Modulation vocoder for analysis, processing and synthesis of audio signals with application to frequency selective pitch transposition

Von der Fakultät für Elektrotechnik und Informatik
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des akademischen Grades

Doktor-Ingenieur

genehmigte

Dissertation

von

Dipl.-Ing. Sascha Disch

geboren am 09. November 1968 in Freiburg im Breisgau

2011

1. Referent:	Prof. Dr.-Ing. Jörn Ostermann
2. Referent:	Prof. Dr.-Ing. habil. Udo Zölzer
Tag der Promotion:	17.03.2011

Vorwort

Diese Dissertation entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Laboratorium für Informationstechnologie (LfI) der Gottfried Wilhelm Leibniz Universität Hannover. Besonderen Dank möchte ich Herrn Prof. Dr.-Ing. Jörn Ostermann aussprechen für die Übernahme des Hauptreferates und die Betreuung der Arbeit, insbesondere für die stete Begleitung der vorliegenden schriftlichen Ausfertigung. Herrn Prof. Dr.-Ing. Bernd Edler möchte ich danken für die persönliche und fachliche Betreuung meiner Tätigkeit am LfI. Herrn Prof. Dr.-Ing. habil. Udo Zölzer danke ich für die freundliche Übernahme des Koreferates.

Mein Dank gilt ebenfalls dem Fraunhofer Institut für Integrierte Schaltungen (IIS) in Erlangen - stellvertretend seien hier Herr Prof. Dr.-Ing. Heinz Gerhäuser, Herr Dr.-Ing. Bernhard Grill, Herr Dipl.-Ing. Harald Popp und Herr Prof. Dr.-Ing. Jürgen Herre genannt - das mich in meinem Vorhaben unterstützt hat, eine Dissertation anzufertigen, und dieses im Rahmen eines Forschungsauftrages ermöglicht hat.

Bedanken möchte ich mich auch bei all meinen Kollegen am Laboratorium für Informationstechnologie (LfI) und des Instituts für Informationsverarbeitung (TNT) für die gute Arbeitsatmosphäre, insbesondere Herrn Dipl.-Ing. Marco Munderloh für viele nette fachliche und persönliche Gespräche und nicht zuletzt kompetente Hilfe bei Rechnersystemfragen. Für umfängliche Einsichten in physikalische Zusammenhänge jeglicher Art geht mein Dank an Herrn Dr.-Ing. Nikolaus Meine.

Bei den Mitarbeitern der Abteilung *Audio* und der Abteilung *Multimedia Echtzeitsysteme* am Fraunhofer IIS möchte ich mich für die gute Zusammenarbeit in der Projektarbeit bedanken. Nicht zu vergessen sind auch die zahlreichen freiwilligen Testhörer des IIS, die es durch ihre musikalische und technische Expertise ermöglicht haben, die Ergebnisse der vorliegenden Arbeit zuverlässig zu quantifizieren. Ihnen allen gilt mein Dank.

Ein ganz herzliches Dankeschön gebühren Frau Dr. techn. Cornelia Falch und Frau Oktavia Ostermann, die die vorliegende Arbeit hinsichtlich englischer Ausdrucksweise zur Korrektur gelesen haben.

Vor allem aber möchte ich mich bei meiner Frau Kirsa für die mit großer Selbstverständlichkeit gewährte tägliche Unterstützung und insbesondere für die Bewältigung aller zusätzlichen Mühen bedanken, die mein jahrelanges wöchentliches Berufspendeln zwischen Hannover und Erlangen mit sich gebracht hat. Einen besonders lieben Dank sage ich auch meiner Tochter Ellen Vera, die bereits an ihrem ersten Lebenstag und an vielen weiteren Tagen immer geduldig auf mich gewartet hat.

Zusammenfassung

In der vorliegenden Dissertation wird ein gehörangepasstes Analyseverfahren entworfen, das Audiosignale blockweise in Teilbandkomponentensätze aus sinusförmigen Trägersignalen und zugehörigen Amplituden- und Frequenzmodulationen zerlegt. Die Zerlegung erfolgt derart, dass die Teilbandkomponenten signaladaptiv an lokalen spektralen Schwerpunkten ausgerichtet werden. Dadurch kann diesen Komponenten eine direkte Interpretation zugeschrieben werden: Die Trägersignale bezeichnen die mittlere Tonhöhe, die durch den Spektralbeitrag der jeweiligen Komponenten bei einem Hörer hervorgerufen wird, die Amplituden- und Frequenz-Modulationen sind bei Frequenzen unter 20 Hz durch die musikalischen Begriffe *Tremolo* und *Vibrato* charakterisiert, bei höheren Frequenzen durch *Rauigkeit*. Passend zur vorgeschlagenen Modulationsanalyse wird ein artefaktarmes Syntheseverfahren entwickelt.

Herkömmliche Verfahren zur Teilbandmodulationanalyse verwenden Filterbänke mit einer, wie in dieser Arbeit gezeigt wird, ungeeigneten festen Bandaufteilung oder lassen die Frage offen, wie eine geeignete Bandaufteilung zwecks nachfolgender Modulationszerlegung beschaffen sein muss, um eine direkt interpretierbare und somit manipulierbare Modulationsdarstellung zu erhalten. Oftmals ist auch keine Synthesemethode zur Rückgewinnung eines Audiosignals angegeben.

Das innovative Potential der signaladaptiven und gehörangepassten Zerlegung wird aufgezeigt, indem das vorgeschlagene Verfahren dahingehend konfiguriert wird, eine neuartige selektive Transponierung einzelner Frequenzbereiche in polyphonen Audiosignalen vorzunehmen. Eine solche Anwendung verändert nachträglich das Tongeschlecht von Audioaufnahmen, beispielsweise von Dur nach Moll.

Die mit dem vorgeschlagenen Verfahren erreichbare und bei der Anwendung des Verfahrens zur selektiven Transponierung von Tonhöhen erzielbare subjektive Audioqualität wird durch Hörtests evaluiert. Die Audioqualität bei selektiver Tonhöhentransponierung wird dabei von Testhörern im Bereich „ausreichend” bis „gut” bewertet, während das reine Analyse/Synthese-Verfahren Audioqualitäten von „gut” bis „sehr gut” erreicht.

Abschliessend wird die vorgeschlagene selektive Transponierung mit den erzielbaren Resultaten des kommerziellen Computerprogrammes „Melodyne editor” von „Celemony” verglichen, das gegen Ende der Entstehung dieser Arbeit als Marktneuheit verfügbar wurde. Das vorgeschlagene Verfahren ist dem Vergleichsverfahren deutlich im Qualitätsaspekt „Transponierung von Melodie und Akkorden” überlegen, während das Vergleichsverfahren mehrheitlich unter des Aspekt „Erhaltung der Klangfarbe” bevorzugt wird.

Stichworte: Audio, Modulation, Transponierung, Polyphonie

Abstract

In this thesis, a perceptually adapted analysis method is devised which decomposes audio signals in a block-wise manner into sets of subband components, each of which is further decomposed into a sinusoidal carrier and its associated amplitude and frequency modulation. The decomposition is configured such that the subband components are aligned with spectral local centers of gravity. Thereby, the components relate to a straight forward interpretation: the carrier signals represent the mean pitch sensation that is perceived by a listener due to the spectral contribution of that component, the amplitude- and frequency modulation correlate at frequencies below 20 Hz with the musical terms *tremolo* and *vibrato*, or at higher frequencies with the sensation of auditory roughness. Fittingly, a synthesis method having low artifacts is proposed.

Conventional methods for subband modulation analysis employ filterbanks with fixed subband positions which, via this work, will be shown to be inadequate. Other publications leave an essential question unanswered: how a suitable partitioning into bands for a subsequent modulation analysis should be done in order to yield a modulation representation that is interpretable and thus manipulable in a direct way. Also, often no synthesis method is described that can retrieve an audio signal from a certain given modulation representation.

The innovative potential of the signal adaptive and perceptually adapted decomposition is demonstrated through the application of the proposed method to a novel frequency selective pitch transposition scheme for polyphonic audio signals. Such an application retroactively changes the key mode of audio recordings, e.g. from a major key to minor key.

The subjective audio quality that can be obtained by said method and its application to selective pitch transposition is evaluated by listening tests. The audio quality of selective pitch transposition is scored by the listeners in a range spanning from «satisfactory» to «good», whereas the perceptual quality of the pure analysis/synthesis scheme alone extents from «good» to «excellent».

Finally, the proposed selective pitch transposition scheme is compared with results obtained by applying the commercial computer program «Melodyne editor» by «Celemony», which became newly available on the market close to the time of finalization of this thesis. The proposed method is clearly preferred in terms of the perceptual quality aspect «melody and chords transposition», while the commercial program is favored by the majority with regard to the aspect «timbre preservation».

Keywords: audio, modulation, transposition, polyphony

Contents

Abbreviations

AAC	advanced audio coding
AM	amplitude modulation
BWE	bandwidth extension
CA	constant amplitude
CASA	computational auditory scene analysis
CB	critical bandwidth
COG	centers of gravity
DESA	discrete-time energy separation algorithm
DFT	discrete Fourier transform
DNA	direct note access
DTF	dynamic tracking filters
ERB	equivalent rectangular bandwidth
ES	envelope shaping
ESA	energy separation algorithm
EWAIF	envelope weighted average of instantaneous frequency
f0	fundamental frequency
FM	frequency modulation
GUI	graphical user interface
HILN	harmonic and individual lines plus noise
HL	harmonic locking
HTD	Hilbert transform demodulation
IDFT	inverse discrete Fourier transform
IF	instantaneous frequency
IHC	inner hair cells
IIR	infinite impulse response
ITU	International Telecommunication Union
IWAIF	intensity weighted average instantaneous frequency
JND	just noticeable difference
JNDF	just noticeable difference of frequency
JNDL	just noticeable difference of level
LPC	linear prediction coefficients
LPSD	linear prediction in the spectral domain
MAS	modulation analysis and synthesis
MIDI	musical instrument digital interface
MODVOC	modulation vocoder
MPEG	Moving Pictures Expert Group

MSE	mean square error
MUSHRA	MUltiple Stimuli with Hidden Reference and Anchor
OHC	outer hair cells
OLA	overlap-add
PIF	linear prediction in the spectral domain
POCS	projection on convex sets
psd	power spectral density
SAM	sinusoidally amplitude modulated pure tone
SB-AMS	sub-band amplitude modulation spectrum
SSM	sound source modeling
STFT	short time fourier transform
TC	tuning curves
TDAC	time domain aliasing cancellation
TNS	temporal noise shaping
VOCODER	voice encoder
VODER	voice operation demonstrator

1 Introduction

1.1 History and usage of audio effects

In modern music productions, audio effects are an integral part of the trademark sound of a certain band, a disc jockey (DJ) or a producer. Historically, the invention of new sound effects was driven by playful or originally unintended use of new technical appliances, starting with electro-mechanical equipment like amplifiers, analog disc recorders or tape recorders and being continued via purely analog to digital signal processing. In the early days, especially the realization of amplitude and frequency modulation effects by post processing in order to «thicken» the sound of a given electro-mechanical instrument, e.g. organ, constituted a major challenge [9].

Innovators, like the well known guitarist Lester W. Polsfuss (known as «Les Paul»), started experimenting in the 1950s with electric guitar pickups and disc recorders, and often included audio effects originating from his inventions in his own music recordings [9].

In the 1960s, bands like «The Beatles» started to explore the new possibilities that opened up due to the invention of multi-track tape recording. Effects like *pitch change*, *echo*, *chorus*, *flanging*, *backward recording* and *time reversed echo* or *reverb* could now be realized by the creative use of these tape recorders.

In the 1970s, electro-mechanical effects were substituted or amended by purely electric analog effect circuitry. Now, the intricate studio effects were available to almost everybody and could easily be integrated in live performances.

The 1980s witnessed the emerge of computer hardware based digital effects, once more augmenting the possibilities of sound manipulation. On one hand, the heritage analog effects were now emulated by digital signal processing, on the other hand companies like Eventide with the «Harmonizer» or MXR introducing the «Pitch Transposer» offered highly sophisticated hardware-based digital effects. For example, the well known hit single «Owner of a lonely heart» by the band «Yes» featured a remarkable guitar solo which was played through the MXR Pitch Transposer and mixed to the original sound.

In 1990, music production changed from essentially recording onto tape towards digital hard disk recording. This trend also affected the way how music was actually produced. Increasingly, studio appliances and dedicated hardware effect boxes were virtualized and integrated into computer recording software suites. Even real music instruments were partly superseded by pre-recorded sampling libraries, mainly for cost cutting reasons, e.g. for traditionally sumptuous orchestral parts.

Contrary to common belief, most modern classic music recordings are also subjected to complex post processing in a sense that the final recording might be composed of different outtakes that are cut, adapted and blended during the record production process. Small musical and technical flaws, like poor timing or intonation are often corrected by computer based editing tools, unwanted background noise is removed.

At present, audio effects are increasingly enhanced by incorporating knowledge about the (local) semantic content of the music signal to be processed. This implies the application of a certain effect in a context adaptive and selective manner. For example, contemporary pitch shifters transpose the notes according to a predetermined musical scale («Auto-Tune» by «Antares» and others), dedicated real-time voice processors generate harmony choir voices that match the chords of e.g. an underlying guitar accompaniment («Vocalist» by «DigiTech»).

While the use of audio effects is most noticeable in modern pop music production, effects are nevertheless applied in some not-so-obvious contexts. For instance, for radio broadcast commercials, the voice of the speaker is often accelerated artificially in order to cut down expensive broadcast time. Sometimes, the pitch is lowered to make the speakers voice sound more appealing and pleasant.

As another example, American cinema films are usually shot with a frame rate of 24 pictures per second. When converted to the European television format PAL/SECAM, the film is simply played back faster at the rate of 25 frames per second resulting in a pitch shift of the accompanying audio sound track of approximately 4%. Since this significantly changes the timbre especially of male voices, often a pitch shifter is applied to the audio track in order to restore the original quality of the voices.

Also, the voice overdub for movie cartoons is often produced using audio effects. In the past, the «chipmunk» effect was achieved by tape recording and subsequent playback at multiples of the original recording speed, thereby inevitably altering the formant structure. Since the availability of digital audio effect technology, these mechanical tricks have hardly been used anymore, but substituted by computer based post-processing. Dedicated voice processing software offers effects like a mutually independent change of pitch and speed, optionally preserving the original characteristic formants. Moreover, the gender, subjective age and mood of the speaker or singer can be manipulated. This considerably lowers the costs of soundtrack production since e.g. one speaker can record the raw sound material for the voices of different cartoon characters, whereas the individual voices can be created later by application of different processing settings.

In summary, music recording, editing and production is increasingly handled by computer software in contrast to traditional methods utilizing hardware-only appliances, like mixing desks and tape recorders. Moreover, modern sound generation itself is often performed by synthesizers or by manipulation of pre-recorded pieces of audio, so-called *samples*, taken from a huge sample database. Consequently, there is an increasing demand to extensively adapt these samples to their intended new musical environment in a flexible way. In this context, advanced digital signal processing is required for the realization of *audio effects* like *pitch shifting*, *time stretching*, or *harmonization* [124], especially their time or frequency selective, signal adaptive variants. All these effects have in common that they substantially alter the musical characteristics of the original audio material under best possible preservation of subjective sound quality. In other words, these edits strongly change the musical content of the audio material but, nevertheless, are required to preserve the *naturalness* of the processed audio sample and thus ensure its *believability.*

1.2 Audio effects and polyphony

While selective pitch transposing and time scaling audio effects are already commercially available and well established for the processing of monophonic content, polyphonic content still poses great challenges to modern signal processing and thus is subject of current scientific investigation activities[1].

Essentially, some kind of source separation could be used to decompose the polyphonic content into monophonic streams, which are then separately processed [81][8]. This usually includes an initial multiple *fundamental frequency* (f0) estimation step [122] and a subsequent grouping of spectral components into several estimated source objects, each containing a fundamental and its associated harmonic overtones, in order to distinguish the spectral contributions of each tonal event. The grouping of spectral components is estimated by e.g. the evaluation of the mutual pitch ratio of the different components and by *common fate* criteria, like *common onset* and *comodulation* [10][95][110], connecting the separation task at hand to *computational auditory scene analysis* (CASA) [11][31][68]. A special variant of this principle is *sound source modeling* (SSM). Also, the separation approach is closely related to the automatic music transcription problem, which targets the automatic derivation of an abstract score notation of a given music recording [67][54]. A precise grouping into source objects is, however, a tedious, impractical and error prone method, especially if the degree of polyphony is high. Thus, the success of the method strongly depends on the musical content of the item to be processed and on the reliability of the various estimation and classification steps.

1.3 Modulation Vocoder

In contrast to the aforementioned source separation method, this thesis follows a new approach based on perceptual properties of the human auditory system. In human auditory perception, the different sound contributions contained in a certain spectral region of a polyphonic mix are fused into a single joint sonic impression given a sufficiently narrow spectral distance of these contributions. The fundamental idea is to jointly process signal components which are also perceived by humans as a sonic entity. Consequently, in this thesis, it is proposed to decompose polyphonic audio material into signal adaptive multiband components prior to a modulation analysis on each component. Most importantly, these multiband components must be aligned with spectral local *centers of gravity* (COG). In this way, the modulation parameters obtained by further analysis

[1] Throughout this thesis, the term *polyphonic* denotes the simultaneous presence of two or more tonal events at one time instant, as opposed to music having just one tonal event at any point in time (*monophonic*).

can be closely related to perceptual parameters. Each of the components is processed *as a unit*, since its content is also fused by human auditory perception. More precisely, the audio signal is decomposed into a set of signal adaptive *carrier frequencies* and their associated *amplitude modulation* (AM) and *frequency modulation* (FM). Most of all, any useful decomposition into components should establish a straight forward and intuitive relationship to audible musical parameters. In case of the decomposition proposed in this thesis, carrier frequencies are directly related to the pitch sensation of a component, coarse modulation corresponds to temporal evolution of sound (e.g. onsets, temporal development or periodic fluctuation of musical events) and, lastly, fine modulation is correlated with the human sensation of *auditory roughness*. Since, for the task of audio signal manipulation, the processed components are required to be reassembled into a perceptually pleasant, modified audio signal, a suitable synthesis method is also an important part of this thesis. If only small changes, or no changes at all, are applied to the components while processing, an acceptable synthesis method is expected to provide *transparent* (indistinguishable from the original) or at least *near-transparent* audio quality. The method proposed in this thesis is termed *modulation vocoder* (MODVOC) and considers analysis, processing and synthesis of audio signals.

Potential fields of application for modulation based audio processing are primarily advanced audio production tools for post processing in recording studios, since there is an increasing demand for rather extreme manipulation of existing audio recordings, be it for *time stretching*, *pitch transposition*, *intonation correction*, *sound morphing*, *voice gender change*, and so forth.

Further in this thesis, the application of MODVOC processing for selective pitch transposition of polyphonic audio material is demonstrated. More precisely, the manipulation of musical key and scale mode of an entire (polyphonic) audio mix is addressed. For this task, the proposed decomposition is well suited, since the carrier frequencies directly correspond to the pitch of the components, while the temporal evolution of each component is captured in its AM and FM, thereby decoupling amplitude modulation (e.g. *tremolo*) and frequency modulation (e.g. *vibrato*) nicely. Finally, the subjective audio quality of the selective pitch transposition application is evaluated by listening tests, employing high-quality rendered synthetic test signals and also natural recordings.

1.4 Overview of chapters

The thesis is grouped into 7 chapters. In Chapter 2, a comprehensive survey of the main properties of the human auditory system and human sound perception is provided. Special focus is put on pitch sensation and temporal modulation perception, since these are the main prerequisites for the subject of the work described in the following.

Next, in Chapter 3, existing different approaches to audio related modulation decomposition as well as analysis and associated synthesis methods are briefly reviewed. This summarizes the current state of the art technology in the field.

Chapter 4 presents the novel MODVOC concept for modulation analysis, modification and synthesis of arbitrary audio signals. The proposed methods and signal processing algorithms are described in detail. Moreover, conceivable basic techniques for signal modification in the modulation decomposition domain are outlined.

Chapter 5 is involved with the proposal of an advanced application of the MODVOC that substantially extends the basic signal modification techniques introduced in the preceding chapter. The application performs a pitch transposition of *selected* spectral parts of an audio signal, thereby enabling musical key and scale mode manipulation of polyphonic audio material. A special case of this class of operation is the conversion of a music signal from the original key mode of *major* to *minor* or vice versa: firstly, the implications of such an application scenario are explained from a music theory point of view and secondly, from the standpoint of digital signal processing.

In Chapter 6, a suitable listening test methodology is proposed that is tailored to assess the subjective quality of audio material that has been modified by rather extreme manipulations. Subsequently, listening test results for musical key and scale mode manipulation performed by the MODVOC are presented. For comparison, similar manipulations have been applied to the test material by using a commercial program. Additionally, results of preference tests on specific aspects of subjective audio quality are provided. Moreover, the analysis and synthesis processing chain itself, without any intermediate modulation processing, is assessed for its perceptual reproduction quality.

Finally, in Chapter 7, a comprehensive summary of the work presented in this thesis is given, conclusions are drawn from the results that have been obtained and future prospects of the research field are briefly touched upon.

2 The human auditory system

An introduction to the main properties of the human auditory system is provided in this chapter. The focus is put on the spectral analysis capabilities of the auditory system and the mechanisms that enable and accompany pitch perception. Examining the resolution and the limitations of human auditory perception, masking effects, critical bands, perceptual scales and just-noticeable-differences are addressed. Furthermore, human cognition of pitch and perception of amplitude modulation along with its relation to the sensation of auditory roughness are discussed.

2.1 Auditory biomechanics

2.1.1 The outer and middle ear

The human auditory system, as depicted in Figure 2.1, consists of the *outer ear*, the *middle ear* and the *inner ear* and subsequent neural processing in the brainstem [125][16]. The outer ear collects and filters the airborne sound and directs the sound waves to the *eardrum*. In the middle ear, an impedance transduction is performed, adapting the airborne vibrations inside the outer ear to the fluid vibrations that are excited at the inner ear inside the *cochlea* via the *ossicles* and the *oval window* (*fenestra ovalis*) of the cochlea. A cross section through the coiled cochlea as indicated in Figure 2.1, is shown in Figure 2.2.

