nimi stanowi odrębny moduł przetwarzania)

- Algorytm powinien pozwalać na efektywną detekcję komponentów tonalnych związanych ze składowymi sinusoidalnymi o modulowanej częstotliwości chwilowej

Schemat blokowy opracowanej metody estymacji tonalności komponentów widma przedstawiono na rys. 2.1.


sygnał wejściowy
Podział na ramki
Widmo fazowe
FFT
Widmo amplitudowe
Wybór maksimów lokalnych
Interpolacja paraboliczna
$T_{\mathrm{cand}}\left[p^{(l)}\right]$
Tworzenie kandydatów (3 ramki sygnału)
Estymacja zmian częstotliwości – widmo fazowe
Estymacja zmian częstotliwości – widmo amplitudowe
$\Delta f_{\Phi}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$
$\Delta f_{\mathrm{M}}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$
Weryfikacja kandydatów i określenie tonalności
Sprawdzanie długości ścieżek i przypisanie tonalności
$\mathrm{FTM}\left[k_N, l\right]$

Rys. 2.1 Schemat blokowy opracowanego algorytmy estymacji tonalności komponentów widma

2.1 PRZETWARZANIE WIDMA AMPLITUDOWEGO

Sygnał wejściowy jest dzielony na ramki ważone oknem Hanna, przy czym długość ramki oraz zakładka stanowią parametr metody. Opcjonalnie każda ramka uzupełniana jest zerami do jej podwójnej długości, co ma wpływ na redukcję obciążenia estymatorów częstotliwości chwilowej opisanych później. W następnej kolejności obliczana jest transformata DFT z wykorzystaniem algorytmu FFT oraz widmo amplitudowe sygnału. W każdym widmie amplitudowym dokonywana jest detekcja jego maksimów lokalnych a następnie określany jest parametr peakiness zdefiniowany jako:

$$g\left[k_{\max}\left[i^{(l)}\right]\right]=A\left[k_{\max}\left[i^{(l)}\right],l\right]-\frac{A\left[k_{\min-}\left(i^{(l)}\right),l\right]+A\left[k_{\min+}\left(i^{(l)}\right),l\right]}{2} \qquad (2.1)$$

gdzie: $k_{\min-}\left[i^{(l)}\right]$ oraz $k_{\min+}\left[i^{(l)}\right]$ są indeksami próbek widma odpowiadającymi minimum lokalnym położonym po obu stronach lokalnego maksima widma oznaczonego jako $k_{\max}\left[i^{(l)}\right]$, $A\left[k_{\max}\left(j^{(l)}\right)\right]=20\log\left(\left|X\left[k_{\max}\left(j^{(l)}\right),l\right]\right|\right)$, $X[k,l]$ – widmo zespolone, i – indeks próbek widma, l – indeks widma.

Jedynie maksima lokalne widma dla których $g\left[k_{\max}\left[i^{(l)}\right]\right]$ >9dB są dalej rozpatrywane jako potencjalne komponenty tonalne. Pozostałym maksimom widma przypisywana jest wartość tonalności równa 0. W kolejnym kroku pozostałe maksima lokalne wraz z sąsiadującym z nimi próbkami widma poddawane są procedurze interpolacji parabolicznej:

$$k_{\mathrm{off}}\left[j^{(l)}\right]=\frac{1}{2}\frac{A\left[k_{\mathrm{s}\max}\left[j^{(l)}\right]-1\right]-A\left[k_{\mathrm{s}\max}\left[j^{(l)}\right]+1\right]}{A\left[k_{\mathrm{s}\max}\left[j^{(l)}\right]-1\right]-2A\left[k_{\mathrm{s}\max}\left[j^{(l)}\right]\right]+A\left[k_{\mathrm{s}\max}\left[j^{(l)}\right]+1\right]} \qquad (2.2)$$

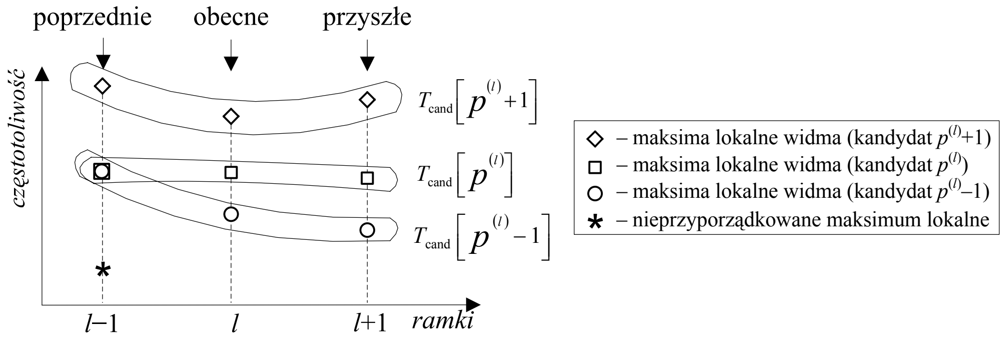
gdzie: $k_{\mathrm{s}\max}\left[j^{(l)}\right]$ oznacza indeks widma odpowiadający wybranemu maksimum lokalnemu, j – indeks wybranego maksima lokalnego. Na tej podstawie dokonywana jest estymacja częstotliwości chwilowej dla wybranego maksimum lokalnego widma zgodnie z zależnością:

$$f_{\mathrm{M}}\left[k_{\mathrm{s}\max}\left[j^{(l)}\right]\right]=\frac{k_{\mathrm{s}\max}\left[j^{(l)}\right]+k_{\mathrm{off}}\left[j^{(l)}\right]}{Z_{\mathrm{p}}N}F_{\mathrm{s}} \qquad (2.3)$$

gdzie: N – długość ramki sygnału, Z_{p} – współczynnik wypełnienia zerami (stosunek długości FFT do długości ramki), F_{s} – częstotliwość próbkowania.

2.2 TWORZENIE KANDYDATÓW DO TRÓJELEMENTOWYCH ŚCIEŻEK TONALNYCH

Kandydat do trójelementowej ścieżki tonalnej tworzony jest z wykorzystaniem trzech kolejnych widm amplitudowych sygnału. Przyjmuje się, iż kandydatem jest ciąg trzech indeksów widma odpowiadających maksimom lokalnym w trzech kolejnych ramkach sygnału, dla których zmiana częstotliwości jest minimalna (patrz rys. 2.2).

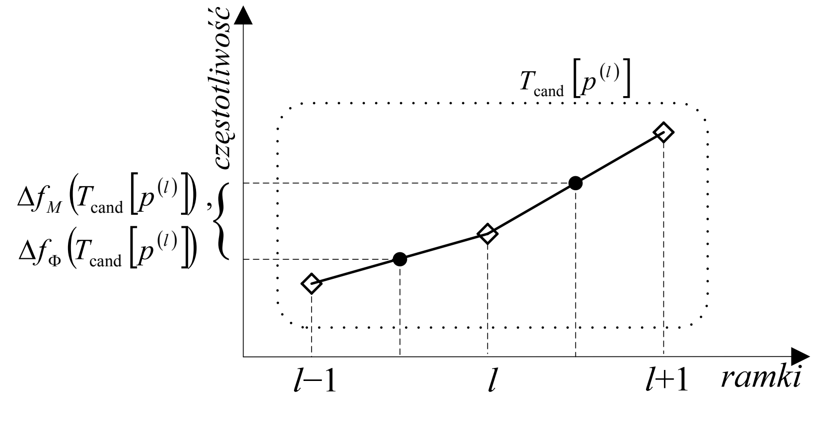

poprzednie
obecne
przyszłe
częstotoliwość
$T_{\mathrm{cand}}\left[p^{(l)}+1\right]$
$T_{\mathrm{cand}}\left[p^{(l)}\right]$
$T_{\mathrm{cand}}\left[p^{(l)}-1\right]$
◇ – maksima lokalne widma (kandydat $p^{(l)}+1$)
□ – maksima lokalne widma (kandydat $p^{(l)}$)
○ – maksima lokalne widma (kandydat $p^{(l)}-1$)
✱ – nieprzyporządkowane maksimum lokalne
$l-1$
l
$l+1$
ramki

Rys. 2.2 Tworzenie kandydatów do trójelementowych ścieżek tonalnych

Kandydaci do trójelementowych ścieżek tonalnych są oznaczani jako $T_{\mathrm{cand}}\left[p^{(l)}\right]$ gdzie $p^{(l)} = 0,1,...,P^{(l)} - 1$ jest indeksem kandydata, a $P^{(l)}$ jest ich całkowitą liczbą.

2.3 ESTYMACJA ZMIAN CZĘSTOTLIWOŚCI

Dla każdego kandydata na ścieżkę tonalną zmiana częstotliwości (patrz rys. 2.3) określana jest z wykorzystaniem dwóch metod: wykorzystującej próbki widma amplitudowego – $\Delta f_{\mathrm{M}}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$, oraz próbki widma fazowego $-\Delta f_{\Phi}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$.


częstotliwość
$T_{\mathrm{cand}}\left[p^{(l)}\right]$
$\Delta f_M\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$,
$\Delta f_{\Phi}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)$
$l-1$
l
$l+1$
ramki

Rys. 2.3 Zmiana częstotliwości związana z kandydatem na ścieżkę tonalną

2.4 ESTYMATOR WYKORZYSTUJĄCY PRÓBKI WIDMA AMPLITUDOWEGO

Estymacja zmian częstotliwości kandydatów na trójelementowe ścieżki tonalne przeprowadzana jest z wykorzystaniem częstotliwości chwilowych obliczonych na podstawie równania (2.3) zgodnie z zależnością:

$$\Delta f_{\mathrm{M}}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)=\frac{f_{\mathrm{M}}\left[k_{\mathrm{s\,max}}\left[j^{(l+1)}\right]\right]-f_{\mathrm{M}}\left[k_{\mathrm{s\,max}}\left(j^{(l-1)}\right)\right]}{2} \tag{2.4}$$

2.5 ESTYMATOR WYKORZYSTUJĄCY PRÓBKI WIDMA FAZOWEGO

Estymator wykorzystujący próbki widma fazowego został opracowany specjalnie na potrzeby proponowanego algorytmu estymacji tonalności). Na początku określany jest dyferencjał fazy drugiego rzędu związany z danym kandydatem na ścieżkę tonalną:

$$\Delta^2\Phi\left(k_{\mathrm{s\,max}}\left[j^{(l+1)}\right],k_{\mathrm{s\,max}}\left[j^{(l-1)}\right]\right)=\Phi\left[k_{\mathrm{s\,max}}\left[j^{(l-1)}\right]\right]-2\Phi\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right]\right]+\Phi\left[k_{\mathrm{s\,max}}\left[j^{(l+1)}\right]\right] \tag{2.5}$$

gdzie: $\Phi\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right]\right]=\arctan\left(\dfrac{\mathrm{Im}\left(X\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right],l\right]\right)}{\mathrm{Re}\left(X\left[k_{\mathrm{s\,max}}\left[j^{(l)}\right],l\right]\right)}\right).$

Jeżeli kandydat na ścieżkę tonalną pochodzi od składnika sinusoidalnego o modulowanej częstotliwości, indeksy widma w kolejnych widmach są różne. Konieczne jest obliczenia dodatkowego czynnika fazowego:

$$\begin{aligned}&\Delta^2\phi\left(k_{\mathrm{s\,max}}\left[j^{(l+1)}\right],k_{\mathrm{s\,max}}\left[j^{(l-1)}\right]\right)=\\&\frac{\pi(N-1)}{Z_{\mathrm{p}}N}\left(k_{\mathrm{s\,max}}\left[j^{(l-1)}\right]-2k_{\mathrm{s\,max}}\left[j^{(l)}\right]+k_{\mathrm{s\,max}}\left[j^{(l+1)}\right]\right)\end{aligned} \tag{2.6}$$

Zmiana częstotliwości określana jest następująco:

$$\begin{aligned}&\Delta f_{\Phi}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)\ =\\&\frac{f_s}{2\pi L}\left(\mathrm{princarg}\left(\Delta^2\Phi\left(k_{\mathrm{s\,max}}\left[j^{(l+1)}\right],k_{\mathrm{s\,max}}\left[j^{(l-1)}\right]\right)\right)+\Delta^2\phi\left(k_{\mathrm{s\,max}}\left[j^{(l+1)}\right],k_{\mathrm{s\,max}}\left[j^{(l-1)}\right]\right)\right)\end{aligned} \tag{2.7}$$

gdzie: L – skok analizy STFT (ang. *Short Time Fourier Transform*), $\mathrm{princarg}(\varphi)=(\varphi+\pi)\mathrm{mod}(-2\pi)+\pi$ jest funkcją rzutującą fazę do zakresu ±π.

Jeśli składnik sinusoidalny sygnału zmienia swoją częstotliwości bardzo szybko w czasie, odpowiadający zmianie częstotliwości skok fazy może przekroczyć zakres ±π co uniemożliwi prawidłową estymacje zmiany częstotliwości. W tym celu dla każdego kandydata określana jest minimalna i maksymalna możliwa zmiana częstotliwości na podstawie indeksów prążków widma stanowiących ciąg odpowiadający danemu kandydatowi. Następnie w sposób iteracyjny dodawany lub odejmowany jest skok częstotliwości odpowiadający wielokrotności F_s/L. Jeśli w wyniku iteracji zmiana częstotliwości określona z wykorzystaniem wartości fazy znajdzie się w przedziale określonym na podstawie indeksów widma tworzących danego kandydata, wartość zmiany częstotliwości modyfikowana jest następująco:

$$\Delta f_{\Phi}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right) \leftarrow \Delta f_{\Phi}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right) + m_j \frac{F_{\mathrm{s}}}{L} \tag{2.8}$$

gdzie m_j=±1, ±2, …,±6. W przeciwnym wypadku m_j=0.

2.6 WERYFIKACJA KANDYDATÓW I OKREŚLANIE TONALNOŚCI

Dla każdego kandydata określana jest różnica pomiędzy estymatą zmiany częstotliwości określoną z wykorzystaniem estymatora przetwarzającego próbki widma amplitudowego oraz fazowego:

$$\delta\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right) = \Delta f_{\mathrm{M}}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right) - \Delta f_{\Phi}\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right) \tag{2.9}$$

Kandydaci dla których $\left|\delta\left(T_{\mathrm{cand}}\left[p^{(l)}\right]\right)\right| < \frac{F_{\mathrm{s}}}{NZ_{\mathrm{p}}}$ uznawani są ścieżki tonalne, pozostałe zaś jako ciągi zawierające komponenty widmowe szumowe. Dla trójelementowych ścieżek tonalnych określana jest wartość FTM (ang. *Frequency-derived Tonality Measure*):

$$\mathrm{FTM}_{\mathrm{trk}}\left(T_{\mathrm{trk}}\left[r^{(l)}\right]\right) = 1 - \frac{NZ_{\mathrm{p}}}{F_{\mathrm{s}}}\left|\delta\left(T_{\mathrm{trk}}\left[r^{(l)}\right]\right)\right| \tag{2.10}$$

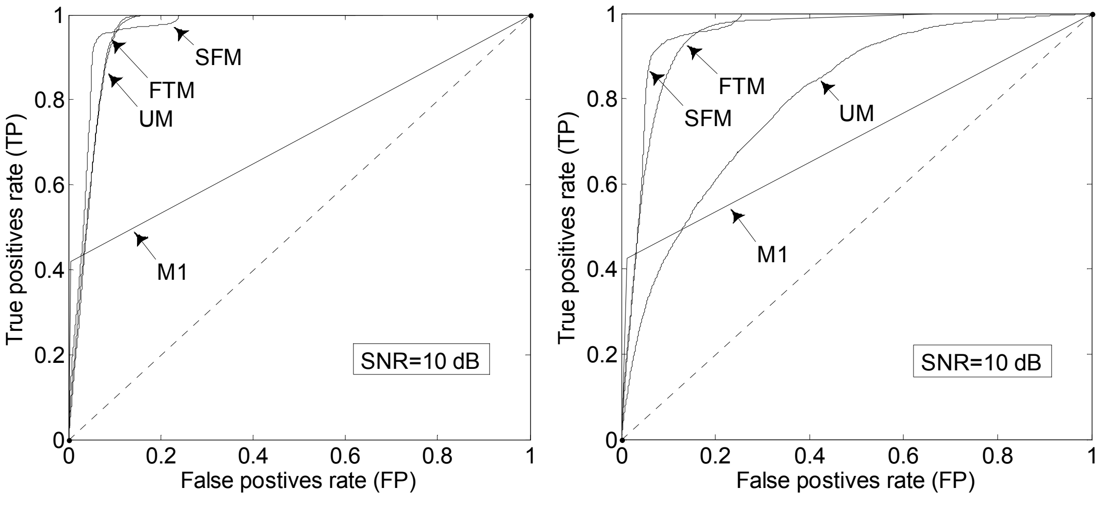
gdzie $r^{(l)}$ jest indeksem ścieżki tonalnej. Wartość $\mathrm{FTM}_{\mathrm{trk}}$ jest przypisywana indeksowi widma odpowiadającemu drugiemu elementowi danej ścieżki tonalnej i oznaczana dalej jako ftm[k_{max}]. Dodatkowo badana jest długość ścieżki tonalnej. Jeśli ścieżka jest odpowiednio długa, wartość $\mathrm{FTM}_{\mathrm{trk}}$ przyporządkowywana jest również ostatniemu elementowi ścieżki która się kończy.

3 BADANIE EFEKTYWNOŚCI OPRACOWANEGO ALGORYTMU

Efektywność opracowanego algorytmu porównano z efektywnością wybranych metod estymacji miary tonalności lub metod detekcji komponentów tonalnych widma wykorzystywanych w modelach psychoakustycznych. Do porównania wykorzystano następujące metody: SFM (ang. *Spectral Flatness Measure*), UM (ang. *Unpredictibility Measure*), M1 – detektor wykorzystywany w pierwszym modelu psychoakustycznym MPEG.

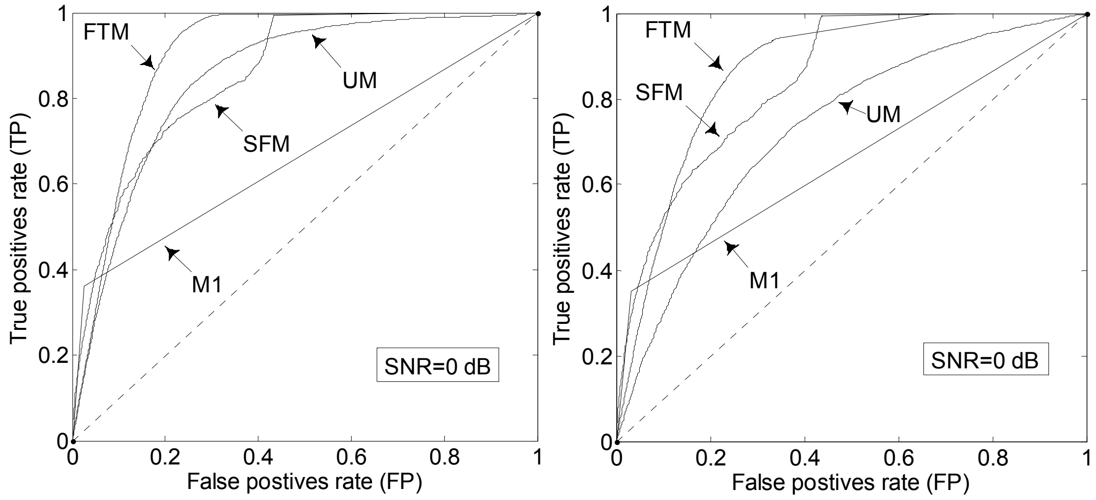
3.1 EFEKTYWNOŚĆ DETEKCJI KOMPONENTÓW TONALNYCH

W celu porównania efektywności detekcji komponentów tonalnych przez wybrane metody wygenerowano szereg sygnałów harmonicznych o różnych częstotliwościach podstawowych i określonych stosunkach sygnału do szumu. Sygnał harmoniczny składał się z komponentów niemodulowanych lub modulowanych częstotliwościowo. Na tej podstawie wygenerowano krzywe ROC (ang. *Reciever Operating Characteristics*) dla sygnałów o stosunku sygnału do szumu odpowiednio 10 i 0 dB (rys. 3.1 oraz 3.2).