2.1.2 The inner ear

The cochlea cross section part indicated by the dashed box in Figure 2.2 is shown in further detail in Figure 2.3. It basically consists of three parts, the *scala vestibuli*, the *scala media* (or: *ductus cochlearis*) and *scala tympani*. The scala vestibuli and scala media are separated by *Reissner's membrane,* and the *basilar membrane* separates the scala media from the scala tympani. The scala vestibuli and the scala tympani are filled with fluid (*perilymph*) and connected at the tip of the cochlea, the *helicotrema.*

In a simplified view, the cochlea can be seen as a passive resonator system with the sound wave entering at the oval window and traveling along the scala vestibuli. Only very low frequencies (a few hundred Hertz) further propagate through the helicotrema at the apex of the cochlea inside the scala tympani towards the round window (*fenestra cochleae*), where they are dampened by the other membrane (*membrana tympani secundaria*). This is to prevent overstimulation of the sensory cells following in the processing chain. In the cochlea, the traveling sound wave excites resonances at distinct locations

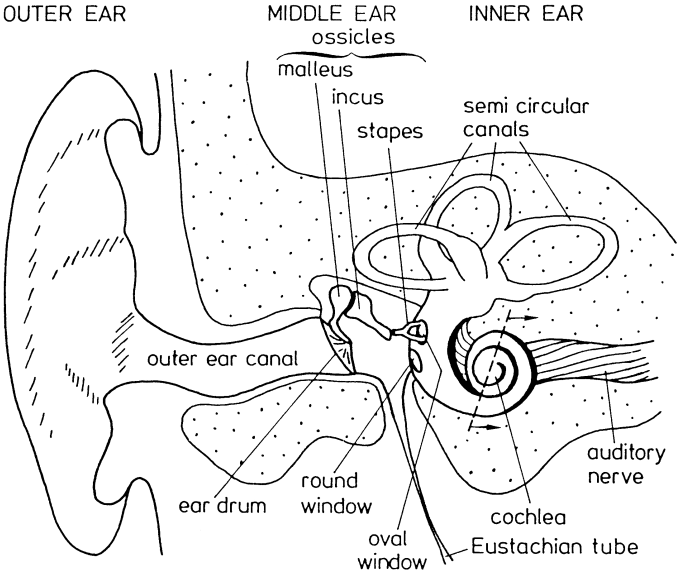


Figure 2.1: Overview of the human auditory system. Reprinted from [125] with kind permission of Springer Science+Business Media.

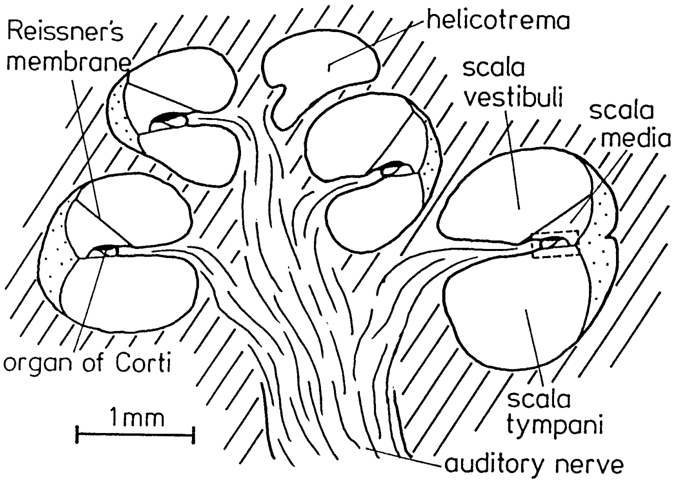


Figure 2.2: Cross section through the cochlea. Reprinted from [125] with kind permission of Springer Science+Business Media.

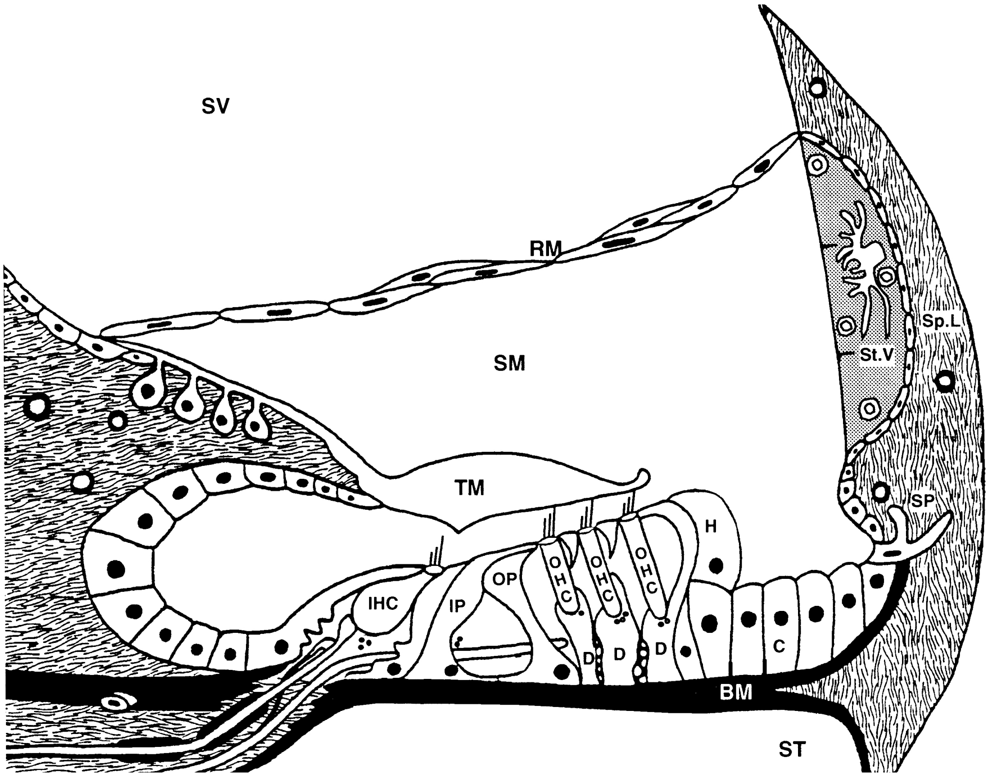


Figure 2.3: The cochlea (section). Pictured is the scala vestibuli (SV), the scala media (SM), the scala tympani (ST), Reissner's membrane (RM), the basilar membrane (BM) and the tectorial membrane (TM). The organ of Corti consists of sensory cells named outer haircells (OHC) and inner hair cells (IHC) and various supporting cells. Reprinted from [16] with kind permission of Springer Science+Business Media.

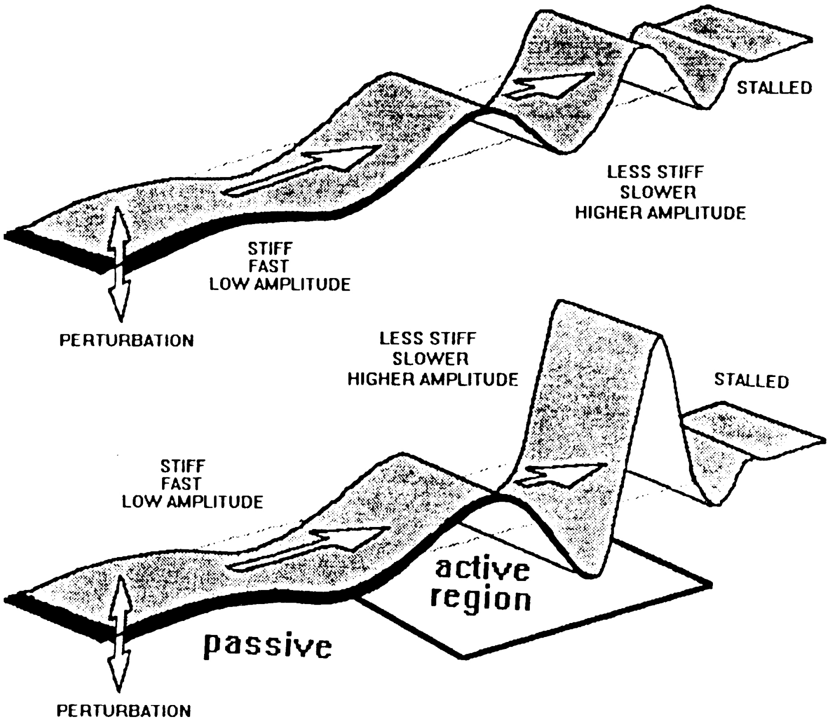


Figure 2.4: Frequency-to-place conversion (tonotopy). Traveling wave on the basilar membrane having graded stiffness in a 'passive' cochlea (upper panel). Traveling wave on the basilar membrane with a combination of graded stiffness and local active amplification producing an 'active' traveling wave, enabling much higher sensitivity and frequency selectivity (lower panel). Reprinted from [16] with kind permission of Springer Science+Business Media.

along the basilar membrane. High frequencies resonate near the oval window where the basilar membrane is rather thick and stiff, and low frequencies near the apex of the cochlea where the basilar membrane is thinner and more elastic. Thus, the cochlea acts as a frequency-to-place converter (*tonotopy*). This is further illustrated in the upper panel of Figure 2.4.

The scala media, which is not directly connected to the other scalae, is filled with a different fluid (*endolymph*). Inside the scala media, the *organ of Corti* is situated on the basilar membrane all along the cochlea. It contains sensory and supporting cells. In the organ of Corti, *outer hair cells* (OHCs) locally amplify the excitation, thereby providing for improved sensitivity and frequency selectivity of the human auditory system. Moreover, the OHCs have a non-linear characteristic in a sense that their positive feedback depends on the absolute energy, in order to prevent self oscillation of the amplifying system [106]. Hence, the auditory system is not a purely passive system, but also contains active elements. The effect is depicted in the lower panel of Figure 2.4.

The transmission of the signal towards the *inner hair cells* (IHCs) is supported by the *tectorial membrane* which covers the apical surface of the sensory cells of the organ of

Corti. Finally, the inner hair cells convert mechanical vibrations to electrical impulses. These are transmitted by the auditory nerve to the brainstem.

2.2 Auditory psychophysics

2.2.1 Bandpass filter model

Auditory biomechanics of the traveling sound wave in the cochlea as sketched in Section 2.1, combined with the concepts of *critical bands* [36], which will be explained in detail in Subsection 2.2.2 and the notion of *excitation patterns* [125] are the foundation of the *place theory* of frequency coding in human auditory perception, characterized by channels tuned to frequency [123]. Figure 2.5 further illustrates this concept. For each location on the unrolled cochlea (right side), the effect of e.g. a sinusoidal burst tone of 1 kHz (upper side) can be modeled by the output of a bandpass filter (left side). Note that different scaling factors have been used to plot the different bandpass filter output signals and the occurrence of increasing filter delays towards the heliocotrema, which are effective for lower frequencies. The parameters of the bandpass filters that are utilized to model the tuning of the auditory channels can be aligned to data that has been obtained from psychoacoustical and physiological experiments and measurements.

2.2.2 Critical bands

Critical bands are closely related to auditory *masking* phenomena. Masking denotes the effect of a decrease in audibility for a *maskee* of a given sound level in the presence of a *masker* of higher sound level that is located in spectral vicinity of the maskee. Moreover, the perceptual phenomena of *beating* tones and *auditory roughness* indicate the inability of the auditory system to resolve inputs which are located within the critical bandwidth of an auditory filter.

In an experiment first published in [36], and subsequently repeated by many researchers [125][123], the threshold for detecting a sinusoidal probe tone (*maskee*) has been measured as a function of the bandwidth of a centered bandpass noise *masker* having constant power density. Consequently, with increasing masker bandwidth the absolute power of the noise increases. For small masker bandwidths, the detection threshold for the maskee also increases proportionally. However, above a certain cutoff bandwidth, a further increase does not lead to a significant rise of the detection threshold. The effect is exemplary illustrated in Figure 2.6.

This can be seen as evidence of the existence of a bandpass filter characteristic in the auditory system that can be envisioned to be dynamically centered around a maskee. Any further increase of the masker bandwidth does not fall in the local spectral scope of such a bandpass filter and hence has no influence on the detection threshold.

This led to the notion of auditory filters that characterize the frequency selectivity of the human auditory system. The bandwidth of these filters is assumed to correspond to the cutoff bandwidth, and is named *critical bandwidth* (CB). The absolute value of a CB

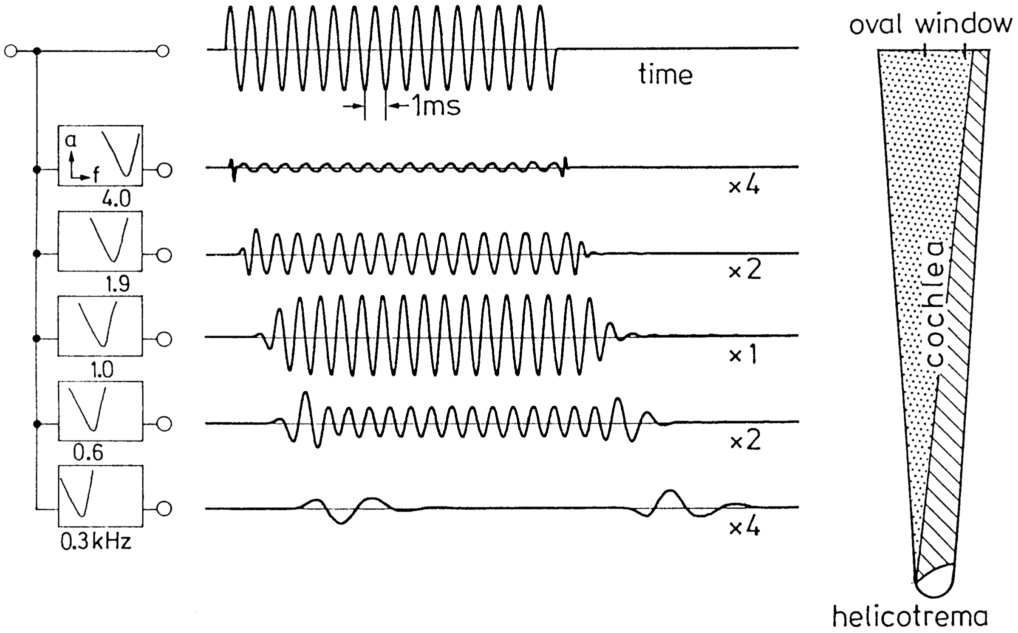


Figure 2.5: Place theory of frequency coding. For each location on the unrolled cochlea (right side), the effect of e.g. a sinusoidal burst tone of 1 kHz (upper side) can be modeled by the output of a bandpass filter (left side). Note that different scaling factors have been used to plot the different bandpass filter output signals and the occurrence of increasing filter delays towards the heliocotrema. Reprinted from [125] with kind permission of Springer Science+Business Media.

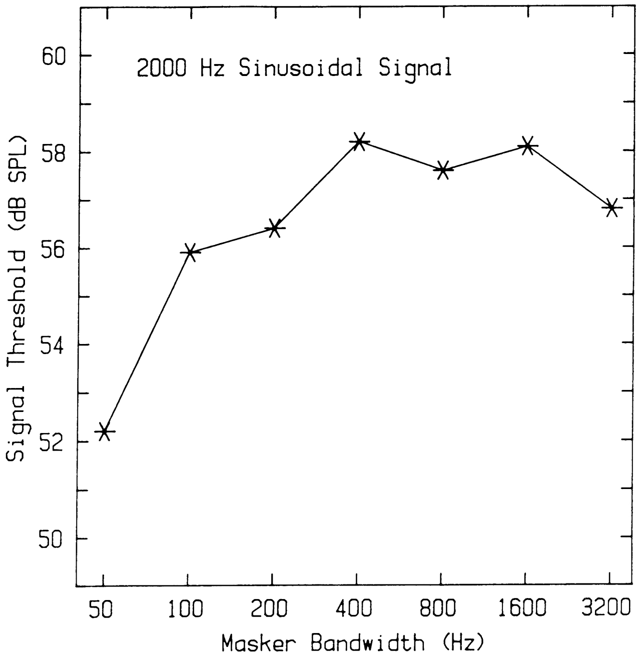


Figure 2.6: Signal detection threshold of a 2 kHz sinusoid masked by centered bandpass noise as a function masker bandwidth. Above approx. 300 Hz the curve flattens off indicating a saturation in masking. Reprinted from [123] with kind permission of Springer Science+Business Media.

on a linear frequency scale depends on the center frequency. Several *perceptual scales* have been proposed, mainly differing due to the design of the underlying experimental setup (see Subsection 2.2.5).

2.2.3 Tuning curves

Tuning curves (TCs), as reprinted in Figure 2.7, are obtained by applying a sinusoidal signal at low level as a maskee and a masker signal being either a sinusoid or preferably a narrow band noise of varying center frequency [123]. For each masker center frequency, the level needed to just mask the maskee signal (indicated by dots) is measured and plotted versus the center frequency (solid curves). The measurement is repeated for several frequencies of the maskee sinusoidal signal. Figure 2.7 additionally includes a graph of the so-called *threshold-in-quiet* as a function of frequency (dashed line). The threshold-in-quiet (or *absolute threshold of hearing*) is defined to be the sound pressure level at a certain frequency that is just detectable in silence.

The TCs indicate the masker level required to produce a given excitation of different auditory filters as a function of frequency. Assuming linearity, the shape of an associated hypothetical auditory filter can be obtained by inverting the TC [123]. Tuning curves can be determined neurophysiologically, by measuring the neural activity of an anesthetized subject, or psychoacoustically, by conducting listening tests [125].

2.2.4 Masking effects

The *simultaneous masking threshold* denotes the sound pressure level of a test sound (maskee) just becoming audible, in spite of the simultaneous presence of an interfering spectrally proximate sound of fixed level (masker). This type of measurements can be conducted with sinusoidal tones and narrowband noise stimuli. Figure 2.8 shows the results of measurements (solid lines) with a probe tone acting as the maskee, and a narrowband noise masker having critical bandwidth. The lower skirt decreases with approximately 100 dB/octave or 27 dB/Bark, the decrease in the upper skirt is less distinct and additionally depends on the level. This is known as the *upward spread of masking*. Figure 2.8 also includes a schematic graph of the threshold-in-quiet (dashed lines).

To some extent, masking threshold curves correspond to *excitation patterns* that are evoked by auditive stimuli in the cochlea. Both have essentially the same shape, but differ in their level. In other words, wherever there is the masking of a maskee, there is also excitation by the masker stimulus [123]. Excitation patterns (and hence also masking curves) are believed to be the combined responses of all auditory filters to a stimulus plotted as a function of their center frequency [123]. Figure 2.9 sketches the (symmetric) responses of five auditory filters to a 1 kHz tone (upper panel). Displayed as a function of their center frequencies, they form the (asymmetric) excitation pattern (lower panel).

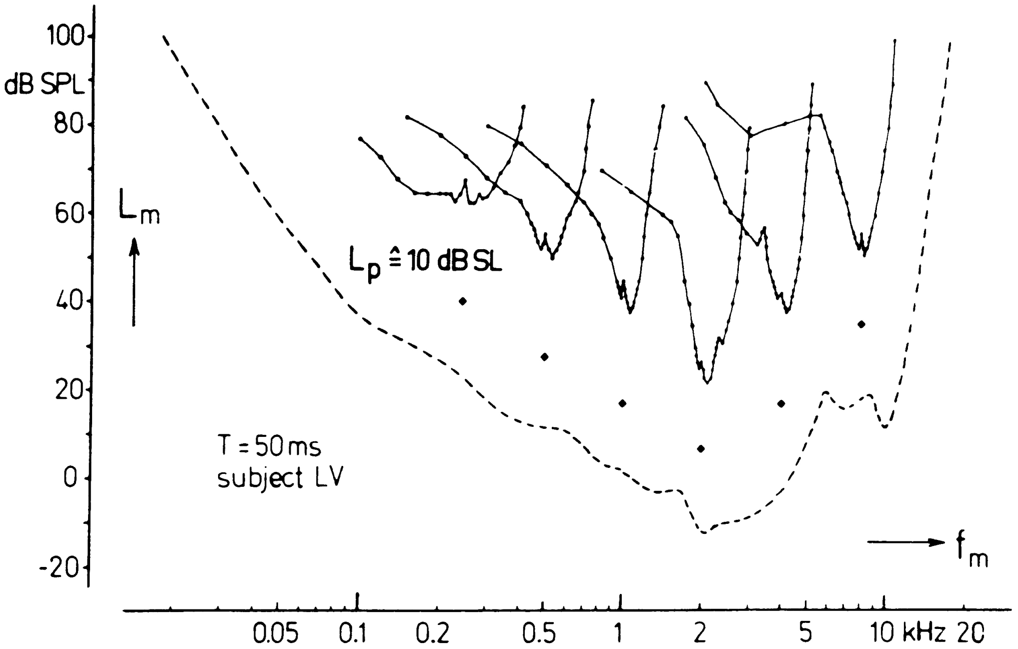


Figure 2.7: Psychophysical tuning curves (TC) indicate the masker sound pressure level L_m required to produce a given excitation of different auditory filters as a function of masking frequency f_m. For each TC, level and frequency of the signal to be masked is plotted (dots below the TC). A measurement of the threshold-in-quiet as a function of frequency (dashed line) is included for reference. Based on data by [116] and reprinted from [123] with kind permission of Springer Science+Business Media.

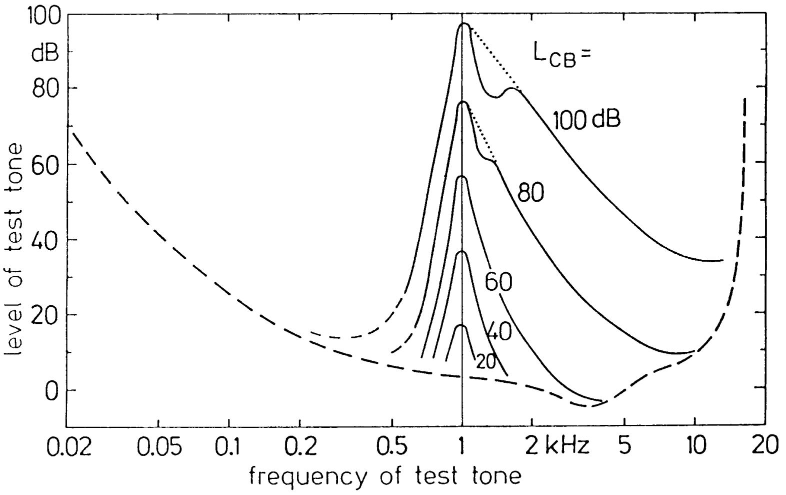


Figure 2.8: Masking threshold (solid lines) of a sinusoidal test tone with varying frequency masked by different noise maskers centered at 1 kHz having sound pressure level L_{CB}(dB). The bandwidth of the noise masker corresponds to the critical bandwidth at the center frequency. Also, the threshold-in-quiet is shown for reference (dashed). Reprinted from [125] with kind permission of Springer Science+Business Media.