Rys. 3.1 Krzywe ROC dla badanych algorytmów (SNR=10 dB, po lewej – sinusoidy o stałej częstotliwości, po prawej – sinusoidy modulowane częstotliwościowo)

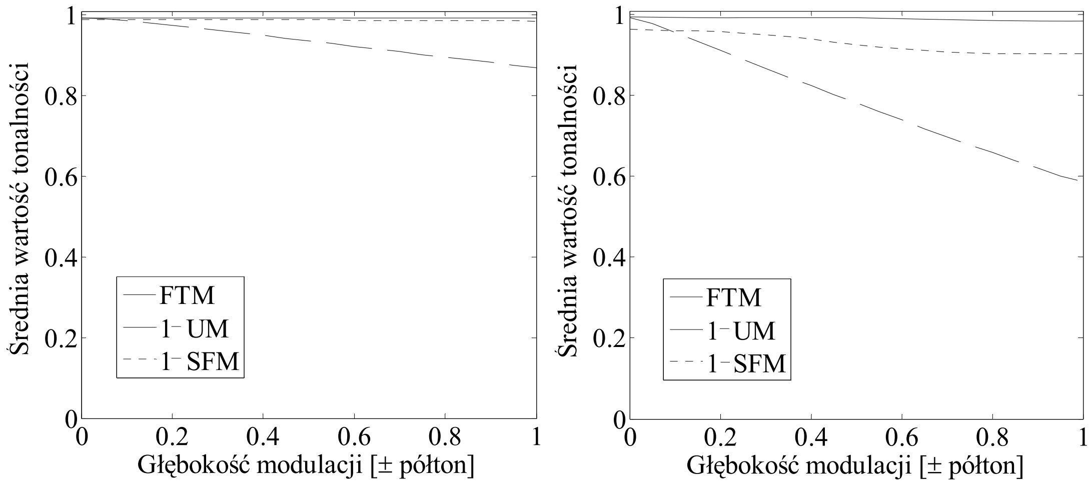
Daje się zaobserwować, iż w przypadku sygnałów zawierających sinusoidy o niezmiennej w czasie częstotliwości efektywność detekcji dla wszystkich metod, z wyjątkiem metody oznaczonej jako M1, jest podobna. W przypadku gdy analizie podlegają modulowane sinusoidy, opracowana metoda jest bardziej efektywna w detekcji komponentów tonalnych od metod pozostałych. W szczególności efektywność metody stosowanej szeroko w kodekach MP3 oraz AAC znacząco spada dla sygnałów zawierających modulowane sinusoidy.



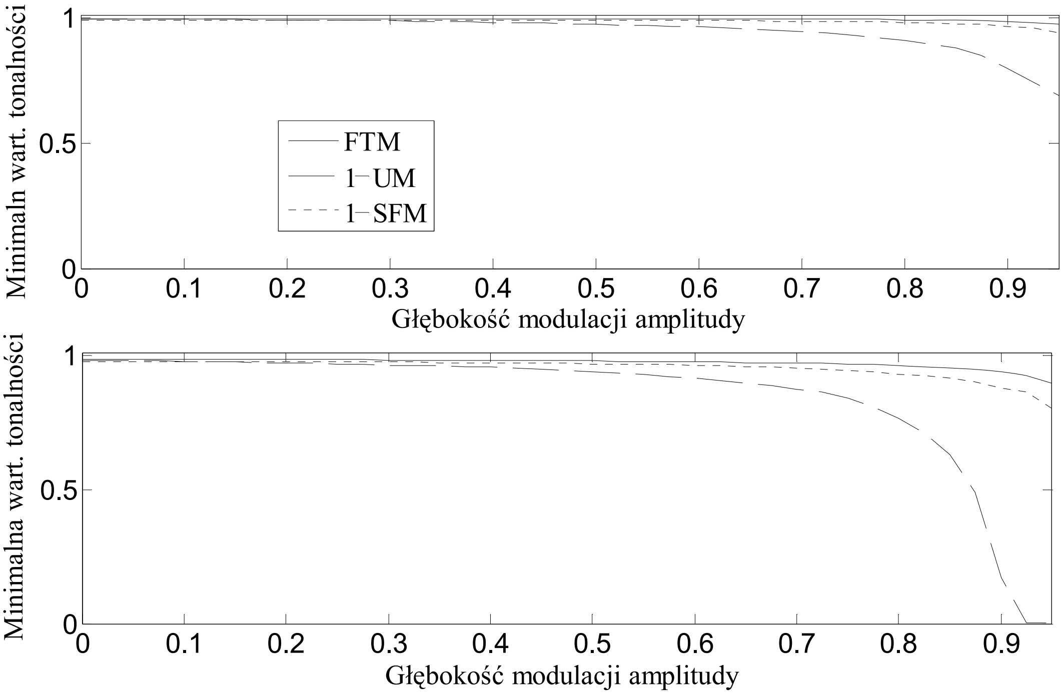
Rys. 3.2 Krzywe ROC dla badanych algorytmów (SNR=0 dB, po lewej – sinusoidy o stałej częstotliwości, po prawej – sinusoidy modulowane częstotliwościowo)

3.2 WPŁYW MODULACJI NA ESTYMATĘ TONALNOŚCI

Zbadano wpływ głębokości modulacji częstotliwości oraz amplitudy na różne miary tonalności, w tym na miarę zaproponowaną w niniejszej rozprawie doktorskiej. Sygnałem testowym był sygnał sinusoidalny o częstotliwości 120 Hz oraz 440 Hz, stosunek sygnału do szumu wynosił 20 dB. Sygnał ten był modulowany z częstotliwością 6 Hz, przy czym maksymalna głębokość modulacji wynosiła ±1 półton. Wpływ głębokości modulacji częstotliwości na miary tonalności przedstawiono na rys. 3.3. Z analizy wyników przedstawionych na rys. 3.3 wynika, iż opracowana metoda jest w dużym stopniu niewrażliwa na efekt modulacji częstotliwości. Warto zauważyć, iż miara UM zwraca wartość poniżej 0,6 dla głębokości modulacji ±1 półton. Algorytm FTM jest również odporny na wpływ modulacji amplitudy co zilustrowano na rys. 3.4



Rys. 3.3 Wpływ głębokości modulacji częstotliwości na średnią miarę tonalności – częstotliwość nośnej wynosi odpowiednio 120 Hz (po lewej) i 440 Hz (po prawej)



Rys. 3.4 Wpływ modulacji amplitudy składnika sinusoidalnego na średnie (górny rys.) oraz minimalne (dolny rys.) wartości tonalności określane z wykorzystaniem metod UM, SFM oraz FTM.

3.3 OGRANICZENIA OPRACOWANEJ METODY

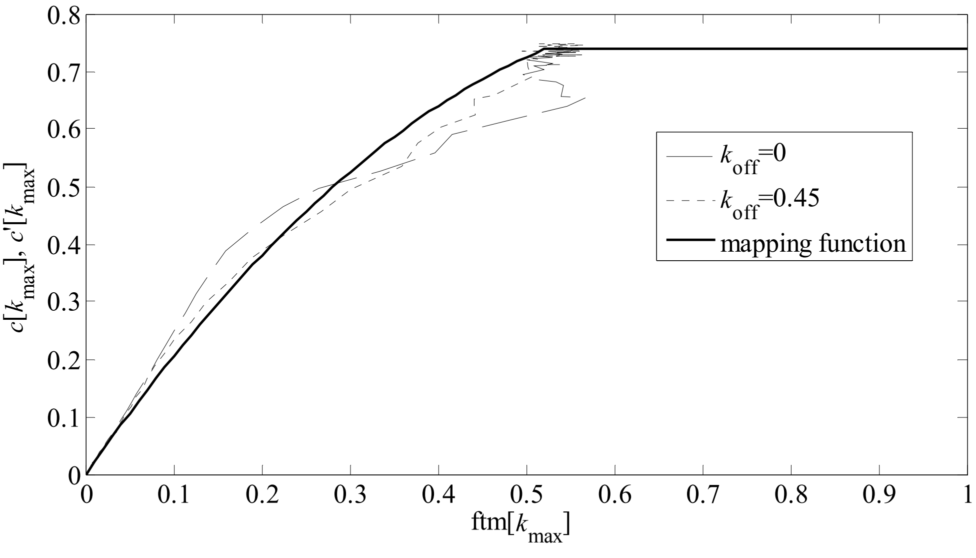
Analiza polifonicznych nagrań muzycznych ujawnia słabość opracowanego algorytmu związanego z estymacją tonalności w przypadku gdy w nagraniu występują sygnały harmoniczne o niskiej częstotliwości podstawowej (np. kontrabas). Dla długości ramek analizy stosowanych w kodekach z grupy MPEG, uzyskiwana rozdzielczość częstotliwościowa analizy widmowej może być zbyt niska by komponenty tonalne związane z harmonicznymi sygnału o niskiej częstotliwości podstawowej mogły być rozpoznane jako maksima lokalne widma. Z drugiej strony badania wskazują, iż metoda UM stosowana standardowo w modelu psychoakustycznym MPEG funkcjonuje prawidłowo dla dolnego zakresu częstotliwości (rys. 3.3). Z tego względu efektywna estymacja tonalności może być przeprowadzona z wykorzystaniem metody hybrydowej bazującej na algorytmie UM w dolnym zakresie pasma słyszalnego (np. do 300 Hz) oraz metody FTM dla pozostałych częstotliwości.

4 WIARYGODNOŚĆ ESTYMACJI PROGU SŁYSZENIA

W dalszych badaniach wykorzystano model psychoakustyczny MPEG stosowany w kodekach MP3 oraz AAC. Stosowany standardowo estymator tonalności UM przypisuje nie tylko wysoką wartość tonalności maksimom lokalnym widma związanym ze składnikami sinusoidalnymi sygnału ale także próbką widma sąsiadującym z maksimum lokalnym widma. Z tego względu konieczne było dopasowanie opracowanego algorytmu FTM tak by mógł być zastosowany jako zamiennik dla algorytmu standardowego. Założono, że dla komponentów sinusoidalnych niemodulowanych model psychoakustyczny powinien generować identyczny próg słyszenia bez względu na to, czy wykorzystano standardowy estymator tonalności czy też algorytmy FTM.

4.1 RZUTOWANIE FTM NA UM DLA MAKSIMÓW LOKALNYCH WIDMA

Zależność miary tonalności w funkcji stosunku sygnału do szumu składników sinusoidalnych jest różna dla metod FTM oraz UM. Z tego względu wyznaczono charakterystykę pozwalającą na modyfikację wartości generowanych przez algorytm FTM w taki sposób by odpowiadały wartościom UM. W tym celu generowano sygnały sinusoidalne o różnych częstotliwościach oraz zmiennym stosunku sygnału do szumu. Wynikową charakterystykę rzutowania FTM na UM dla maksimów lokalnych widma przedstawiono na rys. 4.1.



Rys. 4.1 Funkcja rzutowania wartości FTM na UM (oznaczenia na wykresie odpowiadają oznaczeniom stosowanym w standardzie MPEG)

Wyniki wskazują, iż charakterystyka ta w niewielkim stopniu zależy od częstotliwości komponentu sinusoidalnego. Charakterystykę tą zaproksymowano korzystając z funkcji kwadratowej:

$$c'[k_{\max}] = \begin{cases} -1.5\mathrm{ftm}^2[k_{\max}] + 2.2\mathrm{ftm}[k_{\max}], & \mathrm{ftm}[k_{\max}] \le 0.52 \\ 0.74, & \mathrm{ftm}[k_{\max}] > 0.52 \end{cases}. \tag{4.1}$$

gdzie: $c'[k_{\max}]$ jest wartością tonalności odpowiadającą standardowej mierze tonalności $c[k_{\max}]$ stosowanej w drugim modelu psychoakustycznym MPEG.

4.2 PRZYPISYWANIE MIARY TONALNOŚCI KOMPONENTOM NIE BĘDĄCYM MAKSIMAMI LOKALNYMI WIDMA

W pracy doktorskiej zaproponowano dwie metody pozwalające na przypisanie miary tonalności próbką widma sąsiadującym z maksimami lokalnymi, dla których określono już odpowiednią wartość miary tonalności. Proces ten nazywana jest dalej rozmywaniem miary FTM. Z opisanych metod bazuje na dokładnym modelowaniu zjawiska rozmywania tonalności jakie ma miejsce w przypadku metody UM. Chociaż metoda ta jest dokładna, to wymaga dużej ilości obliczeń. Z tego względu opracowano prostszą metodę heurystyczną pozwalającą na równie dobre modelowanie rozmywania występującego w przypadku algorytmu UM. Metoda ta wykorzystuje parametr peakiness zdefiniowany w równaniu (2.1). W pierwszej kolejności obliczany jest parametr związany z rozpatrywanym maksimum widma:

$$m_g\left[k_{\max}\right]=\left(\frac{g\left[k_{\max}\right]-g_{\mathrm{thd}}}{g_{\mathrm{tnl}}\left[k\right]-g_{\mathrm{thd}}}\right)^2 \tag{4.2}$$

gdzie: g_{thd}=9 dB tak jak poprzednio, natomiast $g_{\mathrm{tnl}}[k]$ dane jest zależnością:

$$g_{\mathrm{tnl}}\left[k\right]=\begin{cases}18+\dfrac{12k}{\left(\left\lfloor\dfrac{Nf_{\mathrm{c1}}}{F_{\mathrm{s}}}\right\rfloor\right)^2}, & k<\left\lfloor\dfrac{Nf_{\mathrm{c1}}}{F_{\mathrm{s}}}\right\rfloor\\ 30, & k\geq\left\lfloor\dfrac{Nf_{\mathrm{c1}}}{F_{\mathrm{s}}}\right\rfloor\end{cases} \tag{4.3}$$

gdzie: f_{c1}=800 Hz jest wartością określoną eksperymentalnie, $\lfloor\ \rfloor$ oznacza zaokrąglenie do wartości całkowitej mniejszej od wartości poddanej tej operacji. W następnym kroku określane są zależności energii pomiędzy próbką widma reprezentującą maksimum lokalne a próbkami sąsiadującymi z nią po obu jej stronach.

$$m_{e-}\left[k_{\max}\right]=\left(\frac{r\left[k_{\max}-1\right]}{r\left[k_{\max}\right]}\right)^{0.5} \tag{4.4}$$

$$m_{e+}\left[k_{\max}\right]=\left(\frac{r\left[k_{\max}+1\right]}{r\left[k_{\max}\right]}\right)^{0.5} \tag{4.5}$$

gdzie: $r[k]$ oznacza próbkę widma amplitudowego w skali liniowej (zgodnie z konwencją stosowaną w standardzie MPEG). W kolejnym kroku parametry zdefiniowane w równaniach (4.2), (4.4) oraz (4.5) są mnożone przez siebie i ograniczane do wartości 1:

$$m_{ge-}\left[k_{\max}\right]=\begin{cases}m_g\left[k_{\max}\right]m_{e-}\left[k_{\max}\right], & m_g\left[k_{\max}\right]m_{e-}\left[k_{\max}\right]<1\\ 1, & m_g\left[k_{\max}\right]m_{e-}\left[k_{\max}\right]\geq 1\end{cases} \tag{4.6}$$

$$m_{ge+}[k_{\max}] = \begin{cases} m_g[k_{\max}]m_{e+}[k_{\max}], & m_g[k_{\max}]m_{e+}[k_{\max}] < 1 \\ 1, & m_g[k_{\max}]m_{e+}[k_{\max}] \geq 1 \end{cases} \tag{4.7}$$

Wartości tonalności $c'[k_{\max} - 1]$ oraz $c'[k_{\max} + 1]$ wyznaczane na podstawie poniższych zależności:

$$c'[k_{\max} - 1] = c'[k_{\max}] + (c_{\mathrm{ns}} - c'[k_{\max}])(1 - m_{ge-}[k_{\max}]) \tag{4.8}$$

$$c'[k_{\max} + 1] = c'[k_{\max}] + (c_{\mathrm{ns}} - c'[k_{\max}])(1 - m_{ge+}[k_{\max}]) \tag{4.9}$$

gdzie: c_{ns}=0,74 jest średnią wartością $c[k]$ w przypadku gdy biały szum Gaussowski jest analizowany z wykorzystaniem metody UM.

4.3 HYBRYDOWY ESTYMATOR TONALNOŚCI

Ze względu na ograniczenia metody FTM dla niskich częstotliwości pasma akustycznego, o których wspomniano w rozdziale 3.3 w dalszych eksperymentach stosowany był hybrydowy estymator tonalności zdefiniowany jako:

$$c''[k] = \begin{cases} \min(c[k], c'[k]), & k \leq \left\lfloor \dfrac{2048 f_{\mathrm{c2}}}{F_{\mathrm{s}}} \right\rfloor \\ c'[k], & k > \left\lfloor \dfrac{2048 f_{\mathrm{c2}}}{F_{\mathrm{s}}} \right\rfloor \end{cases} \tag{4.10}$$

gdzie: f_{c2}=300 Hz zostało określone eksperymentalnie.

4.4 BADANIE WIARYGODNOŚCI ESTYMACJI PROGU SŁYSZENIA

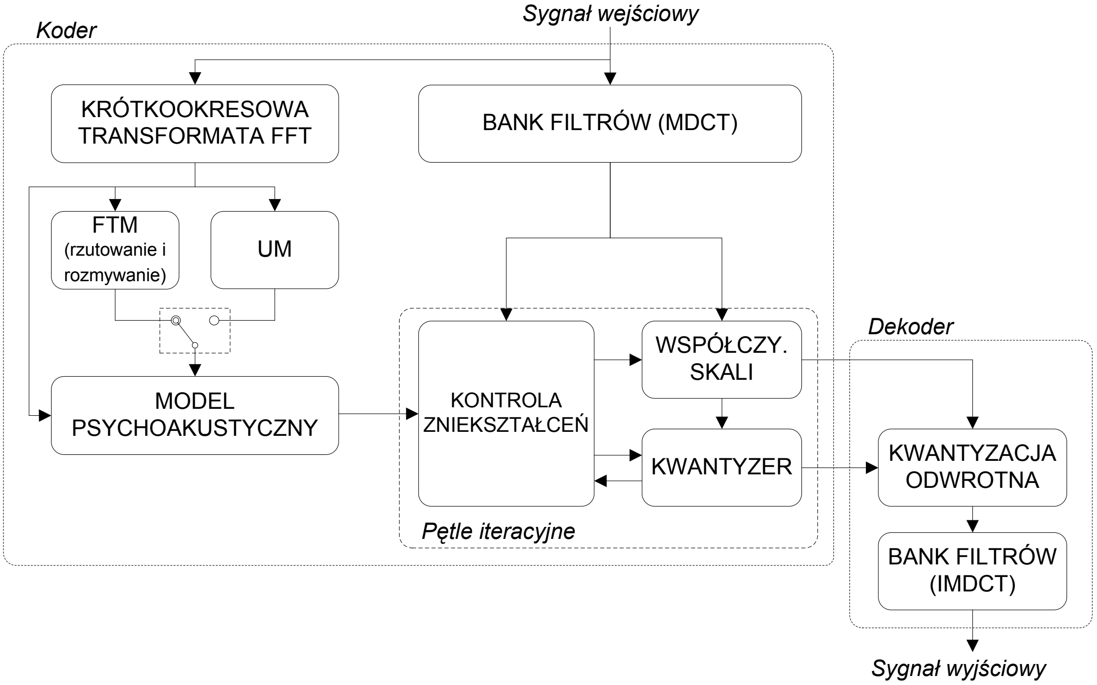
W celu zbadania wiarygodności estymacji progu słyszenia z wykorzystaniem modelu psychoakustycznego pracującego ze standardowym i proponowanym estymatorem tonalności wykorzystano algorytm kodowania bazujący na architekturze kodeka AAC przedstawiony na rys. 4.2.