2.2.5 Perceptual scales

General

Given the bandwidth measurements $B(f)$ of the auditory filters for all frequencies f, perceptual scales $z(f)$ can be constructed through the seamless stacking of numbered bands [43], starting from frequency of 0 Hz and an associated bandwidth of 0 Hz. The perceptual scale can be obtained via integration, according to Equations 2.1.

$$\begin{aligned} dz &= \frac{\Delta z}{\Delta f} df \\ dz &= (1/B(f))\, df \\ z(f) &= \int_0^f df' \frac{1}{B(f')} \end{aligned} \tag{2.1}$$

Bark scale

Masking experiments using fixed frequency probe tones and centered bandpass noise suggest the existence of critical bands in the human auditory system, as introduced in Subsection 2.2.2. The bandpass noise has a constant power density and a varying bandwidth. Hence, the total noise power varies during the experiment.

Figure 2.10 shows the critical bandwidth as a function of frequency. Below a center frequency of 500 Hz, the critical bandwidth is constant at 100 Hz, followed by a rise of

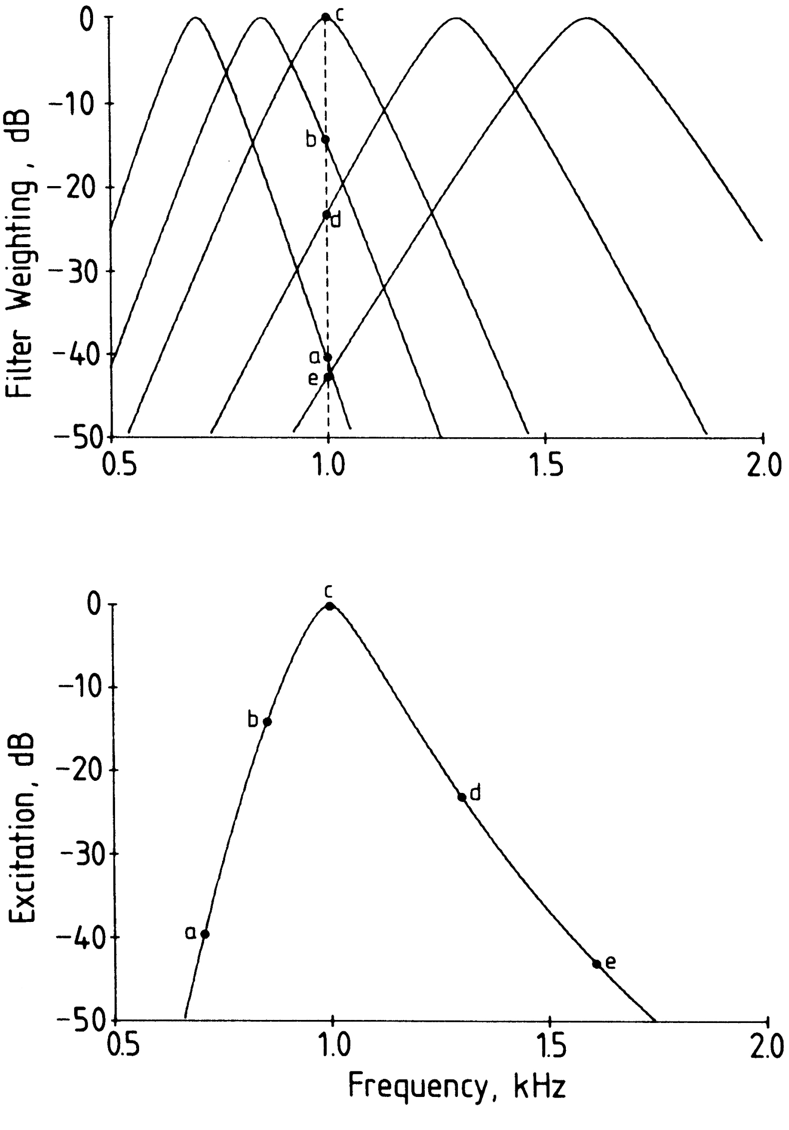


Figure 2.9: Relation of auditory filter shapes and excitation patterns. The responses of five auditory filters to a 1 kHz probe tone are sketched in the upper panel. Displayed as a function of their center frequencies in the lower panel, they form the (asymmetric) excitation pattern. Reprinted from [123] with kind permission of Springer Science+Business Media.

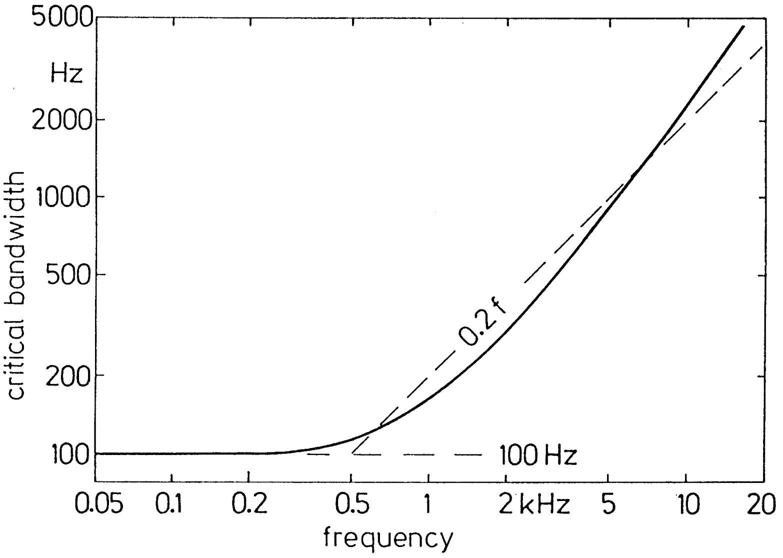


Figure 2.10: Critical bandwidth as a function of frequency of a probe tone. Reprinted from [125] with kind permission of Springer Science+Business Media.

approximately 20 % of the corresponding center frequency. From this, 24 critical bands have originally been defined in a tabulated form that covers the entire human auditory frequency range [125]. Alternatively, Equation 2.2 describes auditory filter bandwidths as a function of frequency.

$$B_G/\mathrm{Hz} = 25 + 75\left(1 + 1.4\ \mathrm{kH}z^{-2} f^2\right)^{0.69} \tag{2.2}$$

Based on the critical bandwidths, the *critical band rate scale* has been derived having the unit *Bark*. The analytical expression in Equations 2.3, including a post-correction step, was given by Traunmüller [111].

$$\begin{aligned} z'/\mathrm{Bark} &= \frac{26.81 f/\mathrm{Hz}}{1960 + f/\mathrm{Hz}} - 0.53 \\ z &= \begin{cases} z' + 0.15\left(2.0 - z'\right) & z' < 2.0\,\mathrm{Bark} \\ z' + 0.22\left(z' - 20.1\right) & z' > 20.1\,\mathrm{Bark} \\ z' & \textit{otherwise} \end{cases} \end{aligned} \tag{2.3}$$

ERB scale

The *equivalent rectangular bandwidth* (ERB) corresponds to the bandwidth of a hypothetical rectangular brick-wall filter that passes the same amount of energy as the corresponding true auditory filter would do. This approach is closely related to the critical bandwidth concept, but, as a result of a different experimental setup, the ERB is widely regarded as being unaffected by certain shortcomings of previous experimental setups such as the parasitic detection of beats or inter-modulation products between the probe signal and masker [72].

In contrast to the critical band measurements that led to the Bark scale, the experimental setup is based on a fixed frequency probe tone located symmetrically at the center of a broadband noise masker, which exhibits a bandstop or notch centered at the signal frequency. While maintaining the total power of the noise masker constant, the signal detection threshold is measured for different widths of the notch. For e.g. a decreasing notch width, increasingly more noise energy leaks through the skirts of the auditory filter and thus rises the detection threshold for the probe tone.

This led to the definition of the ERB provided by Equation 2.4

$$B_{ERB}/\mathrm{Hz} = 24.7\,(1 + 4.37f/\mathrm{kHz}) \tag{2.4}$$

and to the derivation of the *ERB rate scale* given in Equation 2.5 by using Equations 2.1 and 2.4.

$$z/\mathrm{ERB} = 21.3\log_{10}(4.37f/\mathrm{kHz} + 1) \tag{2.5}$$

Comparison of perceptual scales

For higher frequencies, both scales match each other sufficiently well, as the main differences are at low frequencies. The Bark scale suggests a fixed bandwidth of 100 Hz for frequencies below 500 Hz, while the ERB scale has a much finer resolution for low frequencies. This is because the ERB scale not only reflects frequency selectivity due to tonotopy, but also systematically includes the refinement due to correlation based mechanisms. The added sensitivity can be explained by the assumption of a preferred evaluation of temporal cues, performed by a phase locking of the nerve firings to the stimulus waveform [123]. Therefore, the ERB scale better resembles the overall human ability for pitch detection [106]. Figure 2.11 compares the different scales in normalized units plotted versus frequency.

2.2.6 Just noticeable differences for level changes

The *just noticeable difference* (JND) for level variations, often referred to as JNDL, has been measured as plotted in Figure 2.12. The JNDL strongly depends on the absolute sound pressure level and amounts up to 2 dB for levels below 20 dB and slowly descends towards approx. 0.2 dB for high sound pressure levels around 100 dB. If, instead of sound pressure level, the level above threshold-in-quiet is considered, the JNDL is almost independent from frequency [125].

2.2.7 Excitation, masking and JND

Excitation pattern, simultaneous masking and the perception of just noticeable differences can be related to each other. A signal being unmasked is required to alter the

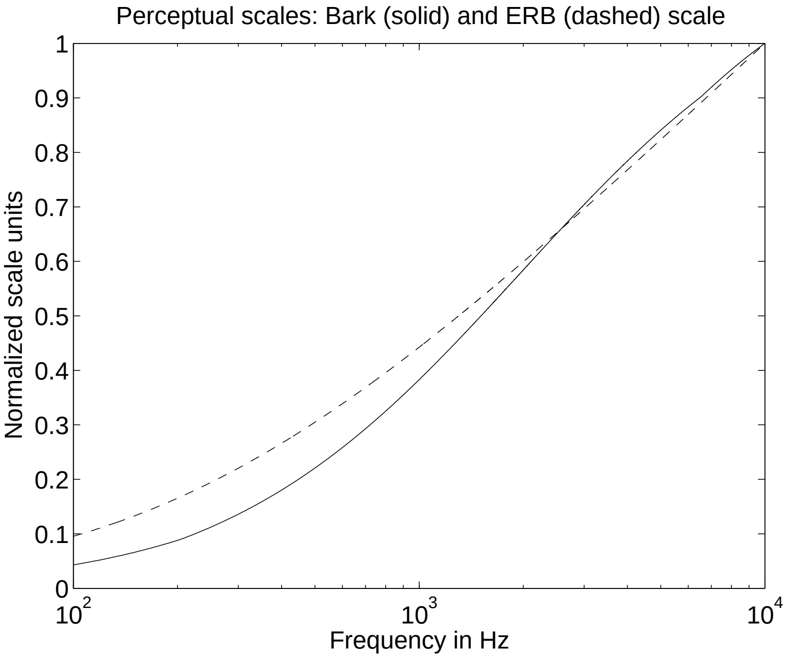


Figure 2.11: Normalized scale units of the Bark scale (solid) and the ERB scale (dashed) as a function of frequency.

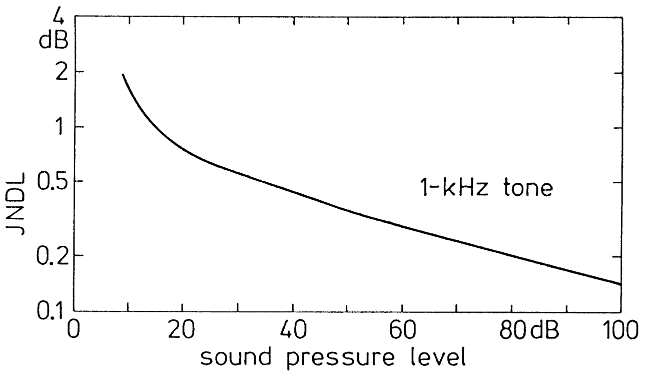


Figure 2.12: Just noticeable differences for level changes (JNDL) as a function of sound pressure level measured with 1 kHz tone. Reprinted from [125] with kind permission of Springer Science+Business Media.

excitation pattern by a minimum of 1 dB at an arbitrary frequency. This frequency is not necessarily centered at the signal frequency, but can be located in its vicinity. For intensity variations, especially at higher sound pressure levels, the upper skirt of the excitation pattern is most likely to be raised by 1 dB, since its steepness varies strongly with level (see Figure 2.8). Thus an increase by a fraction of 1 dB (e.g. 0.25 dB) is already sufficient to raise the excitation level by the necessary 1 dB. This effect is often called *off-center listening.*

Remarkably, the JND measured for frequency variations which will be introduced in Subsection 2.3.1 can also be explained in an analogous way. A frequency shift will become just audible if this shift causes a variation of 1 dB in the excitation pattern. The lower skirt is more sensitive to frequency-to-amplitude conversion due to its greater steepness, as depicted in Figure 2.8. A variation of 1 dB is obtained for a frequency shift of approx. 1/27 Bark.

2.3 Pitch perception

2.3.1 Just noticeable differences for frequency discrimination

Humans can distinguish between roughly 640 different frequencies. The *just noticeable difference for frequency* (JNDF) is mainly dependent on the absolute frequency, but also, to a lesser extent, on the duration and intensity of the stimulus. For stationary pure tones (a duration of more than at least 500 ms), Figure 2.13 displays measurements obtained by different researchers (symbols) and the plot of an analytical expression that has been matched to the experimental data (solid line). Equation 2.6 gives the JNDF for frequency as a function of frequency. Up to 500 Hz the JNDF is constant at approx. 1 Hz. For higher frequencies, the JNDF increases progressively and amounts to approx. 0.2 % of the absolute frequency [106].

$$\Delta f_D\left(f\right) = 1 + \left(\frac{f}{1414Hz}\right)^2 \text{Hz} \tag{2.6}$$

The effect of the stimulus duration on the JNDF is illustrated in Figure 2.14. The JNDF increases considerably with decreasing duration of the stimulus below 500 ms. Hence, the human ability for frequency discrimination ceases.

2.3.2 Spectral pitch

For a sinusoidal stimulus (*pure tone*), the associated human perception is named *spectral pitch.* The sensation of pitch highly correlates with the frequency of a stimulus. Nevertheless, similar to the JNDF, spectral pitch perception also depends on duration, level and global spectral content of the stimulus. For example, Figure 2.15 shows the influence of stimulus intensity on spectral pitch sensation. With increasing level, tones below 2 kHz drop in spectral pitch, while higher tones above 4 kHz rise.

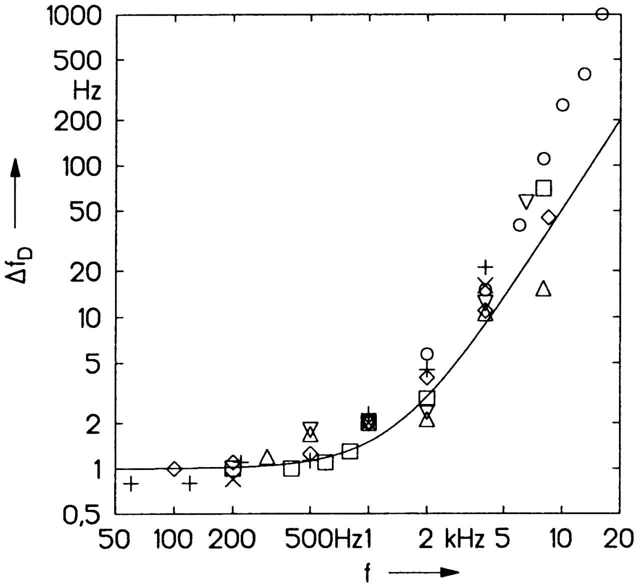


Figure 2.13: Just-noticeable-difference for frequency (JNDF) discrimination as a function of frequency. JNDF measurements Δf_D obtained by various researchers using stationary pure tones are displayed as symbols and a graph of an analytic expression that has been fitted to the data (solid line) is plotted. Reprinted from [106] with kind permission of Springer Science+Business Media.

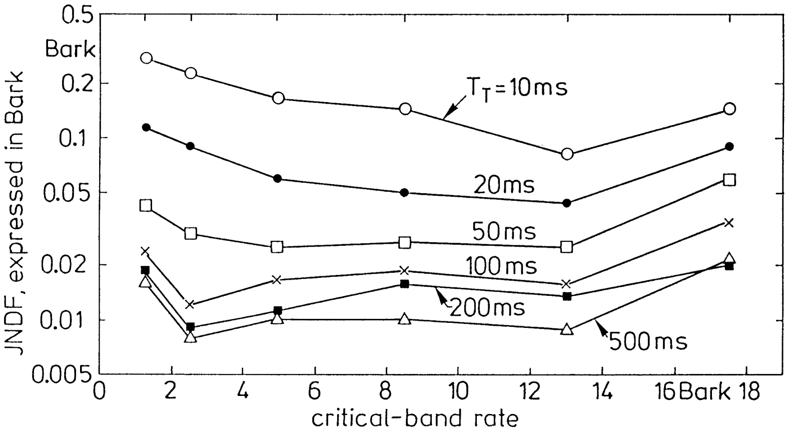


Figure 2.14: JNDF as a function of frequency (both on Bark scale) with parameter denoting the duration of stimulus. Reprinted from [125] with kind permission of Springer Science+Business Media.

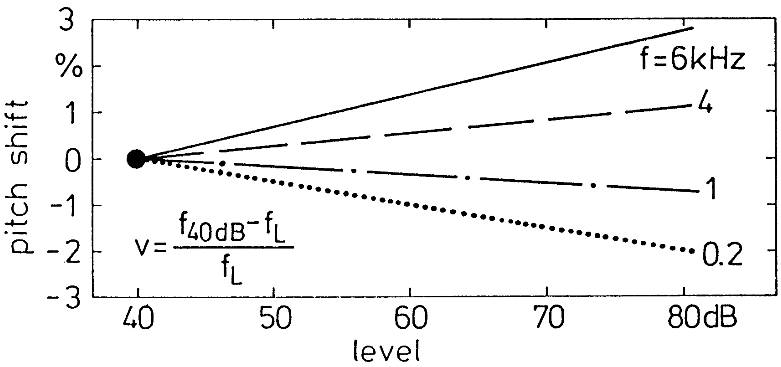


Figure 2.15: Pitch as a function of sound pressure level of stimulus. Reprinted from [125] with kind permission of Springer Science+Business Media.

2.3.3 Virtual pitch

For complex waveforms, the associated human sensation is termed *virtual pitch* [106], since the perceived pitch can be very different from the measured frequency content of the stimulus and can even include components that are not contained physically in the signal (e.g. *missing fundamental*). Figure 2.16 shows different stimuli which all elicit the same pitch sensation albeit with different *pitch strength.* The stimuli are numbered according to their decreasing pitch strength. Panels No. 1 - 4 illustrate examples for stimuli that physically contain the frequency that corresponds to the perceived virtual pitch, panels No. 5 - 6 show the missing fundamental phenomenon and panels 7 - 11 show noise signals that also induce a pitch sensation due to their bandpass center frequency (7, 9), cut-off fringe (8, 11) or amplitude modulation (10).

2.4 Amplitude modulation perception and auditory roughness

2.4.1 Amplitude modulation

An arbitrary audio signal can be expressed in terms of amplitude $E(t)$ and frequency $\omega(\tau)$ in radians (or $f(\tau)$ in Hertz) as denoted in Equations 2.7.

$$\begin{aligned} s(t) &= \Re\left\{E(t)\exp\left(i\left[\int_0^\tau \omega(\tau)\,d\tau + \varphi_0\right]\right)\right\} \\ \omega(\tau) &= 2\pi f(\tau) \end{aligned} \tag{2.7}$$

If defined to be real and positive, $E(t)$ is also termed the *envelope* of the signal $s(t)$. The envelope of a signal is of special interest, since under certain circumstances to be outlined in this section, it directly relates to the auditory perception of relatively slow temporal level variations of a carrier signal. For this reason, it has its own perceptual quality and relevance.

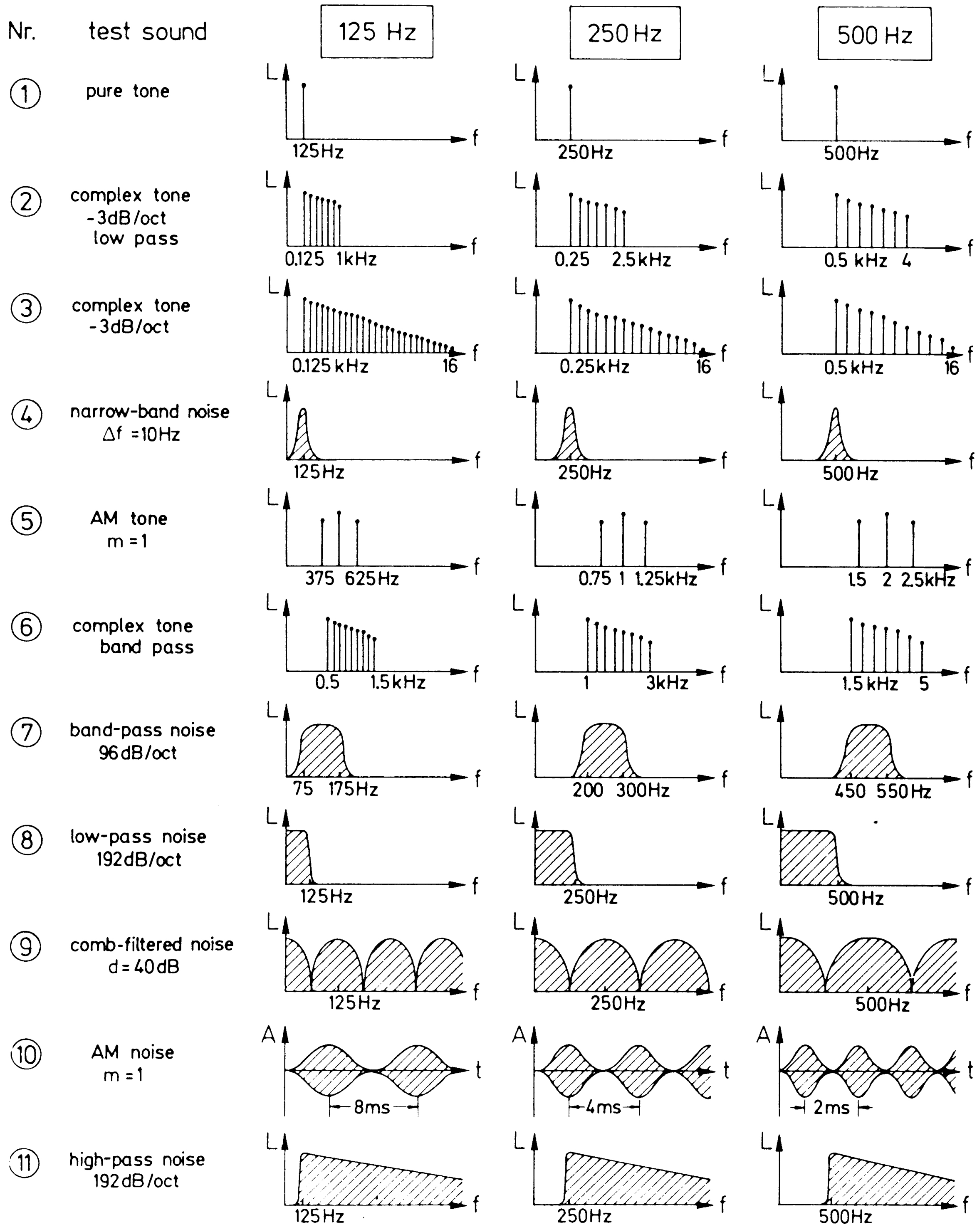


Figure 2.16: Different stimuli that elicit the same pitch, numbered according to decreasing pitch strength. Reprinted from [125] with kind permission of Springer Science+Business Media.