Sygnał wejściowy poddawany jest segmentacji na bloki z wykorzystaniem metody STFT, a następnie określana jest miara tonalności komponentów widmowych z wykorzystaniem algorytmu FTM albo UM. Na tej podstawie estymowany jest próg słyszenia. Kodowaniu podlegają współczynniki MDCT obliczane równolegle do widma

DFT dla ramek sygnału o długości 2048 lub 256 próbek. Próbki widma kwantowane nierównomiernie zgodnie z poniższą zależnością:

$$x_q[k] = \left\lfloor \left| x^{0.75}[k] \right| 2^{0.1875\left(s_{\mathrm{cfc}} - s_{\mathrm{fc}}[k]\right)} + 0.4054 \right\rfloor \tag{4.11}$$

gdzie: $x[k]$ oznacza współczynniki widma MDCT, s_{cfc} (ang. *common scale-factor*) jest wspólnym parametrem dla wszystkich podpasm w których dokonywana jest kwantyzacja, $s_{\mathrm{fc}}[k]$ (ang. *scale-factor*) jest parametrem przypisanym do danego podpasma częstotliwości. Wartości s_{cfc} oraz $s_{\mathrm{fc}}[k]$ określane są w sposób iteracyjny tak by wprowadzany szum kwantyzacji odpowiadał progowi słyszenia w takim stopniu jak to tylko możliwe.



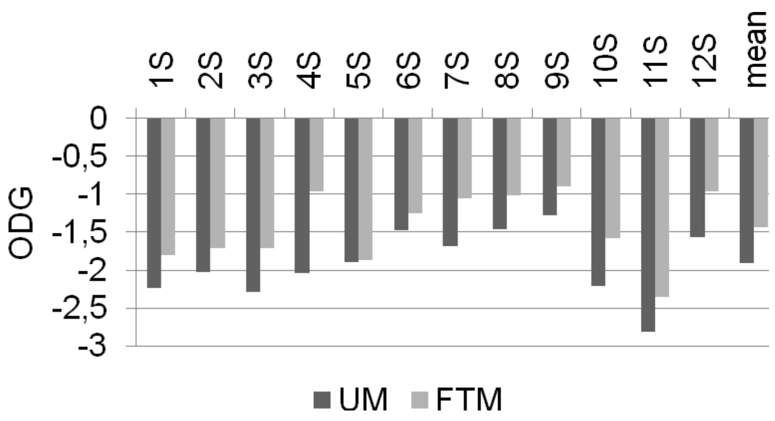
Rys. 4.2 Schemat blokowy kodeka wykorzystanego w badaniach wiarygodności estymacji progu słyszenia

Ze względu na przedmiot badań związany z częścią kodeka odpowiadającą za kodowanie stratne, wykorzystywany algorytm pozbawiony był modułu kodowania bezstratnego. Współczynniki MDCT po przeprowadzeniu kwantyzacji zgodnie z równaniem (4.11) podlegały zdekodowaniu a następnie obliczana była transformata odwrotna IMDCT, na podstawie której wytwarzany był sygnał w dziedzinie czasu.

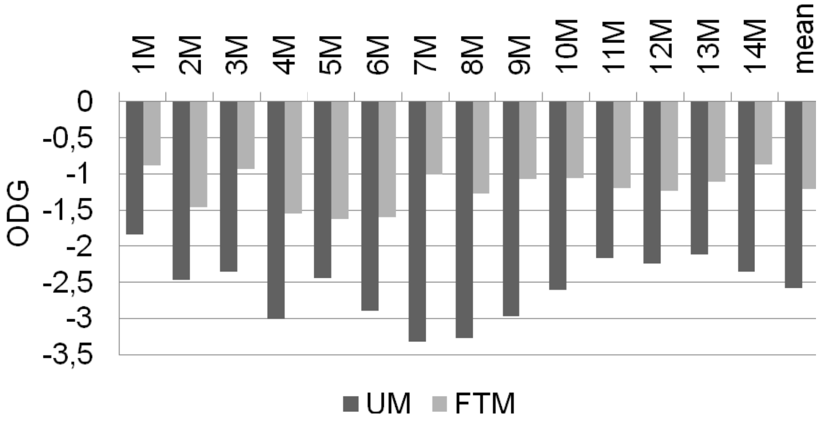
Do testów wykorzystano dwie grupy nagrań muzycznych. Do pierwszej grupy należały nagrania zawierające głównie niemodulowane komponenty tonalne (oznaczone jako

„1S” do „12S”), natomiast do drugiej nagrania zawierające głównie modulowane komponenty tonalne (oznaczone jako „1M” do „14M”). Każde nagranie muzyczne poddano kodowaniu w systemie przedstawionym na rys. 4.2 dwukrotnie. Za pierwszym razem wykorzystano metodę UM do określenia tonalności komponentów widmowych, a za drugim razem algorytm FTM.

W celu określenia stopnia degradacji jakości kodowania ODG (ang. *Objective Difference Grade*) dla obu scenariuszy kodowania wykorzystano zaawansowany algorytm obiektywnej oceny jakości PEAQ (ang. *Perceptual Evaluation of Audio Quality*) opracowany przez firmę *Opticom*. Na rys. 4.3 oraz 4.4 przedstawiono wyniki przeprowadzonych badań odpowiednio dla nagrań zawierających składniki sinusoidalne niemodulowane oraz modulowane.



Rys. 4.3 Oceny ODG uzyskane dla nagrań dźwiękowych zawierających głównie składowe sinusoidalne niemodulowane.



Rys. 4.4 Oceny ODG uzyskane dla nagrań dźwiękowych zawierających głównie składowe sinusoidalne modulowane.

Z analizy wyników zaprezentowanych na rys. 4.3 wynika, iż dla nagrań zawierających składowe sinusoidalne niemodulowane, średnia ocen ODG uzyskanych przy stosowaniu algorytmu FTM jest o 0,47 punktu większa niż w przypadku stosowania metody UM. Niemniej w obu wypadkach stopień degradacji ODG jest w okolicach –1,5 co wskazuje, iż wprowadzane zniekształcenia są niesłyszalne dla słuchacza. Średnia różnica w ocenach ODG wynosi natomiast około 1,4 punktu na korzyść metody FTM dla nagrań muzycznych zawierających głównie składowe sinusoidalne modulowane częstotliwościowo. Biorąc powyższe pod uwagę należy stwierdzić, iż w przypadku gdy nagrania zawierają głównie składowe tonalne modulowane, model psychoakustyczny określa próg słyszenia w sposób bardziej wiarygodny gdy wykorzystany jest estymator tonalności FTM. Ponadto stopień degradacji jakości nagrań zawierających komponenty tonalne modulowane przekracza wartość –2,5 co oznacza, iż wprowadzane zniekształcenia zaczynają być słyszalne.

Na podstawie przeprowadzonych badań poczynione dwie następujące obserwacje:

- Próg słyszenia estymowany przez model psychoakustyczny zdefiniowany w standardzie MPEG jest nieco niższy niż rzeczywisty próg słyszenia;
- Kiedy szum kwantyzacji przekracza próg słyszenia w kilku podpasmach, wprowadzane zniekształcenia są słyszalne jedynie jako niedokuczliwe ograniczenia dynamiki sygnału.

Biorąc pod uwagę powyższe spostrzeżenia dalsze eksperymenty mające na celu wykazanie przewagi metody FTM nad innymi prowadzone były przy założeniu, iż kodek nie dysponuje odpowiednim zasobem bitów do zakodowania sygnału w taki sposób by wprowadzany szum kwantyzacji pozostawał poniżej progu słyszenia.

5 MODUŁ PNS

Stosowanie techniki PNS wymaga określenia, które z podpasm sygnału mogą zostać zsyntetyzowane w dekoderze z wykorzystaniem lokalnego generatora Gaussowskiego szumu białego. Możliwość wykorzystania algorytmu UM jako podstawy detektora pasma szumowych jest ograniczona. Powodem jest niska wiarygodność estymacji tonalności zapewniana przez ten algorytm w przypadku nagrań zawierających modulowane komponenty tonalne.

5.1 DETEKCJA PODPASM SZUMOWYCH

Zarówno kodek MP3 jak i AAC przeprowadzają kodowanie współczynników MDCT zgrupowanych w tak zwane *scale–factor bands*. Z tego powodu detektor podpasm szumowych operuje na zdefiniowanych w standardach podpasmach sygnału wykorzystując trzy parametry:

1. Indeks tonalności określany na podstawie $c''[k]$. Podpasma zawierające komponenty tonalne nie mogą być kodowane zgodnie z techniką PNS.

$$t[m] = \max\left\{t_b\left[b_{\text{low}}[m]\right], \ldots, t_b\left[b_{\text{high}}[m]\right]\right\} \tag{5.1}$$

   gdzie m=1, 2, …, 49 (zakładając F_s=44100 lub 48000 Sa/s) jest numerem podpasma, $t_b[b]$ jest indeksem tonalności wykorzystywanym przez model psychoakustyczny MPEG, b_{low}[m] oraz b_{high}[m] odpowiadają indeksom podpasm wykorzystywanych przez model psychoakustyczny MPEG zawierającym się w danym podpaśmie *scale–factor*.

2. Miara płaskości podpasm *scale–factor*. Podpasma zawierające szum kolorowy nie powinny być kodowane zgodnie z techniką PNS. Wykorzystana miara ta jest w pewnym stopniu podobna do miary SFM i określana jest najpierw dla poszczególnych prążków widma następująco:

$$z_{\text{bin}}[k] = 20\log\left(\frac{r_a[k]}{r_g[k]}\right) \tag{5.2}$$

   gdzie $r_\text{a}[k]$ oraz $r_\text{g}[k]$ odpowiadają widmu amplitudowemu poddanemu filtracji filtrem uśredniającym z wykorzystaniem odpowiednio średniej arytmetycznej i geometrycznej. W następnym kroku obliczana jest miara płaskości podpasm *scale–factor* jako średnia arytmetyczna $z_{\text{bin}}[k]$.