2.4.2 Test stimuli

Suitable stimuli have been defined for measurements of human perception of amplitude modulation. A *sinusoidally amplitude modulated* (SAM) pure tone, which is often referred to in the literature as *SAM stimulus*, can be equivalently described as a *three-tone signal* which, in a spectral view, has a lower and an upper *sideband* located symmetrically around a *carrier* (see Equation 2.8). If ω_m denotes the modulation frequency and ω_c the carrier frequency, the lower and the upper sideband are pure tones of frequency $\omega_c - \omega_m$ and $\omega_c + \omega_m$, respectively.

$$\begin{aligned} x(t) &= (A + a \cdot \cos(\omega_m t + \phi)) \sin(\omega_c t) \\ &= A \sin(\omega_c t) + \frac{a}{2} \sin((\omega_c + \omega_m) t + \phi) + \frac{a}{2} \sin((\omega_c - \omega_m) t - \phi) \end{aligned} \tag{2.8}$$

If expressed in the form of Equation 2.7, the three-tone signal exhibits a pure sinusoidal amplitude modulation $E(t)$ with frequency ω_m and *modulation depth* of $m \in [0....1]$ as denoted in Equations 2.9.

$$\begin{aligned} E(t) &= (A + a \cdot \cos(\omega_m t)) \\ m &= \frac{a}{A} \end{aligned} \tag{2.9}$$

Another important test signal is the *two-tone signal*, which consists of two pure tones with a spectral distance $\Delta\omega$, according to Equation 2.10.

$$\begin{aligned} x(t) &= \cos(\omega_m t) \sin(\omega_c t + \phi) \\ &= \frac{1}{2} \sin((\omega_c + \omega_m) t + \phi) + \frac{1}{2} \sin((\omega_c - \omega_m) t + \phi) \\ &:= \frac{1}{2} \sin((\omega + \Delta\omega) t + \phi) + \frac{1}{2} \sin(\omega t + \phi) \end{aligned} \tag{2.10}$$

Expressing the two-tone signal in the form of Equation 2.7, its amplitude modulation $E(t)$ equals a full-wave rectified cosine shape envelope, as expressed in Equation 2.11.

$$E(t) = 2 \left| \cos\left(\frac{\Delta\omega}{2} t + \frac{\phi}{2}\right) \right| \tag{2.11}$$

Equation 2.12 describes the Fourier series expansion of a full-wave rectified cosine of frequency $\frac{\Delta\omega}{2}$.

$$\left| \cos\left(\frac{\omega}{2} t\right) \right| = \frac{2}{\pi} - \frac{4}{\pi} \sum_{k=1}^{\infty} (-1)^k \frac{\cos(k\Delta\omega t)}{(2k)^2 - 1} \tag{2.12}$$

The fundamental Fourier component $k = 1$ of the envelope has the frequency $\Delta\omega$. This is illustrated in Figure 2.17, where one cycle of the fundamental Fourier component equals the duration of one half-wave of the rectified signal. These amplitude modulations are often also referred to as *beatings*.

A comparison of Equations 2.10 and 2.8 suggests that a two-tone signal can also be seen as a SAM signal, albeit with a suppressed carrier. A further comparison of

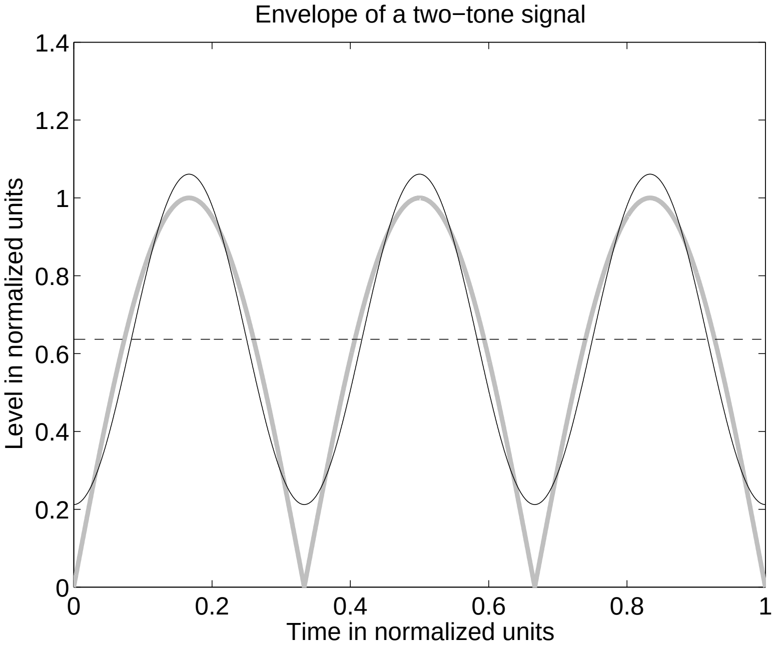


Figure 2.17: Envelope of a two-tone signal (solid gray) and its first two Fourier series terms, the constant (dashed black) and the fundamental component (solid black).

Equations 2.11 and 2.9 reveals that for a given spectral bandwidth, the fundamental of the amplitude modulation associated with the two-tone signal is twice the modulation frequency of a three-tone complex. Therefore, the two-tone signal can be seen as the upper limit of the fundamental modulation frequency present in a bandlimited signal of predefined bandwidth.

2.4.3 Modulation perception

Figure 2.18 illustrates the perceptual effect of a two-tone stimulus according to Equation 2.10, consisting of a pure tone with frequency f_1 and a second pure tone having equal level, albeit different frequency f_2, resulting in a frequency difference of Δf. Three types of auditory sensations can be distinguished. If the absolute frequency difference of the two tones exceeds a threshold Δf_D, both tones are perceived as separate tones. If the absolute frequency difference of the two tones is smaller than the threshold Δf_D, both components are perceptually fused into one sound impression of spectral pitch $(f_1 + f_2)/2$. Depending on the frequency offset Δf, the fused tone is characterized by either audible beatings ($|\Delta f| < 20$ Hz) or auditory roughness (20 Hz $< |\Delta f| <$ 300 Hz). Auditory roughness gradually decreases if Δf approaches the critical band border Δf_{CB}.

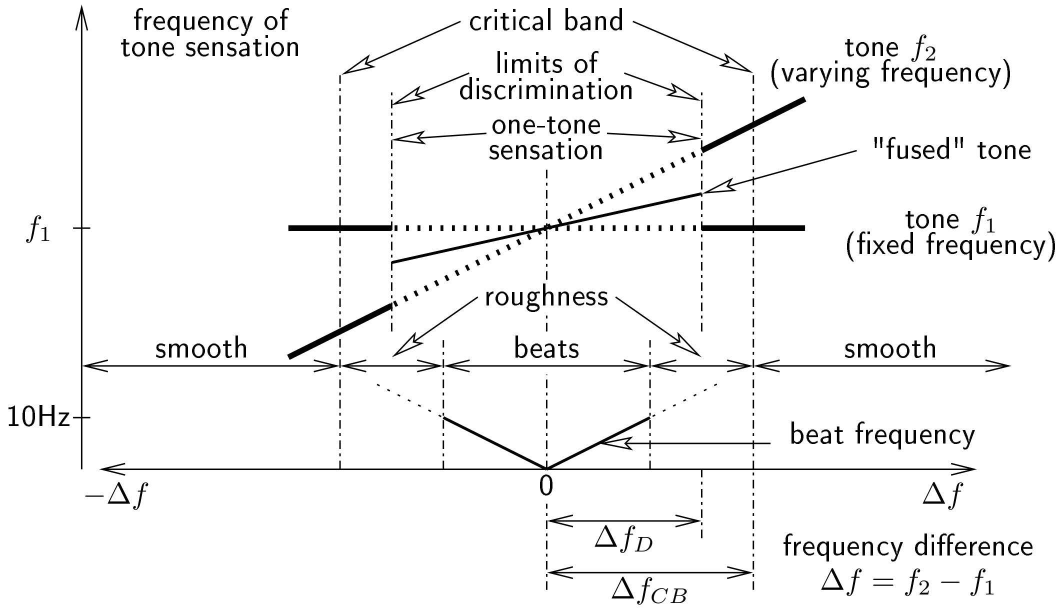


Figure 2.18: Perceptual effect of a pure tone with frequency f_1 and a second pure tone with frequency f_2 having a variable frequency offset Δf from the first tone. Three types of auditory sensations can be distinguished. If the absolute frequency difference of the two tones exceeds a threshold Δf_D both tones are perceived as separate tones. If the absolute frequency difference of the two tones is smaller than the threshold Δf_D, both components are perceptually fused into one fused tone. Depending on the frequency offset Δf, the fused tone is characterized by either audible beatings ($|\Delta f| < 20$ Hz) or auditory roughness ($|\Delta f| > 20$ Hz). Auditory roughness gradually decreases if Δf approaches the critical band border Δf_{CB}. Redrawn after [92].

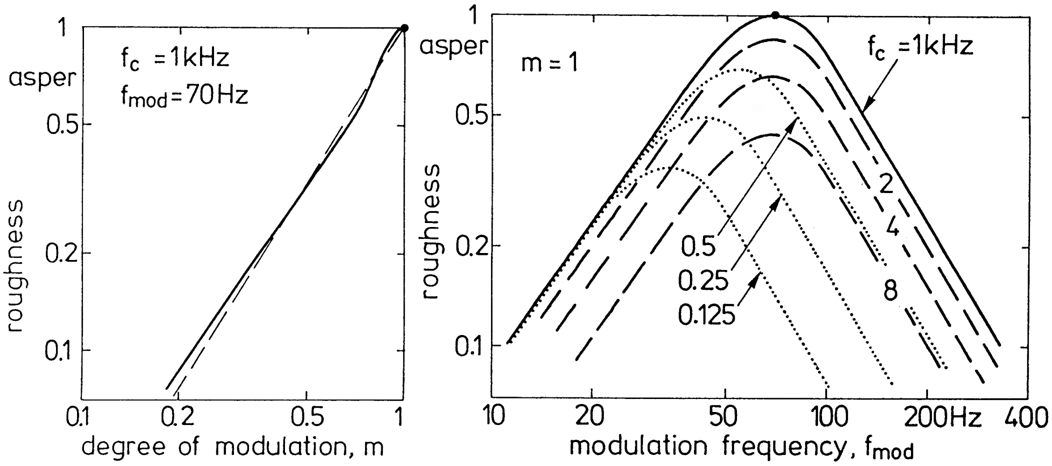


Figure 2.19: Auditory roughness as a function of modulation depth for a 1 kHz tone modulated by 70 Hz (left panel). Auditory roughness as a function of modulation frequency (right panel). Reprinted from [125] with kind permission of Springer Science+Business Media.

2.4.4 Auditory roughness

While the amplitude modulation of a carrier signal by a much lower modulation frequency is perceived as intensity fluctuations, the modulation of the same carrier by higher frequencies ($>$ 20 Hz) is experienced as auditory roughness of the carrier signal [125]. The degree of auditory roughness depends on the carrier frequency ω_c , the modulation frequency ω_m and the modulation depth m of the signal.

In order to quantify the sensation of roughness, a subjective listening test presented sinusoidally modulated pure tones, according to Equation 2.8, to human subjects. The dependency on modulation depth was found to obey the proportionality relation $R \sim m^{1.6}$. Apart from a weak dependency on presentation level, auditory roughness strongly depends on modulation frequency. All carrier frequencies $>$ 1 kHz have a maximum at 70 Hz with modulation depth $m = 1$, albeit with decreasing absolute value towards higher carrier frequencies. These findings might be explained by the limited temporal resolution of the human neural system, which can not follow fluctuations above 300 Hz [106]. For decreasing carrier frequencies below 1 kHz, the maximum is shifted to lower modulation frequencies and the absolute value decreases. The explanation for this is the higher frequency selectivity at frequencies below 1 kHz, since only spectral components that fall into a critical band are perceptually fused, and thus contribute to the sensation of roughness [106].

The unit measuring auditory roughness is named *asper*. One asper was defined to denote the perceived roughness corresponding to a 1 kHz carrier, modulated by 70 Hz with a modulation depth of 1. Zwickers measurements are illustrated in Figure 2.19, which shows auditory roughness dependent on modulation depth (left panel) and on modulation frequency for different carrier frequencies (right panel).

More detailed models for auditory roughness estimation can be found in [6] and, later, an improved model in [17]. The model according to [17] is based on an spectral analysis by Discrete Fourier Transform (DFT) of time segments with a duration of 200 ms. Subsequently, the spectrum is mapped to an excitation pattern of the basilar membrane. The excitation pattern is further decomposed into 47 fixed spectrally overlapped bandpass signals each having a width of 1 Bark, and transformed back into time domain using the original phases. From the bandpass signals, the quadratic mean is calculated and related to the constant component, yielding a *generalized modulation depth* of each bandpass signal. These are converted into bandpass roughness by incorporating a correlation measure between adjacent bands, a carrier frequency dependent factor and a proportionality relation $R \sim m^2$. Finally, the subband roughness measures are added to obtain the global roughness estimation.

The model was successfully evaluated through the comparison with subjective test results for stimuli like two-tone signals, amplitude modulated pure tones (SAM signals), amplitude modulated bandpass noise and broadband noise. The roughness evoked by a two-tone signal is very similar to that of a three-tone signal, if both signals have the same amplitude modulation fundamental frequency [6].

2.5 Summary

In the human auditory system, the inner ear performs a spectral decomposition of the audio signal. This is accomplished by a frequency-to-place transformation in the cochlea (*tonotopy*) and, for the lower frequencies, the additional evaluation of correlation cues.

The frequency resolution of the spectral decomposition is reflected by so-called *critical bands* and by the introduction of auditory scales, like the *Bark* scale or the *equivalent rectangular bandwidth* (ERB) scale. The ERB scale is believed to model both effects, tonotopy and correlation cues, more accurately. The effective spectral decomposition filters in the human ear are aligned with the actual spectral components of the signal and are therefore most appropriately modeled by time-variant filterbanks.

Human pitch perception is a highly abstract process. Therefore, pitch perception is beneficially described by two different terms: *spectral pitch* and *virtual pitch*. Spectral pitch denotes the ability of the hearing system to perceive the frequency of pure sinusoidal tones. To quantify this ability, the *just noticeable difference for frequency* discrimination (JNDF) is an important measure. Moreover, the absolute spectral pitch depends weakly on the duration and level of the stimulus. Virtual pitch relates to the pitch sensation for more complex signals. The virtual pitch does not necessarily relate to components that are contained physically in the audio signal, but can also correspond to the estimated fundamental of a harmonic signal or the frequency of a bandpass center, a cut-off fringe or an amplitude modulation.

The perception of a multi-tone signal also depends on its *amplitude modulation* (AM) properties. If the fundamental frequency of the amplitude modulation is below approx. 20 Hz, the signal is experienced as a perceptually fused signal having a single frequency and, additionally, intensity variations. For the interval of 20 Hz up to 300 Hz, the signal elicits the notion of a fused signal exhibiting *auditory roughness*. Above the upper frequency of roughness perception, both components are perceived as separate entities. In this case, their spectral distance exceeds the width of a critical band.

3 Modulation analysis and synthesis

This chapter presents different approaches to modulation analysis known from literature and the most salient aspects of the distinct methods are discussed. Additionally, the feasibility and properties of associated synthesis techniques are addressed.

3.1 Overview

The decomposition of an audio signal into amplitude modulation (AM), frequency modulation (FM) and, where applicable, carrier components is the subject of investigation for a fairly long time. Due to the ill-posed nature of the problem, there is an infinite number of possible decompositions, such that most publications are concerned with proposing a decomposition that on one hand is unique with respect to certain criteria, and on the other hand yields results that can be physically interpreted.

To provide a systematical review of the literature in this thesis, it is proposed to describe any system for modulation analysis and synthesis of audio signals by the *modulation analysis and synthesis* (MAS) block diagram depicted in Figure 3.1.

For analysis, first the audio input signal is optionally divided by a filterbank or by a transform into subbands. Subsequent to any optional preprocessing, the amplitude modulation and frequency modulation is estimated. Both estimation methods may be mutually dependent. Additionally, one or more carrier signals may either be determined or are already implicitly constituted by the design of the preceding filterbank or transform. For synthesis, there exists a composition stage which is fed by the carriers and the modulation information. The various publications on modulation decompositions differ in various detail aspects of this block diagram. For instance, the initial filterbank or transform stage may be constructed employing

- different filter designs
- uniform or non-uniform band structure
- static or signal adaptive filters

The preprocessing may consist, for example, of a

- Hilbert transform
- minimum phase and maximum phase or all-phase decomposition
- dynamic compression

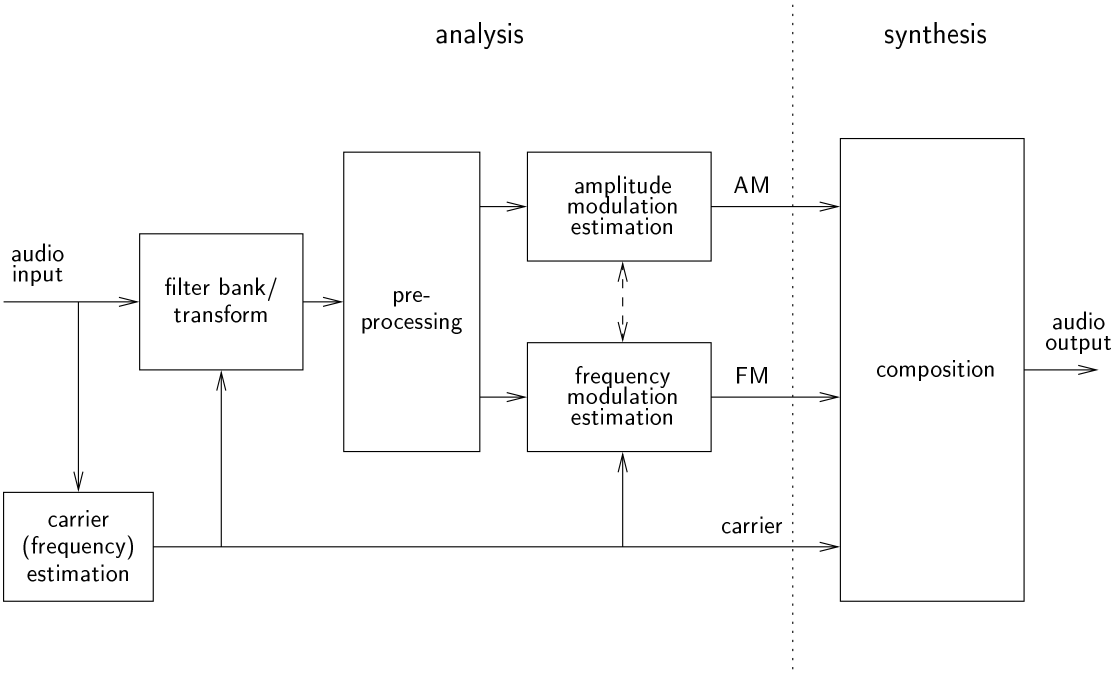


Figure 3.1: Block diagram of modulation analysis and synthesis (MAS).

- auditory model related processing

The parameter extraction stage may be realized by various means. The estimation of AM may be performed by computation of

- Hilbert envelope of the analytic signal (incoherent demodulation)
- synchronous detection (coherent demodulation)
- full or half-wave rectification, lowpass filtering
- energy operators

The estimation of the FM may depend on

- the phase of the analytic signal
- the phase derivative or *instantaneous frequency* (IF)
- energy operators

The proposal of an appropriate synthesis scheme is not within the scope of most publications. If a synthesis method is given, the reconstruction may be dependent on the invertibility of each of the analysis processing steps

- exact

- approximate (pseudo invertible)

In the following, different methods that have been published so far are reviewed with respect to the MAS block diagram.

3.2 Vocoder

3.2.1 Description

A groundbreaking publication was the work of Dudley [25][26], who, for the first time, described a so-called VOCODER (a made-up word from «voice encoder»). The publications proposed the analog synthesis of speech from carriers and modulators. For the time being, the vocoder was intended to solely mimic the human voice and attracted - renamed as VODER (again a made-up word from «voice operation demonstrator») - worldwide attention as a demonstration object on the New York's World Fair 1939. The underlying signal model is illustrated in Figure 3.2. Based on the excitation by pulse-trains or noise as carrier signals, the time variant AM and FM was applied to the carriers in the subsequent processing. The vocoder was controlled manually by intensively trained operators, as depicted in Figure 3.3.

Much later, the vocoder was rediscovered for the purpose of parametric coding of speech by Flanagan [35], hereby extending the basic idea of a voice synthesizer toward a completely automatic digital analysis and synthesis system called the *phase vocoder*. Again, a considerable time later, in the course of the emergence of digital signal processing, the phase vocoder was generalized for subband coding [34], which can be seen as the first disengagement from the limiting source model of solely human speech: in [34], the relations between amplitude envelope (characterizing AM), phase progression (specifying FM) or its derivative, the *instantaneous frequency* (IF), and the *short time fourier transform* (STFT) spectra of an audio signal are discussed. It is noted that the AM and FM of band-limited signals are not limited in their bandwidth, which is undesirable. Moreover, it is already speculated that a perceptually adapted multiband processing would be advantageous in terms of reproduction quality, a proposition which this thesis is also based on. Lastly, some related work by Malah [64] is critically cited: Malah proposes an enhanced vocoder utilizing pitch tracking, which Flanagan legitimately considered to rely on «fragile operations», like voiced-unvoiced decisions and f0 pitch estimation.

In later publications, the phase vocoder [35] was further improved in terms of perceptual quality e.g. by *regions of influence* based phase locking in order to reduce so-called *phasiness* artifacts and poor transient reproduction [57][27][90][91]. Since the modulation components obtained by the phase vocoder are, in general, not band-limited and, due to the static nature of the filterbank, their physical interpretation is not straight forward, successive publications concentrate on the requirements of a physically meaningful decomposition.

Loughlin and Tacer [62] derived desireable properties of any meaningful decomposition, based on an exemplary two-tone test signal, into AM and FM components: for

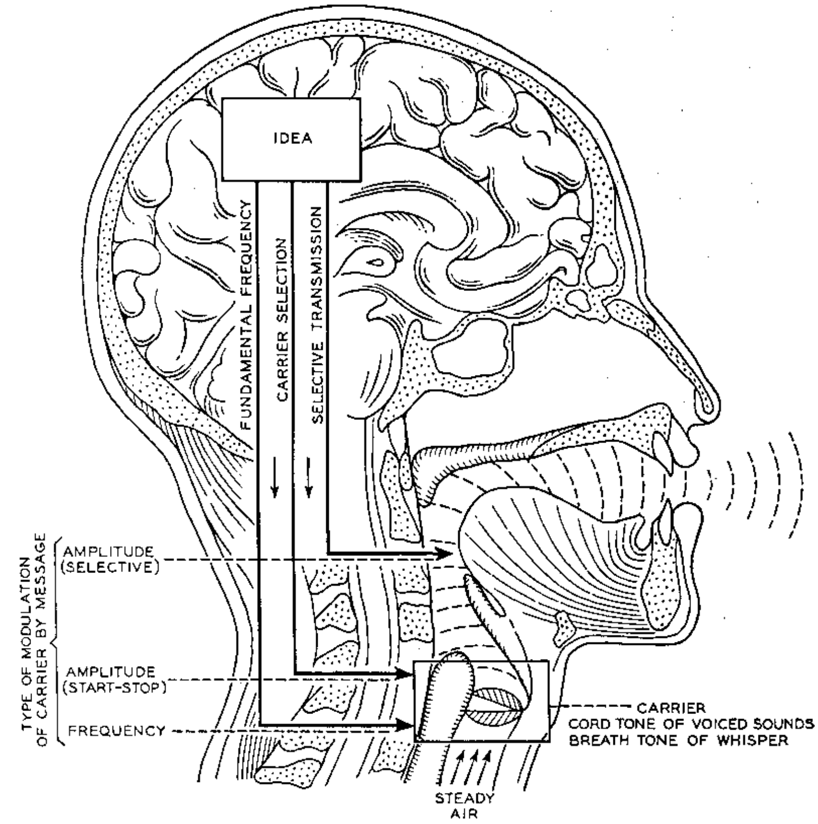


Figure 3.2: The vocoder signal model. Reprinted from [26].

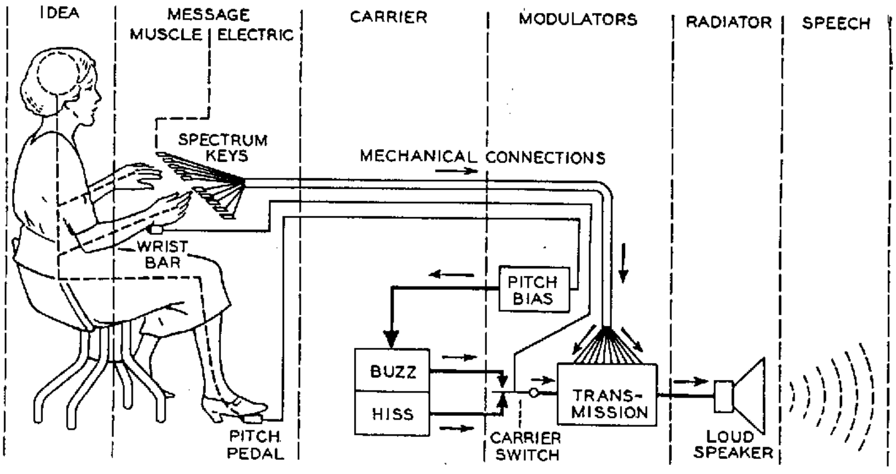


Fig. 8—Schematic circuit of the voder.

Figure 3.3: The vocoder. Reprinted from [26].

AM and FM, they demanded bandwidth limitation and mutual independence. Moreover, they claimed that for a stationary signal, the components are also desired to be stationary. One year later, the authors point out again the inherent contradictions of the established notion of instantaneous frequency being the «the average frequency at each time» [63]. Wei and Bovik [119] extended these considerations being made with respect to two-tone signals to multi-tone signals and concluded the infeasibility of IF for analysis of multi-tone signals due to the lack of interpretability. As a solution, they again propose a suitable spectral decomposition prior to AM/FM estimation, thereby further strengthening the concept of signal adaptive subband filters.

3.2.2 Relation to MAS diagram

The vocoder [25] broadly initiated the idea of distinguishing between carrier and modulator signals. Sinusoidal tone clusters («buzz») and noise («hiss») were used as carriers, which were subsequently amplitude and frequency modulated. So, in a sense, an analysis was performed by humans beforehand, determining suitable carriers and modulation parameters. Synthesis was accomplished through the interaction of the human operator controlling these parameters, and the machine synthesizing the output signal.

The phase vocoder [34] automated the process of parameter analysis and synthesis at the price of a considerable abstraction from the original source model, which had been designed to be physically meaningful, but limited to human speech reproduction. Related to the MAS diagram in Figure 3.1, the phase vocoder utilized a *discrete Fourier transform* (DFT) as an initial static transform; AM was represented by the magnitude of the complex DFT coefficients, FM by the time derivative of the phase for each transform bin, and carrier frequencies implicitly by the spectral locations of the DFT bin centers. The synthesis was computed by integrating the phase derivative, the re-combination of the same with the amplitudes and application of the *inverse discrete Fourier transform* (IDFT). Since the absolute phase is lost due to the phase derivation and subsequent integration, the inversion is only approximate.

3.3 Energy operators

3.3.1 Description

A pioneering publication by Kaiser [53] in 1990 first derived an energy operator, the so-called *Teager-Kaiser-operator*, from differential equations describing a spring-mass system. The intended purpose was to provide a measure for the energy that is needed to excite a certain oscillation. The application of this operator to the time signal indeed yielded the desired measure, yet only for mono-component signals. Given a superposition of two oscillations, the operator is observed to describe the envelope of the sum signal [53]. Based on this, Maragos et al. proposed the use of energy operators for performing an AM/FM decomposition on speech signals [66][65]. In these publications, the authors also showed that a subband decomposition prior to AM/FM analysis aids to obtain meaningful results for multi-component signals. In [42], the rough idea of a speech formant vocoder based on energy operators was initially published. Later, in [83], an approach for speech formant tracking was outlined and subsequently integrated into a complete speech formant vocoder application [84]. For music signals, Sussman and Kahrs demonstrated how the sound of a guitar string can be analyzed, modelled and synthesized through the use of energy operators [104].

The AM/FM separation method that has been applied in the above cited publications is widely known as *energy separation algorithm* (ESA), or its discrete counterpart, the *discrete-time energy separation algorithm* (DESA) [65], and is based on energy operators like, for example, the Teager-Kaiser-operator. In this publication and, earlier, in [82] it is further pointed out, that the calculation of instantaneous frequency and amplitude envelope can be achieved alternatively by *Hilbert transform demodulation* (HTD). The authors conclude that HTD is more robust to noise, but much more demanding in terms of computational complexity. Quadrature operators [4] are a generalization of the Teager-Kaiser-operator. In [5], for example, an enhanced quadrature operator is proposed for the estimation of FM.

3.3.2 Relation to MAS diagram

In the MAS diagram in Figure 3.1, energy operators constitute an alternative way of AM/FM estimation. Energy operators can be combined with any variant of a preceding filterbank or preprocessing. In [5], for example, a static filterbank was used to obtain subband signals, the FM of which was estimated applying a quadrature operator; the FM is utilized for the demodulation of the subband signal to further obtain the AM signal. In [84] energy operators are applied to the output of a signal adaptive filterbank which is steered by speech formant locations. Compared to Hilbert transform based estimation methods, energy operators have much lower computational complexity, but, according to [65][82], can be expected to be more sensitive to noise. Since the presence of noise must be assumed in generic music recordings, the work presented in this thesis relies on Hilbert transform based methods.

3.4 Minimum phase and all-phase decomposition

3.4.1 Description

In 1995, the authors Kumaresan and Rao [55] suggested an alternative method of AM/FM decomposition, building upon the much earlier findings of Voelcker [114][115]. They proposed the decomposition of an analytic signal into a minimum phase and a maximum phase component prior to AM and FM estimation. Later, in 1998, the proposal was modified to decompose the analytic signal into a minimum phase and an all-phase component [56]. Subsequently, AM was estimated by the amplitude envelope of the minimum phase component and FM was shown to be represented by the so-called *positive instantaneous frequency* (PIF), which was obtained via the time derivative of the phase of the all-phase component. The decomposition into minimum phase and all-phase component was accomplished through the application of *linear prediction in the spectral domain* (LPSD). Following up in 2000, the method was extended by the same authors to a complete analysis system for human speech [89]. The system additionally utilized time variant, signal adaptive bandpass filters, the center frequencies of which closely followed the speech formants or other prominent local spectral concentrations of energy. In [109], such a decomposition was applied to a violin tone in order manipulate its vibrato in depth or in time scale. Thereby, the source model was extended from solely human speech to musical tones. Recent publications suggest some additional detail improvements, like the application of a generalized phase locked loop, reassembling the effect of the outer hair cells of the human ear, in order to estimate the PIF [118].

3.4.2 Relation to MAS diagram

With respect to the MAS diagram in Figure 3.1, minimum phase and all-phase decomposition is a preprocessing, that can be utilized in order to obtain better suited modulation parameters. In contrast to IF, the PIF is strictly positive, and therefore is claimed to have a better physical interpretability [56]. It can be combined with any variant of a preceding filterbank. For instance in [89], it was combined with signal adaptive *dynamic tracking filters* (DTF). In particular, the idea of advantageously employing signal adaptive filters centered on local energy concentrations prior to AM/FM decomposition can be seen as the major influential contribution of Rao's publication. Sharing the same view, this thesis is also based on that conviction.

3.5 Subband amplitude modulation spectra

3.5.1 Description

In 2001, Vinton and Atlas [113] proposed a two-stage transform as a key component of a fine grain scalable audio coder that converted audio signals into a spectral decomposition of its subband envelopes. The basic scheme of this transform is depicted in Figure 3.4. For consecutive time blocks of an input signal, an initial time-frequency transform provides sets of complex spectral subband coefficients (rows). A group of successive spectral coefficient sets is stored in a two-dimensional buffer. For each spectral subband, a frame of temporally successive coefficients is assembled (column). This frame is incoherently demodulated, taking the magnitude of each coefficient, and is further analyzed in its spectral content by a second transform. The procedure is repeated for all spectral subbands of the set, finally yielding the *subband amplitude modulation spectrum* (SB-AMS). The phase of the spectral subband coefficients, stored in a second two-dimensional buffer (not shown in Figure 3.4), is retained without any further processing. For the synthesis operation, the processing is reverted. In addition to the SB-AMS data, the phase data is also needed for exact invertability.

The time-frequency transforms, that were originally proposed for the processing, were *time domain aliasing cancellation* (TDAC) transforms due to their *critical sampling* property. This was influenced by the context of bitrate efficient audio coding, in order to keep the quantity of information to be coded already as low as possible in the preprocessing stages. In 2003, Thompson and Atlas published a modified variant of the transform, applying a non-uniform decomposition in the second stage and it was claimed to achieve a better reproduction quality of transients at limited bitrates [108].

From the viewpoint of perceptual and physical meaningfulness, this approach is questionable, since

1. already the first stage does not deliver physical interpretable subband amplitude envelopes due to its critical sampling property

2. the static spectral location of the subbands of the first stage promote mutual AM and FM conversion

3. the content of the phase matrix is barely related to the physical properties of the signal, and thus hard to interpret and modify

4. high quality synthesis of modulation filtered signals is not possible due to the utilization of original, unmodified phase data

The application of oversampled transforms, like overlapping DFT, might be a straight forward measure to successfully address issue number 1. Hence, in the following, problem number 2, AM/FM conversion, will be explained in more detail, since this effect is considered to be an important issue with respect to the physical interpretability of a modulation decomposition.

An illustrative example of such AM/FM conversion is depicted in Figure 3.5. A sinusoidal carrier modulated in its amplitude by a low frequency sinusoid results in a three tone signal, which is symmetric in its magnitude, with respect to the carrier, as sketched in the left top panel. If the carrier frequency coincides with the center frequency of a bin of the initial transform, the subband signal in this bin remains purely amplitude modulated. This is confirmed by the graphical addition of the three phasors in the upper middle panel. The result is a phasor of varying length (pure AM) depicted in upper right panel. If there is an offset frequency between the carrier and center frequency, as outlined in the left bottom panel, the subband signal is asymmetrically damped by the subband filter transfer function and thus appears to also have FM, as sketched in the lower middle and lower right panels. Here, the resulting phasor draws an ellipse, indicating the presence of AM and FM. Similar effects exist for purely frequency modulated signals which, due to the static nature of the filterbank, seemingly exhibit an AM part also.

To remedy issues number 3 and 4, one approach is to restrict the SB-AMS synthesis to AM only and estimate matching phase values using a magnitude-only synthesis approach [88][78][41] based on the principle of *projection on convex sets* (POCS). An application based on this principle is, for example, described in [107].

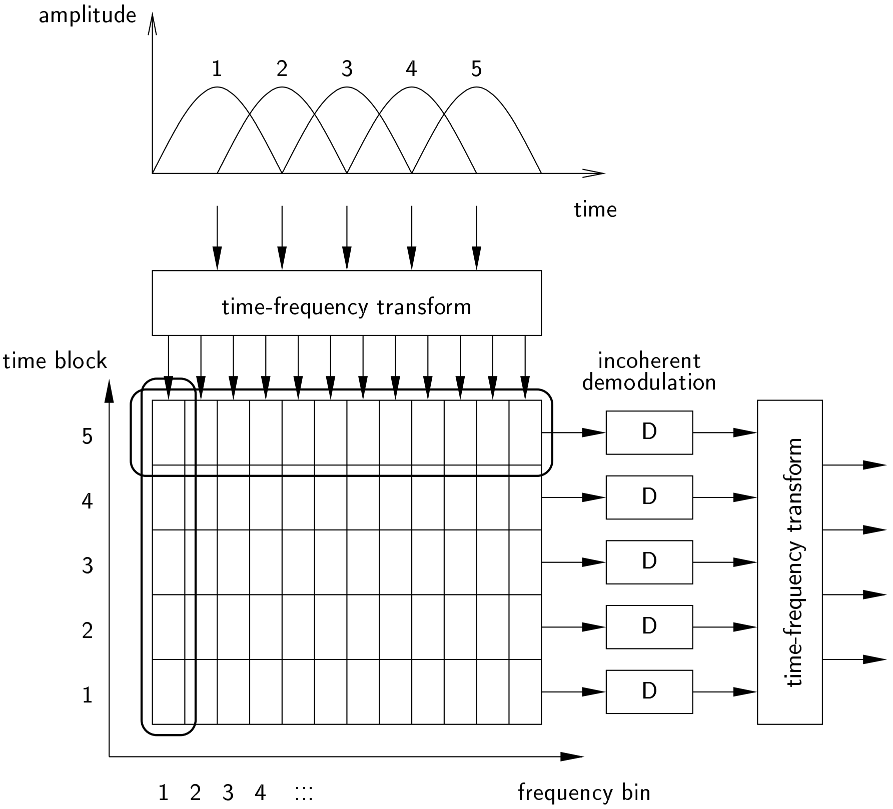


Figure 3.4: Signal processing to obtain *subband amplitude modulation spectra* (SB-AMS). For consecutive time blocks of an input signal, an initial time-frequency transform provides sets of complex spectral subband coefficients (rows). A group of successive spectral coefficient sets is stored in a two-dimensional buffer. For each spectral subband, a frame of temporally successive coefficients is assembled (column). This frame is incoherently demodulated and is further analyzed in its spectral content by a second transform. The procedure is repeated for all spectral subbands of the set, finally yielding SB-AMS data. Redrawn after [113].

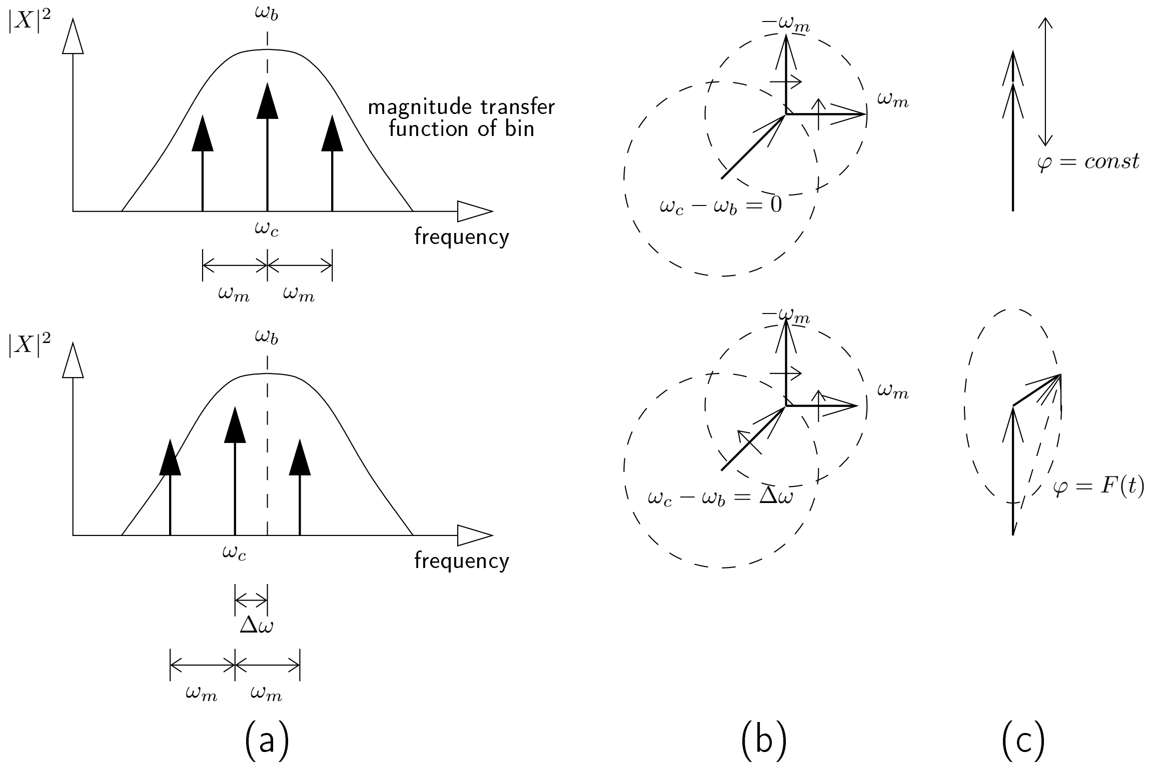


Figure 3.5: Example for demonstration of the mechanism of AM/FM and FM/AM conversion. A sinusoidal carrier ω_c modulated in its amplitude by a low frequency sinusoid ω_m results in a three tone signal, which is symmetric in its magnitude with respect to the carrier, as sketched in the left (a) top panel. If the carrier frequency coincides with the center frequency ω_b of a bin of the initial transform, the subband signal in this bin remains purely amplitude modulated, hence $\varphi = const$. This is confirmed by the graphical addition of the three phasors in the upper middle (b) panel. The result is a phasor of varying length denoting pure AM as depicted in upper right (c) panel. If there is a frequency offset $\Delta\omega$ between carrier and center frequency, as outlined in the left (a) bottom panel, the subband signal is asymmetrically damped by the subband filter transfer function and thus appears to also exhibit FM indicated by $\varphi = F(t)$, as sketched in the lower middle (b) and lower right (c) panel. Here, the resulting phasor draws an ellipse, indicating the presence of AM and FM.

Extended SB-AMS methods, which all have the fact that they use static bandpass filters in common, are published in [96], [99] and [98]. However, the central flaw of static decomposition into subbands is not addressed. In [97], Schimmel outlines a formalized framework for modulation filtering, comparing incoherent and coherent demodulation. In coherent demodulation, the estimation of AM is dependent on the estimated FM. As a main result, it is claimed that coherent demodulation is to be preferred over incoherent demodulation. Most importantly, however, are the initial thoughts to depart from the assumption of fixed carriers towards signal adaptive carriers: a novel approach is proposed which uses frequency tracking in order to adapt the initial subband decomposition filters. Influenced from the preceding publications on the matter, Li [61] later focused on signal adaptive coherent demodulation, thereby addressing the AM/FM conversion issue. His method, however, required a-priori knowledge of certain audio signal properties: the fundamental frequency f0 and the total number of carriers had to be known beforehand. Furthermore, the method was limited to the source model of voiced speech.

3.5.2 Relation to MAS diagram

SB-AMS denotes a system for modulation analysis and synthesis. Viewed in relation to the MAS diagram in Figure 3.1, it consists of an initial static uniform filterbank and subsequent AM estimation through the application of incoherent demodulation. The final AM estimate is represented in the modulation spectral domain. FM is not explicitly estimated, but is contained unresolved in the phase matrices. SB-AMS provides exact invertibility, but, due to the static filterbank and the lack of FM estimation, no means for perceptually adapted signal manipulation. An investigation of the shortcomings of this particular SB-AMS led to the approach presented in this thesis, of using a signal adaptive transform followed by both AM and FM estimation.