3. Dewiacja standardowa energii w podpasmach. Podpasma sygnału widma określonego dla 2048 próbek sygnału wejściowego zawierające szum niestacjonarny nie powinno być kodowane zgodnie z techniką PNS. Odchylenie standardowe energii w podpasmach określane jest z wykorzystaniem 8 kolejnych widma obliczanych dla ramek sygnału zawierających 256 próbek sygnału na podstawie zależności:

$$\sigma[m]=\sqrt{\frac{1}{8}\sum_{l_\mathrm{s}=1}^{8}\left(e_{nl}[m,l_\mathrm{s}]-\overline{e_{nl}}[m,l_\mathrm{s}]\right)} \qquad (5.3)$$

gdzie $e_{nl}[m,l_\mathrm{s}]$ oznacza energię m-tego podpasma obliczonego dla widma l_s=0, 1, …, 7 uzyskanego dla krótkiej ramki sygnału (256 próbek), $\overline{e_{nl}}[m,l_s]$ jest średnią energią w danym podpasmie obliczoną dla 8 kolejnych widma.

Podpasmo *scale-factor* może zostać zakodowane zgodnie z techniką PNS jeśli spełnione jest następujący warunki:

$$t[m]\le t_\mathrm{thd} \wedge z_\mathrm{sfb}[m]<2 \wedge \sigma[m]<6\,. \qquad (5.4)$$

gdzie t_thd = 0,01 dla trybów pracy kodeka z wysoką przepływnością bitową, oraz t_thd = 0,05 dla trybów pracy z niską przepływnością bitową.

5.2 BADANIE MODUŁU PNS

W celu przeprowadzenia badania modułu PNS wybrane nagrania muzyczne przetwarzano w taki sposób, by podpasma spełniające kryterium podane w (5.4) zastępowane były syntetycznym szumem o energii odpowiadającej energii podpasma oryginalnego. Dla każdego nagrania określano wpływ każdego z parametrów wykorzystanych w detektorze pasm szumowych na ostateczną decyzję dotyczącą tego czy możliwe jest zastosowanie techniki PNS. W tym celu obliczano parametr

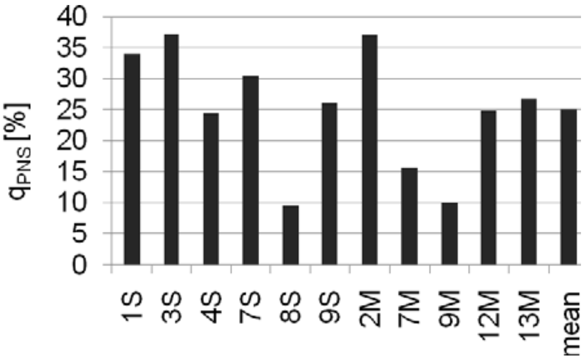
$$q_* = \frac{n_{\mathrm{sfb}*}}{n_\mathrm{asfb}}100\,\% \qquad (5.5)$$

gdzie n_asfb jest całkowitą ilością podpasm w sygnale, które były brane pod uwagę, $n_{\mathrm{sfb}*}$ określa ilość podpasm spełniających jeden z warunków cząstkowych: $t[m]<t_\mathrm{thd}$, $z_\mathrm{sfb}[m]<2$, $\sigma[m]<6$ lub wszystkie jednocześnie. Gwiazdka w równaniu (5.5) jest odpowiednio zastępowana symbolami t, z_sfb, σ lub PNS. Na rys. 5.1 przedstawiono q_PNS dla poddanych analizie nagrań muzycznych podzielonych na grupy zawierające głównie komponenty tonalne modulowane (np. 2M, 7M, itd.) oraz komponenty tonalne niemodulowane (np. 1S, 3S, itd.).

Stosunek q_PNS dla podpasm sygnału leżących powyżej częstotliwości 3,1 kHz silnie zależy od zawartości nagrania poddanego analizie i waha się od około 9 do 37% dla wybranych nagrań. Dla wszystkich badanych nagrań średnia wartość q_PNS wynosi 25%.

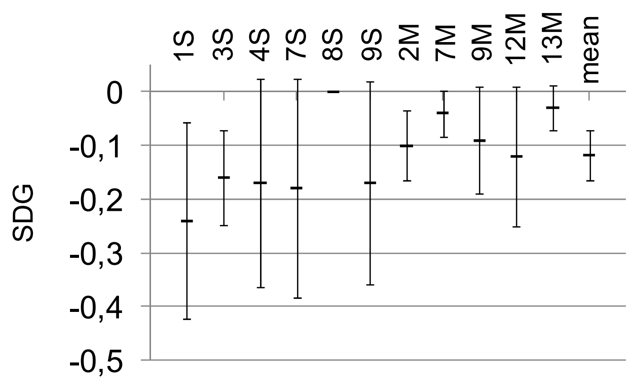
Zasób bitów oszczędzony dzięki wykorzystaniu techniki PNS może zostać wykorzystany do kodowania pasm tonalnych.

W celu określenia wpływu stosowanie techniki PNS na subiektywną jakość kodowania nagrań muzycznych, przeprowadzono pilotażowe testy odsłuchowe zgodnie z procedurą ITU-R BS.1116. Każdy z ekspertów miał do dyspozycji 3 nagrania muzyczne w każdym z zadań. Pierwsze nagranie było wskazanym (jawnym) sygnałem oryginalnym (referencyjnym). Wśród dwóch kolejnych nagrań, jedno było sygnałem referencyjnym (niejawnym), a drugie nagraniem w którym pasma szumowe zostały zakodowane zgodnie z techniką PNS.



Rys. 5.1 Stosunek wyrażony poprzez iloraz ilości podpasm dla których zastosowano technikę PNS do wszystkich rozpatrywanych podpasm sygnału dla wybranych nagrań muzycznych

Zadaniem eksperta było w pierwszej kolejności wskazanie które z dwóch nagrań jest nagraniem oryginalnym, a następnie określenie stopnia degradacji jakości SDG (*Subjective Difference Grade*) drugiego z nagrań w skali od 0 (brak degradacji jakości) do 5 (bardzo duża utrata jakości) w porównaniu do oryginału. Na rys. 5.2 przedstawiono średnie oceny SDG dla badanych nagrań wraz z przedziałami ufności na poziomie 95%.

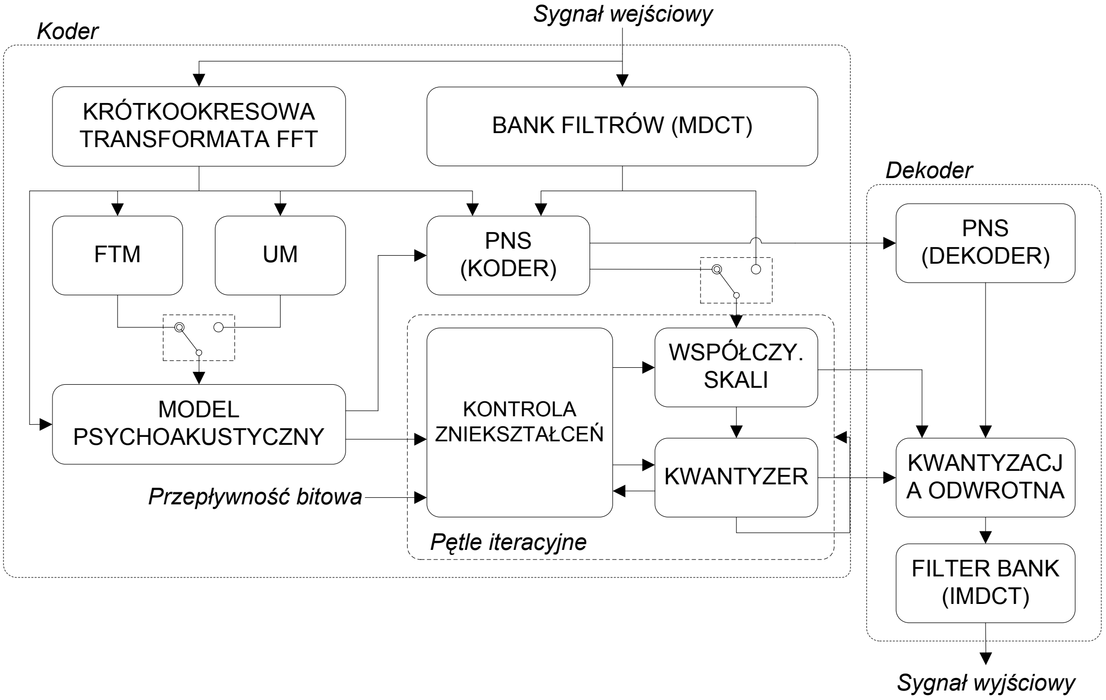


Rys. 5.2 Oceny SDG uzyskane przy stosowaniu modułu PNS wykorzystującego algorytm FTM

Z analizy rys. 5.2 wynika, iż średnia wartość SDG dla żadnego z nagrań nie spada poniżej –0,25 punktu, co oznacza, iż utrata jakości jest właściwie niesłyszalna. Potwierdza to, iż detektor pasm szumowych wykorzystujący algorytm FTM funkcjonuje prawidłowo.

6 WPŁYW METODY ESTYMACJI TONALNOŚCI NA JAKOŚĆ KODOWANIA

Wszystkie eksperymenty przeprowadzono w systemie przedstawionym na rys. 6.1 mogącym pracować w każdym z pięciu opisanych w tab. 1.