3.6 Centers of gravity

3.6.1 Description

Initially, Feth et al. [33] developed a model intended to describe the human ability to discriminate between the elements of a two tone complex that is applied as a stimulus. He referred to the observation first published by von Helmholtz [117], that the perceived pitch of a two tone complex is dependent on the amplitude relation of both components and is increasingly shifted towards the pitch of the component with dominating amplitude. This model was named *envelope weighted average of instantaneous frequency* (EWAIF). Equation 3.1 defines the EWAIF from the analytic signal $s(t) + i\hat{s}(t)$ as the integral of its IF, denoted by $f(t)$, weighted by its Hilbert envelope $e(t)$ and normalized by the integral of the Hilbert envelope. It is apparent that this integrating measure assigns the most weight on those points in time where the IF corresponds to a high amplitude of the associated temporal envelope. In contrast, spikes in the IF, e.g. of a two-tone signal [63], at points in time where the envelope is approximately zero, are suppressed in the EWAIF result.

$$EWAIF[s\left(t\right)] = \frac{\int_0^T e\left(t\right) f(t) dt}{\int_0^T e\left(t\right) dt} \tag{3.1}$$

Feth demonstrated that the discrimination ability can be predicted by calculating the EWAIF difference between two complementary stimuli. Later, Anantharaman et al. [1] proposed a slightly modified measure named *intensity weighted average instantaneous frequency* (IWAIF), which is closely related to the original EWAIF concept. The main advantage is a lower computational complexity and a better accuracy [1]. In contrast to EWAIF, the new measure can be calculated directly from the Fourier transform $S\left(f\right)$ of the signal $s\left(t\right)$, as defined by Equation 3.2

$$IWAIF\left[s\left(t\right)\right] = \frac{\int_0^\infty f \left|S\left(f\right)\right|^2 df}{\int_0^\infty \left|S\left(f\right)\right|^2 df} \tag{3.2}$$

Besides the interpretation analogous to EWAIF, substituting «amplitude» by «intensity», the authors provided another straight forward interpretation of their measure: describes the COG of the power spectrum. The COG corresponds to the «mean» frequency that is perceived by a human listener. In [120], this definition is used to determine the bandwidth of the so-called COG effect in vowel-like sounds. In [80], the authors suggest a modulation decomposition for application in cochlea implants and justify the proposed AM/FM decomposition strategy along the lines of that COG interpretation. They conclude that the incorporation of slowly varying FM into the signal processing significantly improves vowel intelligibility.

3.6.2 Relation to MAS diagram

The notion of the COG/IWAIF concept is, in general, useful for the estimation and interpretation of IF. It can be applied simply as an FM estimation technique, like in [80], where it was combined with a static filter bank. On the contrary, in this thesis the COG calculation is utilized as a method for the signal adaptive design of a filter bank that separates distinct spectral regions prior to AM/FM estimation, to obtain modulation estimates that are physically interpretable and that minimize the effect of AM/FM conversion (also see Subchapter 3.5.1).

3.7 Multiresolution spectro-temporal analysis

3.7.1 Description

Shamma et al. [121] developed an advanced model of the human auditory perception. In 1990 they published a model describing the nature of the outer stages (cochlea to middle brain) of the auditory system of mammals. The output of the model results in a perceptually adapted spectrogram. Additionally, a method is given how this non-linear model can be approximately reverted. The synthesis employs an iterative magnitude-only reconstruction method based on POCS. Following up, the model was amended by a second stage, modeling the functionality of the auditory cortex. The second stage implements a further decomposition into temporal-spectral modulation components [29][30]. Some applications based on this new model are proposed: the automatic detection of speech signals [71] and the improvement of speech intelligibility using modulation filtering [70]. Since the second processing stage is invertible, the entire model remains approximately invertible [70].

3.7.2 Relation to MAS diagram

Multiresolution spectro-temporal analysis aims at mimicking the human auditory system. For the initial spectral decomposition, a static filter bank is applied. The preprocessing consists of perceptually adapted dynamic compression and incorporates a hair-cell model. Since only magnitude information is further processed within the higher stages, the method only has approximate invertibility and relies on an iterative synthesis method. Modeling the human auditory system, including the cortical stages, already leads to a rather abstract representation of the audio signal. Thus, in this thesis, the view of [28] is shared, that multiresolution spectro-temporal analysis could be most beneficially applied within highly semantic scenarios like e.g. *computational auditory scene analysis* (CASA). The MODVOC described in this thesis presupposes a much lower semantic level of human auditory modeling since it predominantly relies on a physically motivated perceptual fusion premise.

3.8 Parametric coders with AM model

3.8.1 Description

Parametric audio coding schemes rely on a decomposition of the audio signal into short-term stationary sinusoids and, optionally, spectrally weighted and shaped noise. Such analysis/synthesis schemes were published, for example, in [69], [101] and [39]. In [13], such traditional sinusoidal modelling techniques are labeled by the term *constant amplitude* (CA). The necessary parameters are estimated for (overlapping) time blocks spanning approximately twenty milliseconds. Sinusoidal components are linked between adjacent blocks via parameter matching to ensure phase continuation. Since the absolute number of sinusoidal parameters varies from block to block, the linking of these components is usually subjected to a birth, continuation and death scheme [100]. An extended model named *harmonic and individual lines plus noise* (HILN) was proposed by Purnhagen and others [87][86], and standardized within the *Moving Pictures Expert Group* (MPEG). All CA methods have the fact that they approximate the tonal components of the signal by short term stationary sinusoids, or harmonic clusters thereof, in common.

Other publications extend the CA approach through the inclusion of amplitude modulated sinusoids in the model. In [40], a method is proposed that iteratively decomposes an audio signal into modulated sinusoidal components. As an approximation, these AM components are parametrized by truncated power series. The coefficients of the power series are estimated, minimizing a *mean square error* (MSE) criterion. In 2004, Christensen et al. [15] also outlined a coder based on the decomposition of audio signals into a sum of otherwise stationary sinusoids, but having a common temporal envelope. Optionally, this coder included an initial non-uniform multiband analysis. Through listening test results, the authors verified the benefit of a multiband analysis prior to AM modeling, especially for complex music mixtures originating from different sources. In [14], another coder having a similar configuration was proposed, that incorporated an initial signal decomposition into uniformly spaced subbands. The authors reported an improved perceptual reproduction quality for transient signals.

In [13], Christensen et al. published a method that extends parametric audio coding, based on sinusoids, by optionally allowing for amplitude modulation (AM) of each of the individual sinusoids. Compared to the traditional CA approach, they reported a significant listener preference for their AM/CA coder.

3.8.2 Relation to MAS diagram

Parametric coders having an AM model mainly aim at incorporating the short-term variability of sinusoidal components into their underlying model for e.g. improved reproduction of transient portions of the signal. Therefore, they can not be regarded as genuine modulation analysis/synthesis systems and do not fit well in the MAS diagram. At most, the approach of [15] can be interpreted within the MAS diagram. The method comprised an initial filter bank and an AM estimator in each subband. The weighted sum of stationary sinusoids in each subband can be viewed as a composite carrier. Thus, FM is not estimated explicitly, but modeled by the superposition of individual sinusoids.

Nevertheless, these techniques were introduced, since for the MODVOC system presented in Chapter 4 certain elements of such systems were adopted. Specifically, this applies for the idea of block-wise processing, along with component linkage by parameter matching and the application of a birth, continuation and death scheme.

3.9 Summary

In this chapter a *modulation analysis and synthesis* (MAS) scheme was proposed in order to conveniently range the different contributions to this research field that can be found in the literature. The MAS scheme consists of a (multiple) component carrier frequency estimation, a front-end filterbank to obtain multiband components, a component preprocessing and an *amplitude modulation* (AM) and *frequency modulation* (FM) estimation followed by a synthesis stage.

Secondly, publications that share a common history or have a similar approach were pooled into groups. Each group of publications was discussed, pointing out the underlying main ideas. These groups comprise vocoder related schemes, energy operators, minimum phase/all-phase decomposition, *subband amplitude modulation spectrum* (SB-AMS) approaches, *center of gravity* (COG) based methods, multiresolution spectro-temporal analysis and parametric coders with an AM model. Using the MAS diagram, the salient aspects of the different groups of publications were related to the novel MODVOC method proposed in this thesis. Said method adopts the multiband approach from vocoder schemes and combines it with the idea of signal adaptivity postulated in the minimum phase/all-phase decomposition group of publications. The need for signal adaptivity in multiband modulation decomposition is further supported by a description of the shortcomings that are immanent in SB-AMS schemes. To incorporate signal adaptivity into modulation decomposition in a perceptually meaningful way, the COG approach is adopted for the MODVOC and extended to a simultaneous estimation of multiple carrier frequencies. The AM/FM estimation of the MODVOC is based on Hilbert transform processing, rather than energy operator techniques, due to the superior stability under noisy conditions. Finally, the MODVOC synthesis method utilizes a parameter matching reminiscent of techniques found in parametric coders.

4 Modulation vocoder

The novel modulation vocoder that has been developed as the central contribution of this thesis is presented in this chapter. The MODVOC consists of an analysis, a modification and a synthesis procedure, based on multiband modulation components. Firstly, the basics of the approach are described. In the following, the analysis and synthesis are described in detail and different possibilities of signal processing in the modulation parameter domain are proposed. Finally, an additional technique is presented that addresses the processing of transient signal portions in the audio signal.

4.1 Introduction

The MODVOC denotes a system for modulation based multiband signal analysis, processing and synthesis [19]. Referring to the MAS diagram of Chapter 3, the analysis consists of a bank of bandpass filters that separate the broadband input signal into multiband components. These bandpass filters are designed signal adaptively to be aligned with spectral local *centers of gravity* (COG). The estimated COG, at the same time, correspond to the carrier frequencies of the multiband components. For each component, a preprocessing employs the Hilbert transform to obtain the analytic bandpass signal. The AM estimation is achieved by computing the absolute amplitude of the analytic signal (Hilbert envelope). The FM estimate is yielded by calculation of the *instantaneous frequency* (IF) derived from the phase of the analytic signal that is heterodyned by the carrier.

The synthesis consists of a block-wise reassembly of the multiband components. For each time block, the carrier is modulated in its frequency by the associated FM and subsequently integrated to obtain the signal phase by which a sinusoidal oscillator is loaded. The resulting signal is modulated in its amplitude by the associated AM. Finally, the contributions of all components are accumulated in order to synthesize the broadband output signal. To ensure continuity between successive blocks, the components of consecutive blocks are bonded by parameter matching and blending.

4.2 MODVOC analysis

4.2.1 Principle

Modulation analysis/synthesis schemes which decompose a broadband signal into a set of components, each comprising of a carrier, AM and FM information have several degrees of freedom, since this task is, in general, an ill-posed problem. To arrive at a meaningful modulation representation, additional requirements must be met. The approach of this thesis is to satisfy the condition that the extracted information is perceptually meaningful and interpretable, in a sense that modulation processing applied on the modulation information should produce perceptually smooth results, avoiding undesired artifacts introduced by the limitations of the modulation representation itself. The extracted carrier information alone must allow for a coarse, but perceptually pleasant and representative «sketch» reconstruction of the audio signal, and any successive application of AM and FM related information must refine this representation towards full detail, finally reaching perceptual *transparency*. The SB-AMS methods introduced in Chapter 3, for example, do not satisfy this constraint. The proposed multiband modulation analysis of this thesis is subject to the following requirements

- interpretability of the parameters
- scalability towards perceptual transparency
- high perceptual quality

These criteria are considered by the following processing paradigms

- incorporation of signal adaptivity
- provision of seamless spectral coverage
- utilization of a perceptually adapted representation

Signal adaptive front-end filters ensure that each component contains a spectral segment that can be regarded as a sonic entity. The carrier frequency of such a component represents the mean pitch elicited in a human listener by this spectral contribution. A seamless spectral coverage by the set of bandpass filters allows for scalability towards perceptual transparency. The properties of the human auditory perception is accounted for by designing the filters on a perceptually adapted ERB scale [73] representation of the spectrum.

In detail, the multiband modulation analysis dissects the audio signal into a signal adaptive set of analytic bandpass signals, each of which is further divided into a stationary sinusoidal carrier and its time-varying AM and FM. The set of bandpass filters is computed such that the full-band spectrum is covered seamlessly and the filters are each aligned with local COG. Accordingly, the carrier frequencies are defined to correspond to the local COG. In a sense, the bands aligned with local COG positions are equivalent to the classic *regions of influence* based phase locking of standard phase vocoders [57][27][90][91]. The bandpass signal envelope representation and the traditional region of influence phase locking both preserve the temporal envelope of a bandpass signal: either intrinsically or, in the latter case, by ensuring local spectral phase coherence during synthesis.

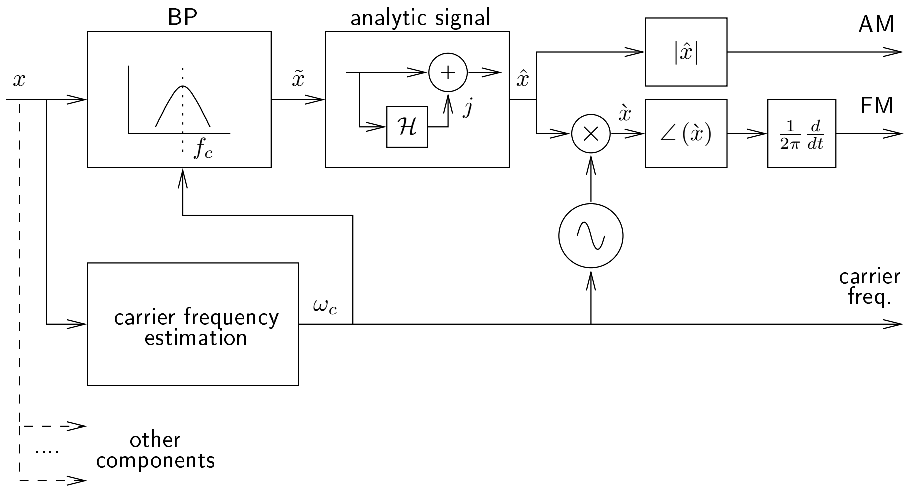


Figure 4.1: Modulation analysis.

4.2.2 Block diagram

A block diagram of signal decomposition into carrier signals and their associated modulation components is depicted in Figure 4.1. In this diagram, the schematic signal flow for the extraction of one of the multiband components is shown. All other components are obtained in a similar fashion. First, a broadband input signal x is fed into a bandpass filter that has been designed to signal adaptively yield an output signal, $\widetilde{x}$. Next, the analytic signal is derived via the Hilbert transform, according to Equation 4.1.

$$\widehat{x}(t) = \tilde{x}(t) + j\mathcal{H}(\tilde{x}(t)) \tag{4.1}$$

The AM is given by the amplitude envelope of $\widehat{x}$ (Equation 4.2)

$$AM(t) = |\widehat{x}(t)| \tag{4.2}$$

while the FM is obtained through the phase derivative of the analytic signal heterodyned by a stationary sinusoidal carrier with angular frequency ω_c (Equations 4.3). The carrier frequency is determined to be an estimate of the local COG. Hence, the FM can be interpreted as the IF variation at the carrier frequency f_c.

$$\begin{aligned} \grave{x}(t) &= \widehat{x}(t) \cdot \exp(-j\omega_c t) \\ FM(t) &= \frac{1}{2\pi} \cdot \frac{d}{dt}\angle(\grave{x}(t)) \end{aligned} \tag{4.3}$$

Since the estimation of local COG and the signal adaptive design of the front-end filterbank is one of the key parts of the proposed modulation analysis, it it described in Section 4.3.

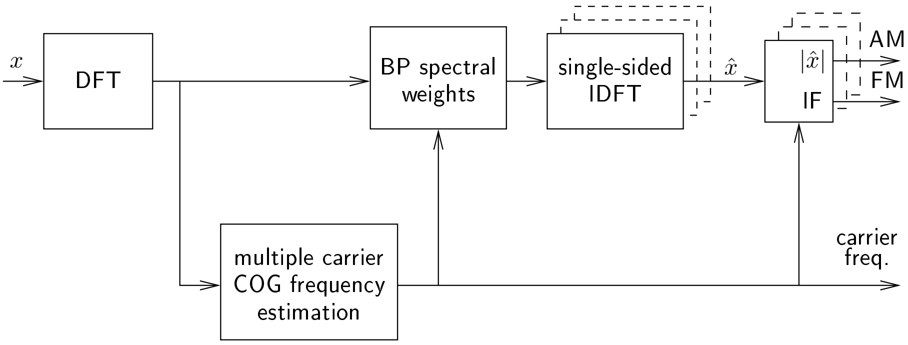


Figure 4.2: Implementation - Modulation analysis.

4.2.3 Implementation

Practically, in a discrete time system the component extraction is carried out jointly for all components, as illustrated in Figure 4.2. The proposed processing scheme supports real-time computation. The processing of a certain time block is only dependent on the parameters of previous blocks. Hence, no look-ahead is required in order to keep the overall processing delay as low as possible. The processing is computed on a block-by-block basis, using e.g. 75 % analysis block overlap and the application of a *discrete fourier transform* (DFT) on each windowed signal block. The window is a *flat top* window, according to Equation 4.4. This ensures that the centered $N/2$ samples that are passed on for the subsequent modulation synthesis utilizing 50 % overlap are unaffected by the skirts of the analysis window. A higher degree of overlap may be used for improved accuracy at the cost of increased computational complexity.

$$window\left(i\right)_{analysis} = \begin{cases} \sin^2\left(\frac{2i\pi}{N}\right) & 0 < i < \frac{N}{4} \\ 1 & \frac{N}{4} \le i < \frac{3N}{4} \\ \sin^2\left(\frac{2i\pi}{N}\right) & \frac{3N}{4} \le i < N \end{cases} \tag{4.4}$$

Given the spectral representation, a set of signal adaptive spectral bandpass weighting functions that is aligned with local COG positions is calculated. After applying the bandpass weighting to the spectrum, the signal is transferred into the time domain and the analytic signal is derived using Hilbert transform. These two processing steps can be efficiently combined by calculating a single-sided IDFT on each bandpass signal. Given the discrete time bandpass signal, the estimation of the IF (Equation 4.3) is implemented by phase differencing, as defined in Equation 4.5, where * denotes the complex conjugate. This expression is convenient, since it avoids phase ambiguities and, therefore, the need for phase unwrapping.

$$FM\left(n\right) = \angle\left(\hat{x}\left(n\right)\hat{x}\left(n-1\right)^\star\right) \tag{4.5}$$

4.3 Center of gravity estimation

4.3.1 Principle

The COG estimation and bandpass segmentation algorithm recently proposed in [20], extends the single COG model introduced in Section 3.6 towards the estimation of multiple local COG that are distributed in the frequency domain. The algorithm consists of an initial COG spectral position candidate list that is iteratively updated through refined estimates. In the refinement process, adding, deleting or fusing of candidates is incorporated, thus the method does not require a-priori knowledge of the total number of final COG estimates. The iteration is implemented by two loops. All necessary operations are performed on a spectral representation of the signal. The details are outlined in the following.

4.3.2 Preprocessing

For each signal block, a *power spectral density* (psd) estimate is obtained by computing the DFT spectral energy. Next, a mapping of the psd is performed onto a perceptual scale prior to COG calculation and segmentation, in order to facilitate the task of segmenting a spectrum into perceptually adapted non-uniform and, simultaneously, COG aligned bands. Thereby the problem is simplified to aligning a set of approximately uniform segments with the estimated local COG positions of the signal. As a perceptual scale, the ERB scale [73], is applied, which provides better spectral resolution at lower frequencies than e.g. the BARK scale. The mapped spectrum is calculated by interpolating the uniformly sampled spectrum towards spectral samples that are spaced following the ERB scale (Equation 4.6).

$$ERB\left(f\right) = 21.4 \log_{10}\left(0.00437 f + 1\right) \tag{4.6}$$

Subsequently, in order to remove the global trend inherent in real-world audio signal spectra, the mapped psd is normalized on its trend, which is calculated by linear regression, minimizing a least squares criterion. An example of an ERB mapped psd (gray) and its linear trend (black) is depicted in Figure 4.3.

Prior to division, both quantities are temporally smoothed by applying first order *infinite impulse response* (IIR) filters $H\left(z\right)$, each having a time constant of approximately $\tau = 200ms$, as defined by Equations 4.7, where T is the DFT subband sample period given by the input sample period times the temporal stride of the DFT.

$$\begin{aligned} H\left(z\right) &= \frac{1}{1 - a_1 z^{-1}} \\ a_1 &= \exp\left(-\frac{T}{\tau}\right) \end{aligned} \tag{4.7}$$

These preprocessing steps prevent a global bias towards low frequencies in the subsequent COG position iteration and stabilize the estimated positions for temporally successive blocks, respectively.

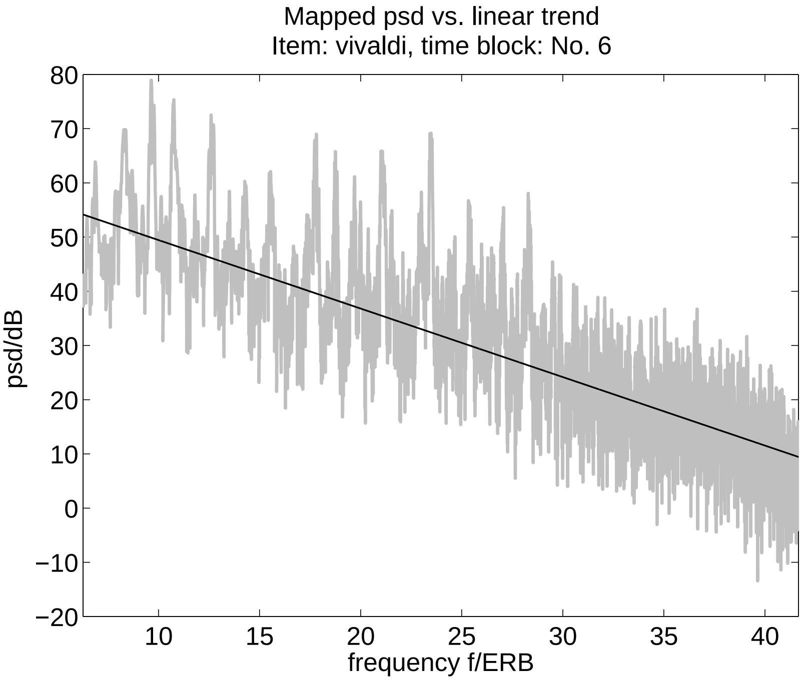


Figure 4.3: Mapped psd (gray) and linear trend (black).

4.3.3 Iterative center of gravity estimation

The iterative COG estimation flowchart is depicted in Figure 4.4. For each time block k, a sorted position candidate list c is initialized with a uniformly spaced grid of N candidate positions $c(n)$, having a spacing S (Equation 4.8). Most importantly, the parameter S sets the spectral resolution of the estimates obtained in the course of the iteration process. In other words, the parameter S determines what is considered to be the local scope of the COG estimation.

$$\begin{aligned} c\,(n) &= nS \\ n &\in [1, 2..., N] \end{aligned} \tag{4.8}$$

The iteration process consists of two loops. The first loop calculates the position offset $posOff\,(n)$ of the candidate position $c\,(n)$ from the true local COG, by applying a negative-to-positive linear slope function of size $2S$, weighted by the weights $g\,(i)$, to each candidate position n on the preprocessed psd estimate of a signal block (Equations 4.9).