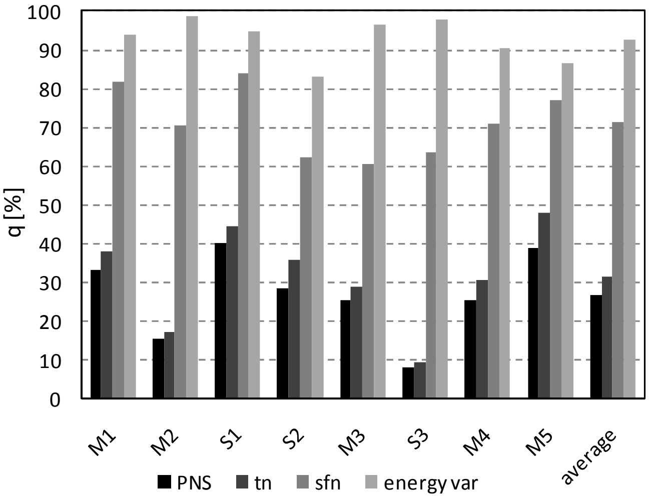


Rys. 6.1 Schemat blokowy eksperymentalnego kodeka

Tabela 6.1 Tryby pracy eksperymentalnego kodeka sygnałów fonicznych

Tryb	Moduł PNS	Estymacja tonalności	Opis
1	Nieaktywny	UM	Wszystkie pasma sygnału kodowane są na podstawie progu słyszenia określonego przez model psychoakustyczny wykorzystujący metodę UM.
2	Nieaktywny	FTM	Wszystkie pasma sygnału kodowane są na podstawie progu słyszenia określonego przez model psychoakustyczny wykorzystujący metodę FTM.
3	Nieaktywny	–	Wszystkie pasma sygnału kodowane są na podstawie progu słyszenia określonego przez model psychoakustyczny nie wykorzystujący żadnej metody estymacji tonalności. (wszystkie komponenty widmowe traktowane są jako szumowe)
4	Aktywny	UM	Pasma szumowe kodowane są z wykorzystaniem techniki PNS. Pasma szumowe kodowane są na podstawie progu słyszenia określonego przez model psychoakustyczny wykorzystujący metodę UM. Zasób bitów oszczędzony dzięki stosowaniu PNS wykorzystany jest do kodowania pasm tonalnych.
5	Aktywny	FTM	Pasma szumowe kodowane są z wykorzystaniem techniki PNS. Pasma szumowe kodowane są na podstawie progu słyszenia określonego przez model psychoakustyczny wykorzystujący metodę FTM. Zasób bitów oszczędzony dzięki stosowaniu PNS wykorzystany jest do kodowania pasm tonalnych.

Wybrano grupę nagrań dźwiękowych zawierających głównie komponenty tonalne niemodulowane oznaczone jako S1 – S3 oraz nagrania zawierające komponenty tonalne modulowane M1 – M5. Każde z nagrań zakodowano i zdekodowano z wykorzystaniem systemu przedstawionego na rys. 6.1 pracującego w trybach 1, 2, 3 oraz 5 dla dwóch docelowych przepływności bitowych: 48 oraz 64 kbps. Wykorzystywany system kodowania nie zawiera modułu kodowania bezstratnego (kodowanie Hamminga) i nie jest formowany wyjściowy strumień bitów. Jest to uzasadnione tym, iż praca poświęcona jest metodą stratnego kodowania sygnałów fonicznych, a zaproponowana architektura pozwala na badanie właśnie tej części algorytmu kodowania. Z tego względu przepływności bitowe były szacowane na podstawie entropii perceptualnej określanej zgodnie z zależnością podaną w standardzie MPEG. Opracowano również metodę dystrybucji bitów w przypadku gdy zasób bitów jest niewystarczający do utrzymania szumu kwantyzacji poniżej progu słyszenia. Dla trybów pracy kodeka, w których moduł PNS jest aktywny określono parametry q_t, q_{zsfb}, q_σ oraz q_{PNS}. Wyniki tych badań przedstawiono na rys. 6.2.



Rys. 6.2 Efektywność modułu PNS dla t_{thd}=0,01

Z analizy rys. 6.2 wynika, iż tonalność podpasm sygnału ma kluczowe znaczenie w procesie detekcji podpasm kodowanych zgodnie z techniką PNS. Chociaż warunek dla parametru σ jest spełniony dla średnio ponad 90% przypadków, ma on istotne znaczenie w przypadku gdy nagranie zawiera wiele dźwięki artykułowane przez instrumentalistów i wokalistów z dużą ekspresją. Ma to miejsce w przypadku nagrania S2.

6.1 TESTY ODSŁUCHOWE

Testy odsłuchowe przeprowadzono zgodnie z zaleceniem ITU-R BS.1534 definiującym procedurę MUSHRA (MUlti Stimulus with Hidden Reference and Anchor). Testy przeprowadzono z wykorzystaniem interfejsu graficznego funkcjonującego w środowisku MATLAB, karty dźwiękowej posiadającej dynamikę konwersji cyfrowo-analogowej ponad 100 dB oraz wysokiej jakości słuchawek studyjnych AKG K240 MKII. W testach wzięło udział 17 ekspertów. Każdy ekspert przechodził fazę szkolenia, w której miał możliwość zapoznania się z interfejsem graficznym, mógł posłuchać wszystkie wykorzystywane w dalszej części testu nagrania, a także otrzymywał wskazówki co do sposobu dokonywania oceny. Przed przejściem z fazy nauki do fazy testu następowała przerwa. W każdym zadaniu testowym ekspert oceniał 8 nagrań muzycznych przetworzonych przez kodek pracujący w czterech wybranych trybach oraz dla dwóch przepływności bitowych. Oprócz wspomnianych nagrań ocenie podlegało również nagranie ograniczone pasmowo do 3,5 kHz – tak zwany *Anchor* oraz ukryte nagranie referencyjne. Umieszczanie nagrań na liście do odsłuchania odbywała się w sposób losowy. Jawnie wskazany sygnał referencyjny mógł być odtwarzany dowolną ilość razy podobnie jak wszystkie pozostałe nagrania. Na rys. 6.3 przedstawiono interfejs systemu pozwalającego na odsłuchiwanie i ocenę nagrań dźwiękowych.

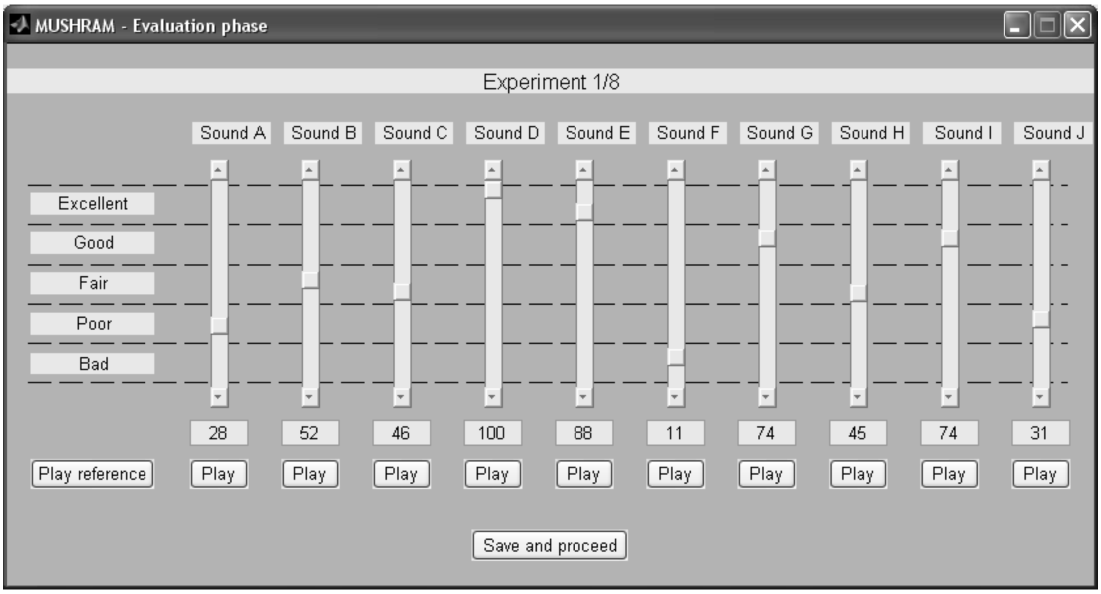
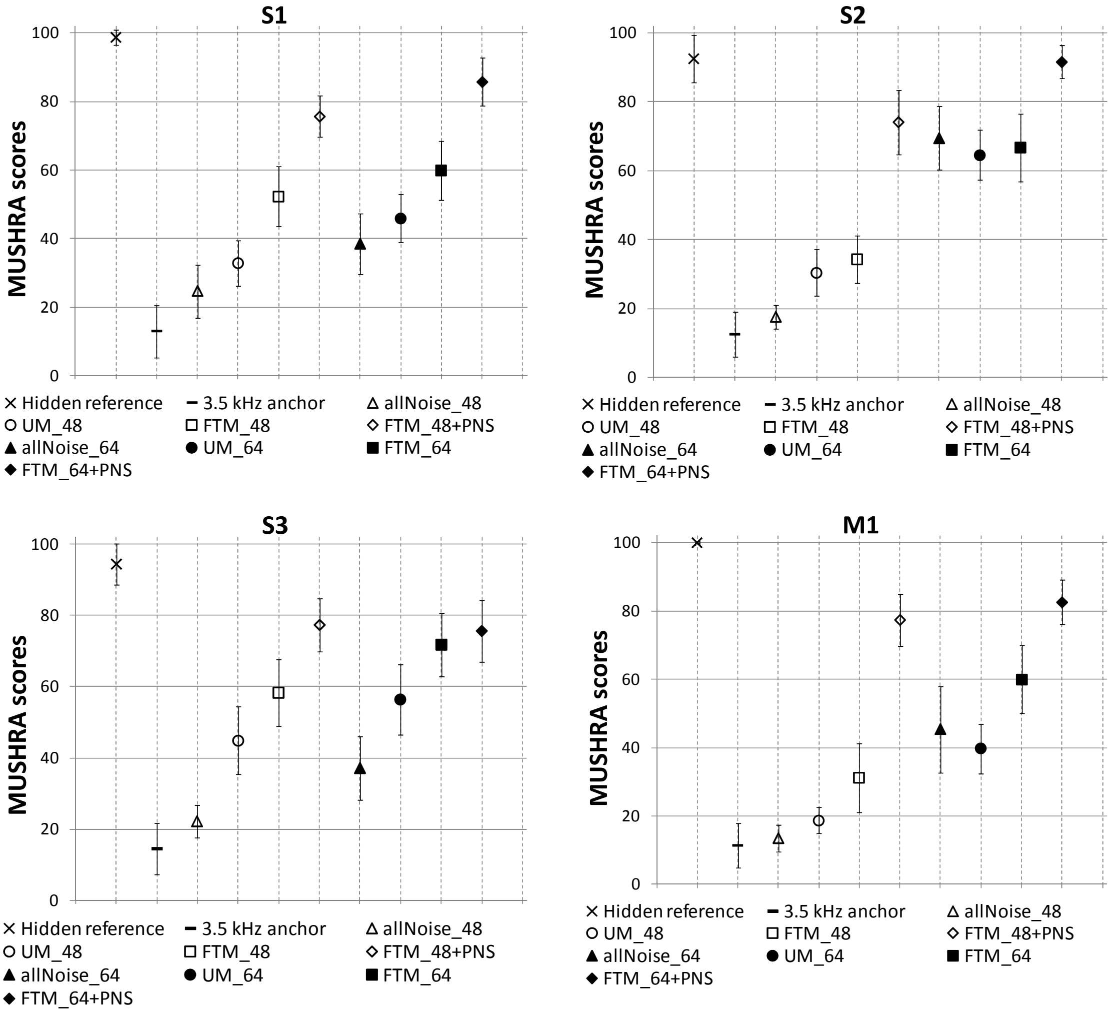
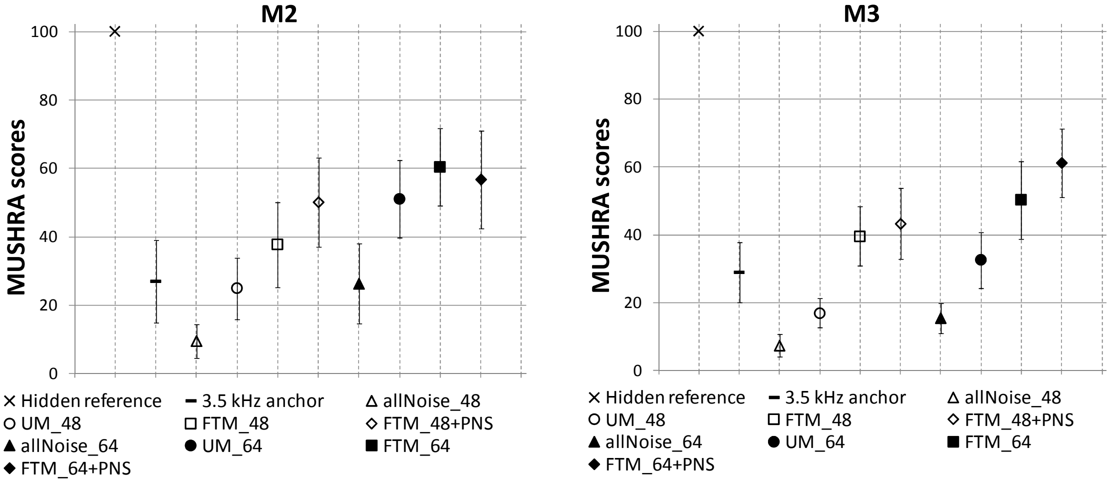