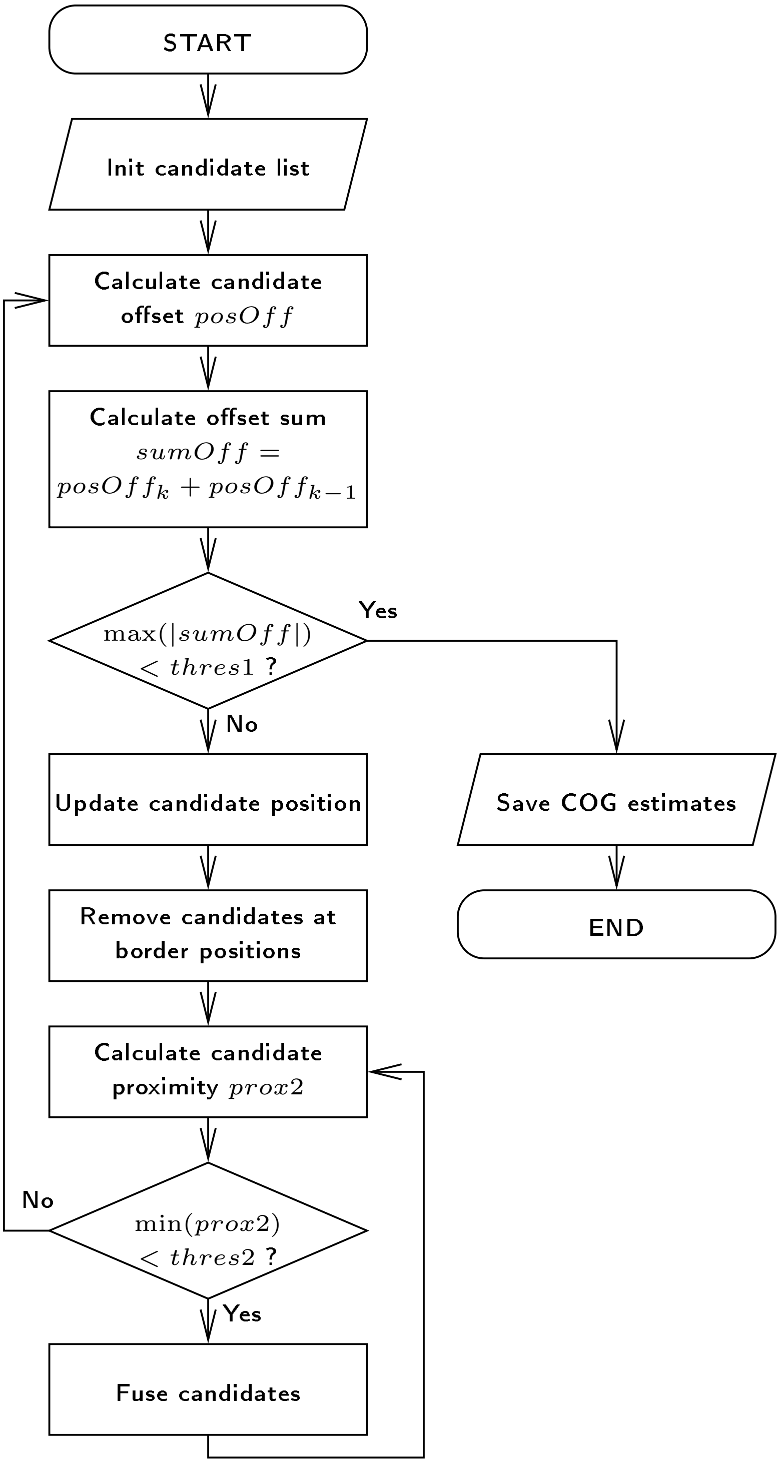

START
Init candidate list
Calculate candidate offset $posOff$
Calculate offset sum $sumOff = posOff_k + posOff_{k-1}$
$\max(|sumOff|) < thres1$?
Yes
No
Save COG estimates
END
Update candidate position
Remove candidates at border positions
Calculate candidate proximity $prox2$
$\min(prox2) < thres2$?
No
Yes
Fuse candidates

Figure 4.4: Flowchart of iterative COG estimation.

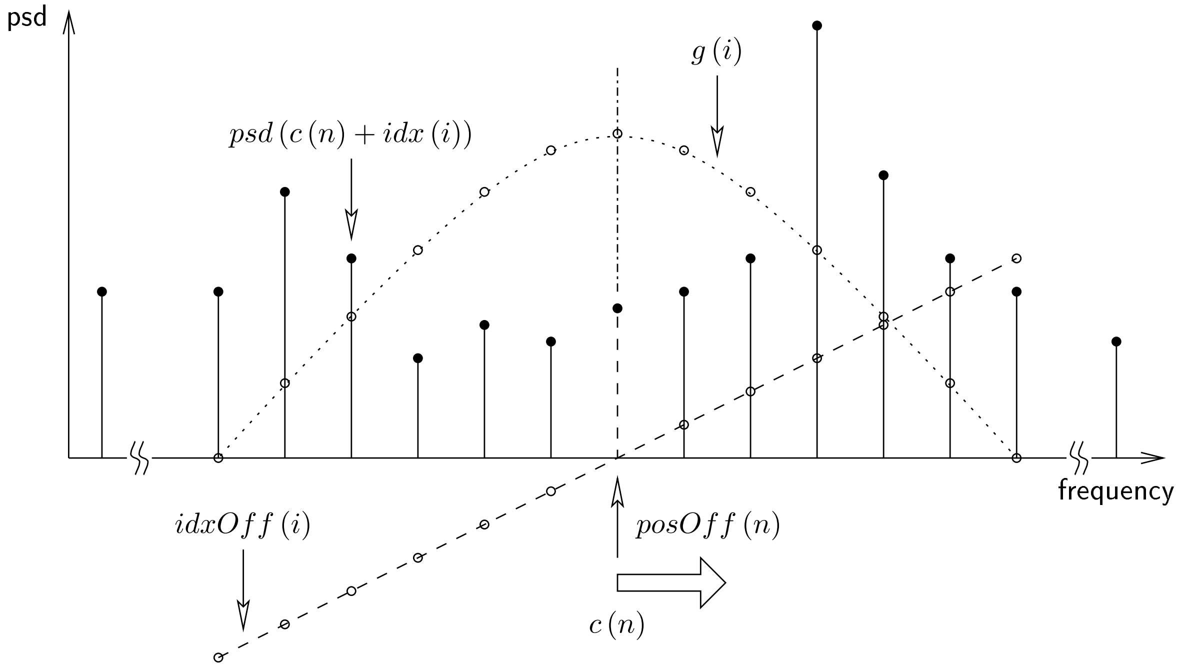


Figure 4.5: Iterative COG estimation.

$$\begin{aligned}
posOff\,(n) &= \operatorname{round}\left(\frac{\sum_i \left(w_n\,(i)\cdot idxOff\,(i)\right)}{\sum_i w_n\,(i)}\right)\\
w_n\,(i) &= \operatorname{psd}\left(c\,(n)+idx\,(i)\right)\cdot g\,(i)\\
idxOff\,(i) &= i-S+0.5\\
idx\,(i) &= \operatorname{round}\left(idxOff\,(i)\right)\\
i &\in [0,1,2...,2S-1]
\end{aligned} \tag{4.9}$$

In Figure 4.5, the candidate position offset $posOff\,(n)$ procedure is visualized. The stem plots correspond to the local psd samples $psd\,(c\,(n)+idx\,(i))$, centered at the candidate position $c\,(n)$. The window function is represented by values $g\,(i)$ and the linear slope function is denoted by $idxOff\,(n)$.

Next, all candidate positions from the list are updated by their offset position (Equation 4.10).

$$c\,(n) := c\,(n) + posOff\,(n) \tag{4.10}$$

Each candidate position that violates the border limitations is removed from the list, as indicated by Equations 4.11, and the number of remaining candidate positions N is decremented by 1.

$$\begin{aligned} &\text{if } \left(c\left(n\right) < S\right) \vee \left(c\left(n\right) > NS\right) \rightarrow \\ c\left(x\right) &:= c\left(x+1\right)\ \forall x \in \left[n+1, ..., N-1\right] \\ N &:= N-1 \end{aligned} \tag{4.11}$$

If the absolute value of the sum of the actual and the previous position offsets of a candidate, as defined in Equation 4.12, is smaller than a predefined threshold, this candidate position $c\left(n\right)$ is not updated in further iterations, but remains in the list and is, thus, subjected to the subsequent candidate fusion mechanism.

$$sumOff\left(n\right) = posOff_k\left(n\right) + posOff_{k-1}\left(n\right) \tag{4.12}$$

If the $\left|sumOff\left(n\right)\right|$ of all candidates is smaller than a predefined threshold (Equation 4.13), the first iteration loop is exited, hereby terminating the iteration process. All remaining candidates from the list constitute the final set of COG position estimates. Note that using this type of condition also ends the iteration if the position offset toggles back and forth between two values, hereby always ensuring proper termination.

$$\max\left(\left|sumOff\left(n\right)\right|\right) < thres1 \tag{4.13}$$

The second loop iteratively fuses the closest (according to a certain proximity measure) two position candidates that violate a predefined proximity restriction due to the position update provided by the first loop, into one single new candidate, thereby accounting for perceptual fusion. The proximity measure $prox2$ is the spectral distance of the two candidates (Equations 4.14).

$$\begin{aligned} prox2 &< thres2 \\ prox2 &= \left|c\left(n\right) - c\left(n+1\right)\right| \\ thres2 &:= S \end{aligned} \tag{4.14}$$

Each newly calculated joint candidate is initialized to occupy the energy weighted mean position of the two former candidates (Equations 4.15).

$$\begin{aligned} c\left(n\right) &:= \text{round}\left(\frac{w\left(n\right)c\left(n\right) + w\left(n+1\right)c\left(n+1\right)}{w\left(n\right) + w\left(n+1\right)}\right) \\ w\left(n\right) &= \sum_i w_n\left(i\right) = \sum_i \left(\text{psd}\left(c\left(n\right) + idx\left(i\right)\right) \cdot g\left(i\right)\right) \\ c\left(x\right) &:= c\left(x+1\right)\ \forall x \in \left[n+1, ..., N-1\right] \\ N &:= N-1 \end{aligned} \tag{4.15}$$

Both former candidates are deleted from the list and the new joint candidate is added to the list. Consequently, the number of remaining candidate positions N is decremented by 1. The inner loop iteration terminates if no more candidates violate the proximity restriction.

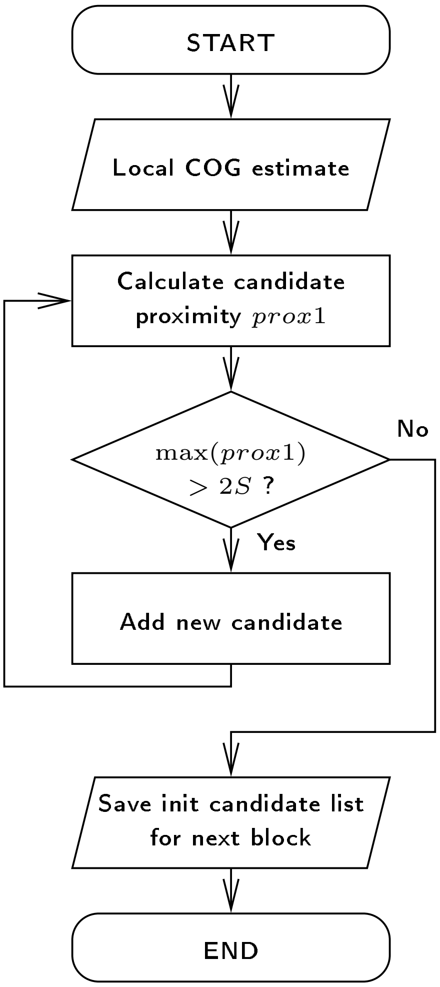

START
Local COG estimate
Calculate candidate proximity $prox1$
$\max(prox1) > 2S$?
No
Yes
Add new candidate
Save init candidate list for next block
END

Figure 4.6: Flowchart of improved initialization.

4.3.4 Improved initialization

To speed up the iteration process, the initialization of each new block can advantageously be done using the COG position estimate of the previous block, since it is already a fairly good estimate of the current positions. This applies due to the block overlap in the analysis and the temporal smoothing in the preprocessing, and, hence, the assumption of a limited change rate in temporal evolution of COG positions is well justified.

Still, it has to be ensured that a sufficiently large number of initial position candidates exist to capture the possible emergence of new COGs. Therefore, position candidate gaps in the estimate spanning a distance greater than $2S$ are filled by new COG position candidates (Equations 4.16), thus warranting that potential new COGs are within the scope of the position update function. Figure 4.6 shows a flow chart of this extension to the algorithm.

The apposition of additional candidates to the list is accomplished with a loop that terminates if no more gaps larger than $2S$ are found.

$$\begin{aligned}
\text{if } prox1 &> 2S \;\rightarrow \\
prox1 &= c(n+1) - c(n) \\
c(x+1) &:= c(x) \;\forall x \in [N, N-1, ..., n+1] \\
c(n+1) &:= \text{round}\left(\frac{c(n) + c(n+1)}{2}\right) \\
N &:= N+1
\end{aligned} \tag{4.16}$$

4.3.5 Design of bandpass filter set

After having the COG estimates in the ERB adapted domain determined, a set of N bandpass filters is calculated in the form of a spectral weighting functions $weights_n$, of length M, according to Equations 4.17. The bandpass filters are designed to have a pre-defined roll-off of length $2 \cdot rollOff$, with sine-squared characteristics. To achieve the desired alignment with the estimated COG positions, the design procedure described in the following is applied.

Firstly, the middle positions between adjacent COG position estimates are calculated, where $m_L(n)$ denotes the lower midpoint, and $m_U(n)$ the upper midpoint of a COG position $c(n)$ relative to its neighbors. Then, at these transition points, the roll-off parts of the spectral weights are centered such that the roll-off parts of neighboring filters sum up to one. The middle section of the bandpass weighting function is chosen to be flat-top equal to one, the remaining sample points are set to zero. The filters for $n = 0$ and $n = N$ only have one roll-off part and exhibit lowpass or highpass characteristics, respectively.

$$\begin{aligned}
weights_n(m) &= \begin{cases} \sin^2(k_L(m)) & m_L(n) - rollOff < m < m_L(n) + rollOff \\ 1 & m_L(n) + rollOff \le m \le m_U(n) - rollOff \\ \sin^2(k_U(m)) & m_U(n) - rollOff < m < m_U(n) + rollOff \\ 0 & otherwise \end{cases} \\
m &\in [0, 1..., M-1] \\
m_L(n) &= \text{round}\left(\frac{c(n) - c(n-1)}{2}\right) \\
m_U(n) &= \text{round}\left(\frac{c(n+1) - c(n)}{2}\right) \\
k_L(m) &= (m - m_L(n) + rollOff)\,\frac{\pi}{4 \cdot rollOff} \\
k_U(m) &= (m - m_U(n) - rollOff)\,\frac{\pi}{4 \cdot rollOff} + \frac{\pi}{2}
\end{aligned} \tag{4.17}$$

In designing the roll-off part, a trade-off has to be made with respect to spectral selectivity and temporal resolution. Also, allowing multiple filters to spectrally overlap may

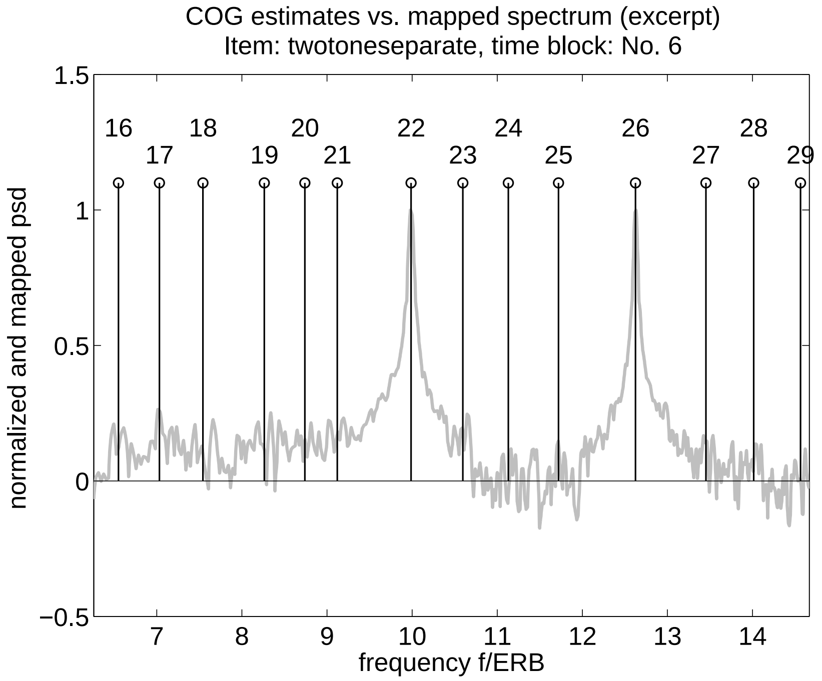


Figure 4.7: Two separate tones - Local centers of gravity (black, stem plot) vs. mapped spectrum (gray, line plot).

add an additional degree of freedom to the design restrictions. The trade-off may be chosen in a signal adaptive fashion for e.g. improving on the reproduction of transients.

Lastly, the COG positions and the spectral weighting functions are mapped back to the linear domain by solving Equation 4.6 for f, obtaining Equation 4.18. Finally, the spectral weights on a linear scale are yielded, which are to be applied to the original DFT spectrum of the broadband signal.

$$f\left(ERB\right) = \frac{1}{0.00437}\left(10^{\frac{ERB}{21.4}} - 1\right) \tag{4.18}$$

4.3.6 Examples of carrier estimation and spectral segmentation

Figures 4.7, 4.8, 4.9 and 4.10 visualize results obtained by the proposed iterative local COG estimation algorithm of Subsection 4.3.3 that has been applied to different test items.

The test items are two spectrally well separated pure tones, two closely spaced tones that cause beat effects, plucked strings («MPEG Test Set - sm03») and orchestral music («Vivaldi - Four Seasons, Spring, Allegro»). In these figures, the perceptually mapped, smoothed and globally detrended spectrum is displayed (gray, line plot) along with the COG estimates (black, stem plot). The COG estimates are numbered in ascending order. The estimates no.22, no.26 of Figure 4.7 and estimates no.18 and no.19 of Figure 4.9 correspond to sinusoidal signal components. Other than that, estimate no.22 of Figure

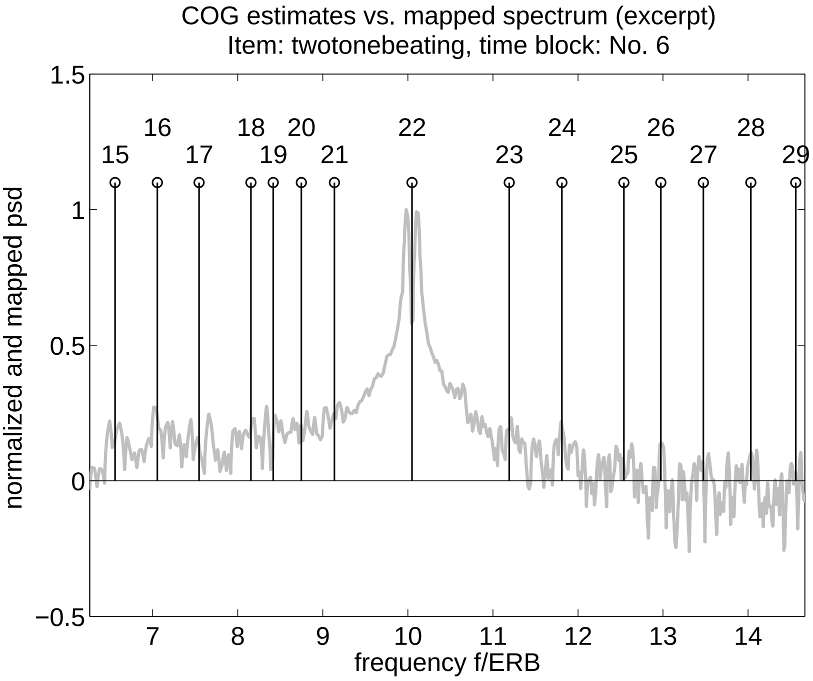


Figure 4.8: Two beating tones - Local centers of gravity (black, stem plot) vs. mapped spectrum (gray, line plot).

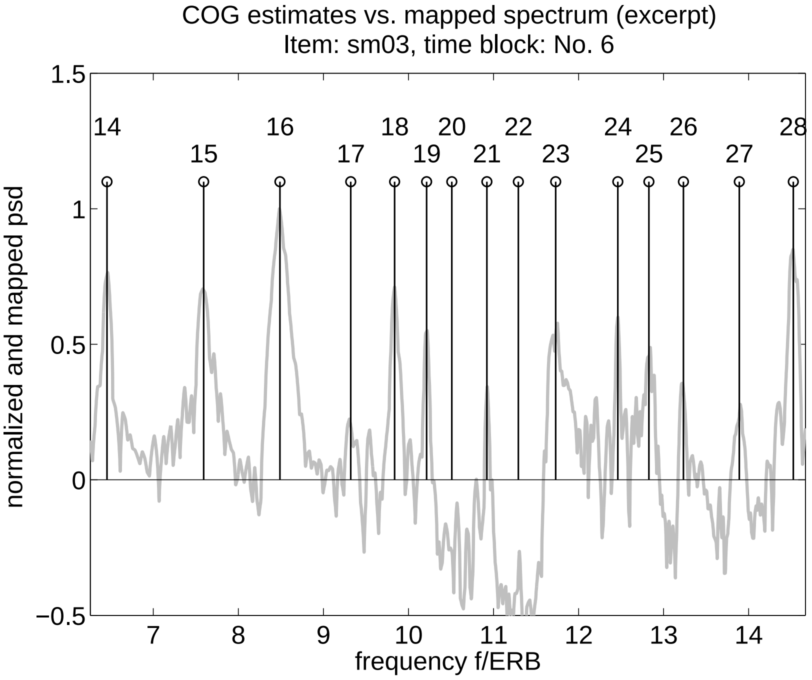


Figure 4.9: Plucked strings - Local centers of gravity (black, stem plot) vs. mapped spectrum (gray, line plot).