Fig. 6.3 Interfejs aplikacji do oceny jakości nagrań zgodnie z metodą MUSHRA

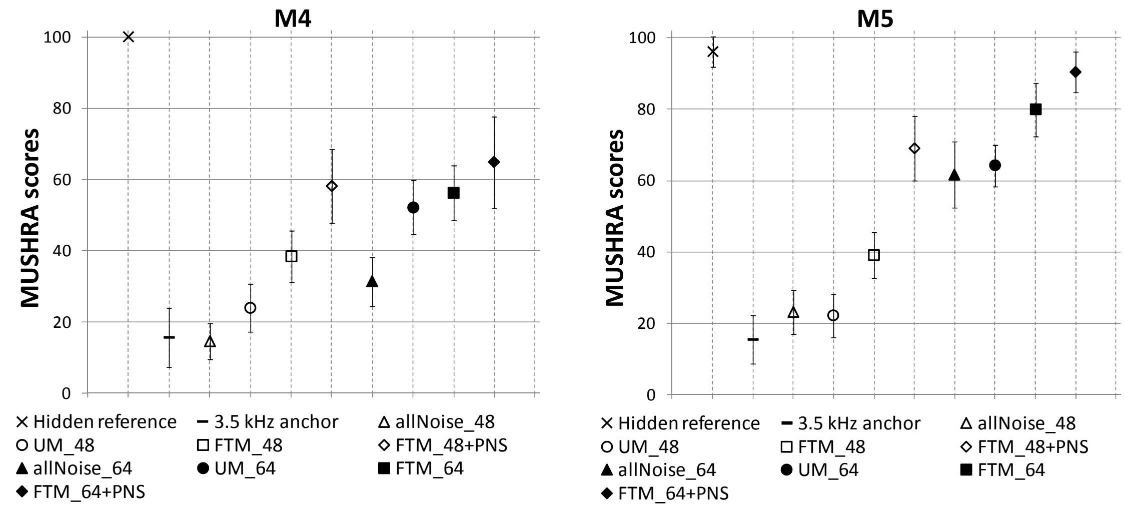
Ponieważ wśród ocenianych nagrań muzycznych znajdował się również ukryty sygnał referencyjny możliwa była weryfikacja wiarygodności ekspertów na podstawie zdolności do prawidłowego wskazywania sygnału referencyjnego. Ponadto przeprowadzono analizę statystyczną dostarczającą informacji o tym, w jaki sposób oceny przydzielane przez danego eksperta różniły się od średnich ocen pozostałych ekspertów. Dwoje ekspertów miało wyraźnie obniżoną zdolność wskazywania ukrytych sygnałów referencyjnych, a także ich oceny znacząco różniły się od ocen pozostałych ekspertów. Z tego względu oceny przez nich przydzielone nie były brane pod uwagę przy dalszej analizie. Średnie oceny jakości w skali MUSHRA wraz z przedziałami 95% ufności dla każdego z nagrań przedstawiono na rys. 6.4, 6.5 oraz 6.6.



Rys. 6.4 Wyniki testów odsłuchowych MUSHRA (nagrania S1 – S3 oraz M1)



Rys. 6.5 Wyniki testów odsłuchowych MUSHRA (nagrania M2 oraz M3)



Rys. 6.6 Wyniki testów odsłuchowych MUSHRA (nagrania M4 oraz M5)

Przedstawione wyniki testów odsłuchowych pozwalają poczynić następujące spostrzeżenia:

- Jakość kodowania z wykorzystaniem modelu psychoakustycznego zintegrowanego z zaproponowanym w rozprawie estymatorem tonalności FTM jest o 10 do 20 (średnio o 15) punktów w skali MUSHRA wyższa niż w przypadku stosowania metody UM dla nagrań zawierających modulowane komponenty tonalne. Podobna tendencja jest widoczna w przypadku dwóch spośród trzech nagrań zawierających niemodulowane komponenty tonalne.
- Model psychoakustycznym wykorzystujący metodę UM funkcjonował zupełnie nieefektywnie dla nagrań M1 oraz M5, gdzie uzyskana ocena jest podobna do oceny otrzymanej w przypadku gdy nie zastosowano żadnej metody estymacji tonalności. W przypadku stosowania metody FTM jakość wspomnianych próbek jest znacząco wyższa.
- Oceny uzyskane dla nagrań zakodowanych z przepływnością 48 kbps i aktywnym modułem PNS są równie wysoki lub nawet wyższe niż w przypadku kodowania tych samych próbek z przepływnością 64 kbps i nieaktywnym modułem PNS.
- Średnia ocena jakości kodowania w przypadku stosowania modułu PNS dla przepływności 48 kbps jest o 22 punkty MUSHRA wyższa niż wtedy gdy moduł PNS jest nieaktywny. W przypadku przepływności równej 64 kbps, zysk ze stosowania metody PNS jest równy średnio 12 punktów w skali MUSHRA.

Wynika stąd, że metoda ta jest bardziej efektywny w przypadku gdy kodek pracuje w trybie niskich przepływności bitowych.

7 PODSUMOWANIE

Przedstawione wyniki badań wskazują, iż autorowi udało się osiągnąć wszystkie założone cele główne oraz dodatkowe rozprawy doktorskiej. Do autorskiego wkładu w dziedzinę kodowania sygnałów fonicznych można zaliczyć:

- algorytm estymacji tonalności komponentów widmowych pozwalający na wiarygodną estymację tonalności zarówno komponentów tonalnych niemodulowanych jak i modulowanych. Ponadto zaproponowano dwie metody jego sprzężenia z modelem psychoakustycznym stosowanym w kodekach MP3 oraz AAC,
- detektor podpasm sygnału zawierających wyłącznie składowe szumowe, które mogą być kodowane zgodnie z techniką PNS. Detektor ten wykorzystuje zaproponowany w rozprawie estymator tonalności, który stanowi integralną część modelu psychoakustycznego. Z tego względu jego złożoność obliczeniowa jest niższa niż w przypadku gdy stosowana jest dedykowana metoda detekcji podpasm szumowych,
- wyniki badań pozwalające określić wpływ wykorzystanego estymatora tonalności na wynikową jakość kodowania sygnałów fonicznych z wykorzystaniem metod perceptualnych. Wyniki te wskazują na to, iż wybór metody estymacji tonalności ma znaczący wpływ na jakość zapewnianą przez dany system kodowania. Ponadto wykazano, że stosowanie zaproponowanego w rozprawie estymatora tonalności komponentów widmowych w miejsce estymatora zdefiniowanego w standardzie MPEG pozwala na ograniczenie zniekształceń wprowadzanych w procesie kodowania perceptualnego.

Biorąc pod uwagę osiągnięte cele można stwierdzić, iż obie tezy rozprawy zostały udowodnione.