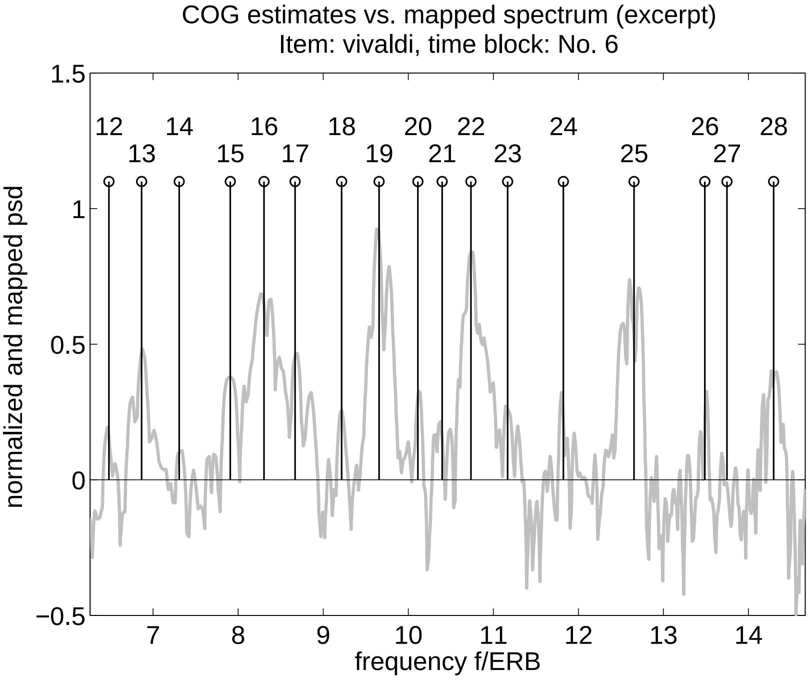


Figure 4.10: Orchestral music - Local centers of gravity (black, stem plot) vs. mapped spectrum (gray, line plot).

4.8, estimates no.23 and no.25 of Figure 4.9 and most estimates of Figure 4.10 capture spectrally broadened or beating components. These spectrally broadened components were nevertheless well handled through the proposed algorithm since they were also grouped into perceptual units.

In Figures 4.11 and 4.12, the original non preprocessed psd of the signal block is depicted (gray) and a set of bandpass filters (black) is sketched as outlined in Subsection 4.3.5. It is clearly visible that each filter is aligned with a COG estimate and pairwise smoothly overlaps with its adjacent subband filters.

4.4 MODVOC synthesis

4.4.1 Principle

Like the analysis, the synthesis is performed on a block-by-block basis. Since only the centered $N/2$ portion of each analysis block is evaluated for synthesis, it results in a synthesis overlap factor of 50%. A blending of successive blocks is applied in the parameter domain rather than on the readily synthesized signal, in order to avoid phase cancellation effects between adjacent time blocks. The blending is controlled by a component bonding mechanism.

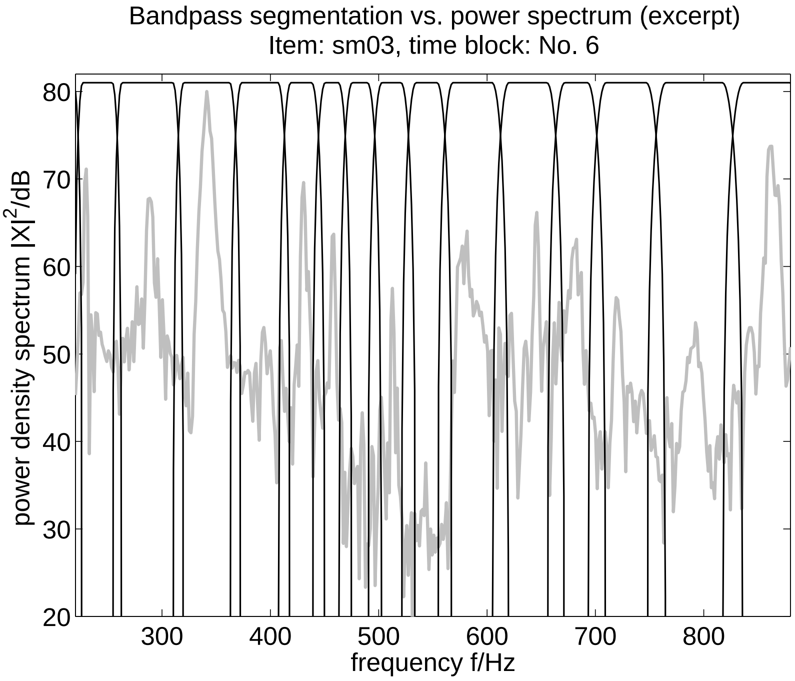


Figure 4.11: Plucked strings - Bandpass filters (black) aligned with local centers of gravity vs. power spectrum (gray).

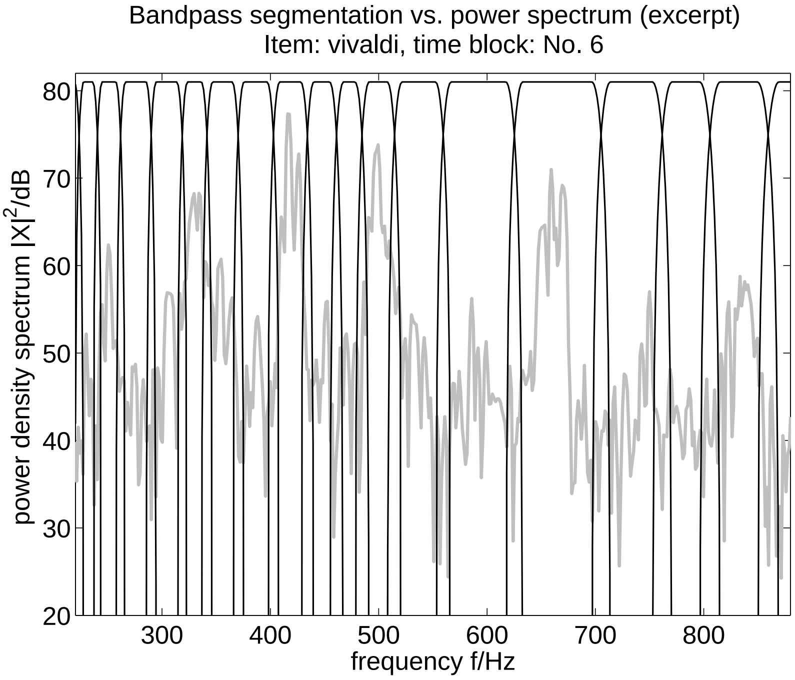


Figure 4.12: Orchestral music - Bandpass filters (black) aligned with local centers of gravity vs. power spectrum (gray).

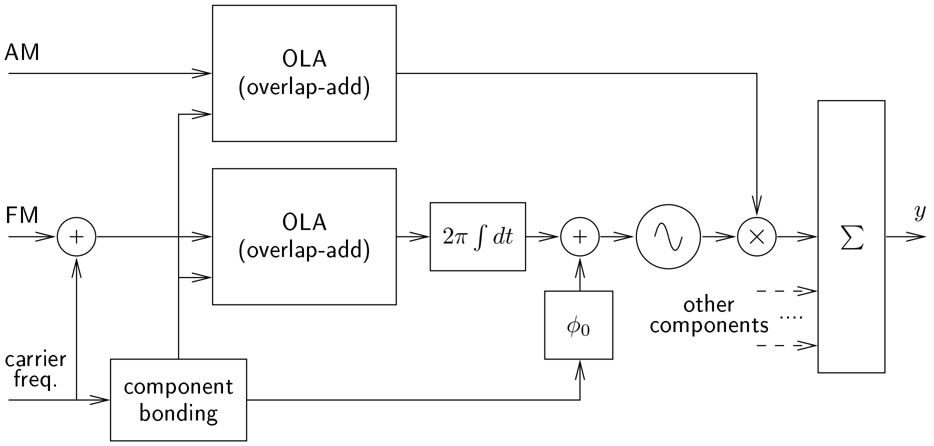


Figure 4.13: Modulation synthesis.

4.4.2 Block diagram

The signal is synthesized on an additive basis of all components. Successive blocks are blended by *overlap-add* (OLA), which is controlled by the bonding mechanism. The processing chain for one component is shown in Figure 4.13. First, the FM signal is added to the stationary carrier frequency and the resulting signal is passed on to an OLA stage, the output of which is subsequently temporally integrated. A sinusoidal oscillator is fed by the resulting phase signal. The AM signal is processed by a second OLA stage. Next, the output of the oscillator is modulated in its amplitude by the AM signal to obtain the additive contribution of the component to the output signal. Finally, the contributions of all components are summed to obtain the output signal y.

4.4.3 Component bonding

The component bonding ensures a smooth transition between the borders of adjacent blocks, even if the components are substantially altered by a modulation domain processing. The bonding performs a pair-wise match of the components of the actual block to their predecessors in the previous block. Additionally, the bonding aligns the absolute component phases of the actual block to the ones of the previous block. If no modulation processing is intended, the OLA controlled by the bonding may be disabled and only the absolute phase of each component has to be adjusted to reassemble the original phase in order to obtain perfect reconstruction.

The bonding is determined by an iterative algorithm steered by the spectral vicinity of component carriers measured on an ERB scale. A flowchart is depicted in Figure 4.14. For each element of a list containing the actual component carriers, the spectral distances to all predecessor candidates are calculated. The predecessor of an actual component is chosen to be the component having minimum spectral distance of component carriers, if this minimum distance is below a pre-defined bonding threshold. Subsequently, the bonded component pair is removed from the list. The iteration terminates

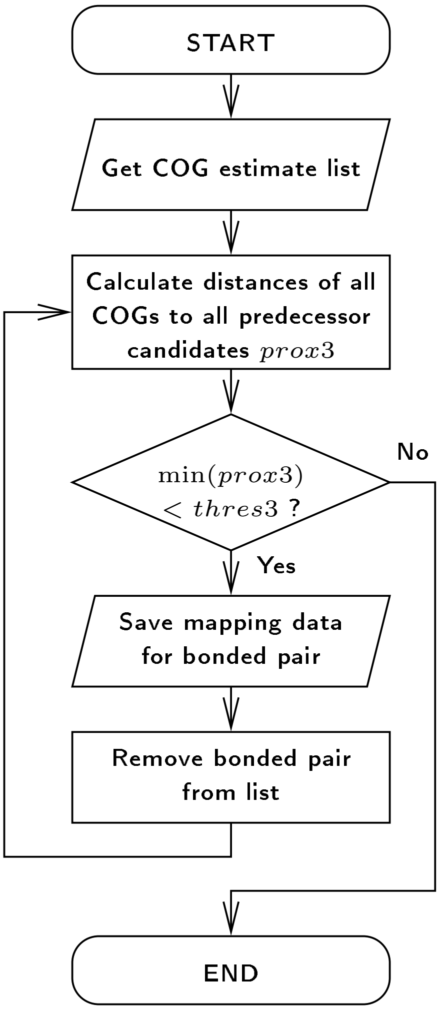

START
Get COG estimate list
Calculate distances of all COGs to all predecessor candidates $prox3$
$\min(prox3)$ $< thres3$?
No
Yes
Save mapping data for bonded pair
Remove bonded pair from list
END

Figure 4.14: Flowchart of bonding algorithm.

if all distances of non-bonded actual components to predecessor candidates are above the bonding threshold. Components of the previous block without an associated successor are terminated via fade-out within that block. Components of the actual block without an assigned predecessor are faded-in. Bonded component pairs are linked by OLA. This principle is reminiscent of the birth, continuation and death scheme common to standard parametric coders, as described in Section 3.8.

4.4.4 Examples of component bonding

In the following, component bonding is visualized using excerpts of two items containing plucked strings and classical orchestral music, respectively. Figures 4.15 and 4.18 show the spectrograms of two different input signals. From the spectrogram, it is clearly visible that the orchestral item is harmonically much more complex and dense than the single instrument item. Nevertheless, the COG aligned carriers can be reliably tracked for both items. Figures 4.16 and 4.19 sketch the carriers of the modulation components that are linked by the bonding algorithm, as described in Subsection 4.4.3. For each time block, the components are numbered in ascending spectral order. The block numbers correspond to the numbers of the carrier estimation examples given in Subsection 4.3.6. Solid lines indicate components that are linked together through bonding, while dotted lines indicate components that are faded-in or faded-out, due to the lack of an adequate bonding partner. Figures 4.17 and 4.20 display, again, the same input signal spectrograms, but superimposed by the component bonding data. It can be seen that the carrier locations indicated by line plots coincide with local energy concentrations visualized by light color in the spectrogram.

4.5 MODVOC processing

Having interpretable modulation components at hand, new and interesting processing methods become feasible. A great advantage of the modulation decomposition presented in this thesis is that the proposed analysis/synthesis method implicitly assures that the result of any manipulation - to a large extent independent from the exact nature of the processing - will be perceptually smooth, e.g. free from clicks and repetitions of transients. In the following, two main examples of modulation processing possibilities are suggested

- manipulation of auditory roughness
- alteration of musical pitch

By smoothing or filtering the AM and FM modulation components, the auditory *roughness* [125][105] of a signal can be altered. In the AM signal, there is a coarse structure related to onset and offset of musical events etc. and a fine structure related to faster modulation frequencies of approx. 30 - 300 Hz. Since for carriers up to 2 kHz this fine structure is strongly correlated to the roughness properties of an audio signal [105][17],

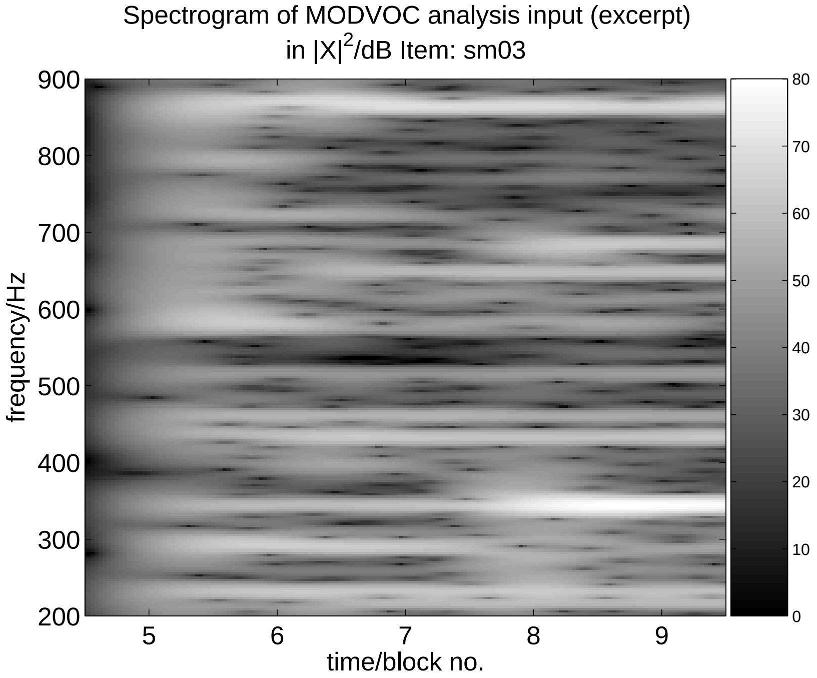


Figure 4.15: Plucked Strings - Spectrogram of MODVOC analysis input.

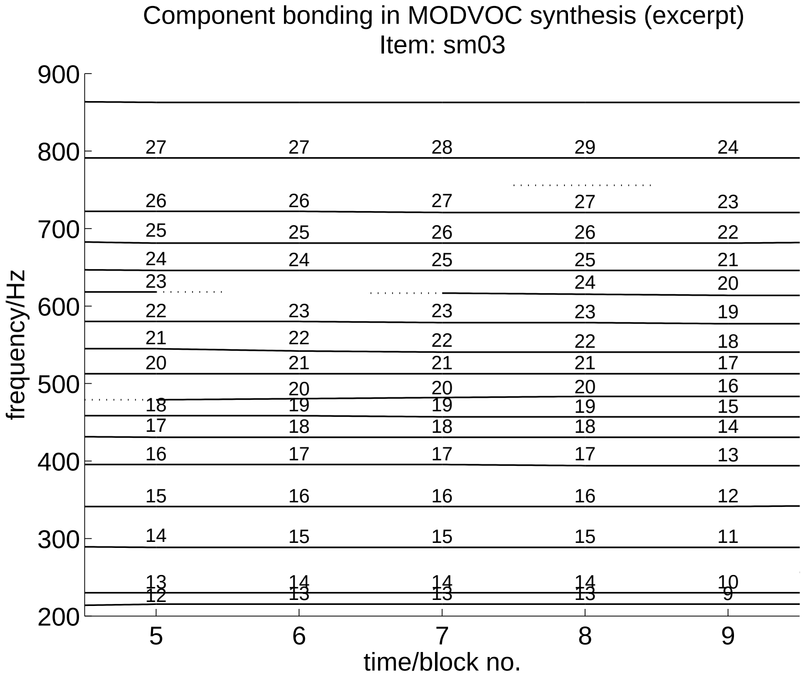


Figure 4.16: Plucked Strings - Component bonding in MODVOC synthesis: bonded carriers (solid lines), fade-in or fade-out carriers (dotted lines).

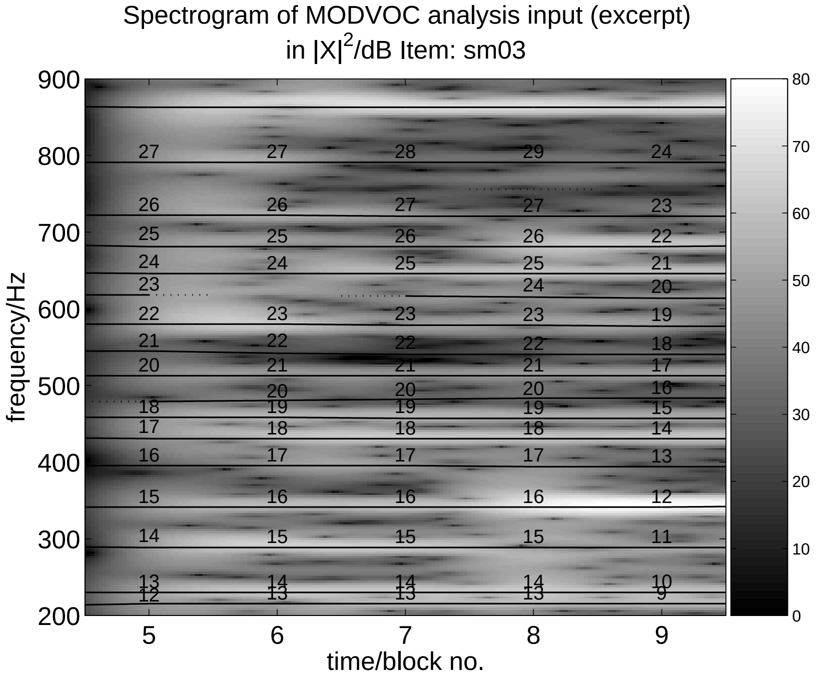


Figure 4.17: Plucked Strings - Spectrogram of MODVOC analysis input superimposed by bonding data.

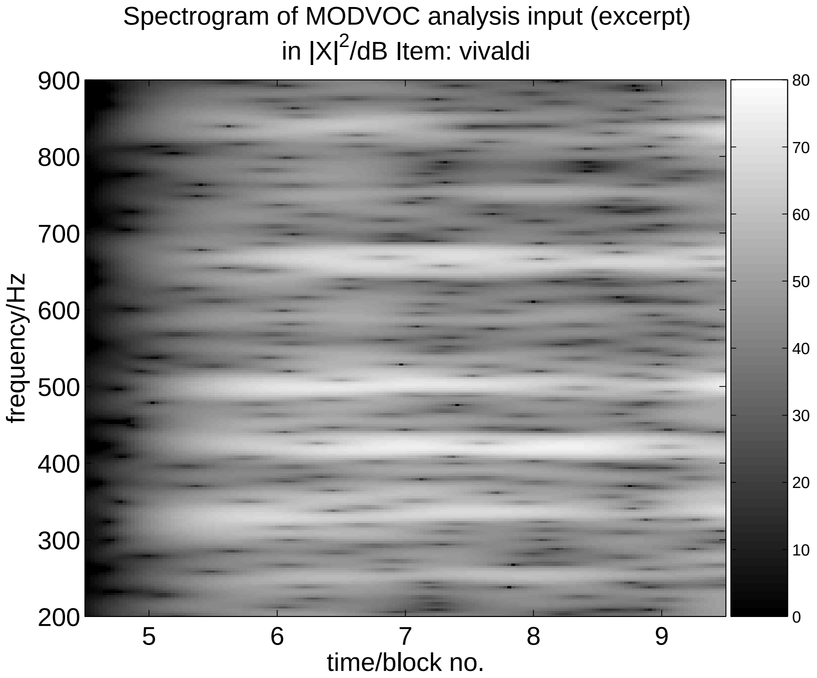


Figure 4.18: Orchestral music - Spectrogram of MODVOC analysis input.

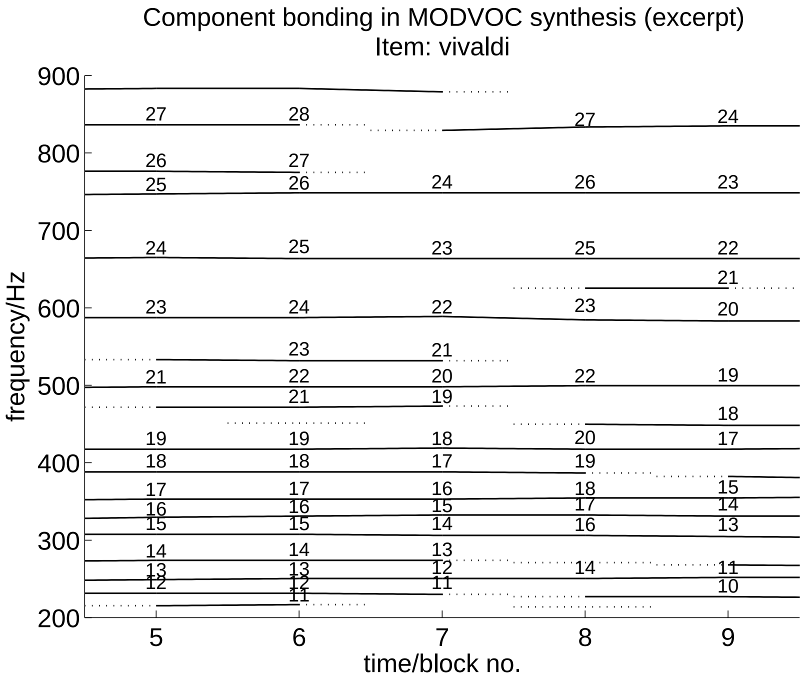


Figure 4.19: Orchestral music - Component bonding in MODVOC synthesis: bonded carriers (solid lines), fade-in or fade-out carriers (dotted lines).

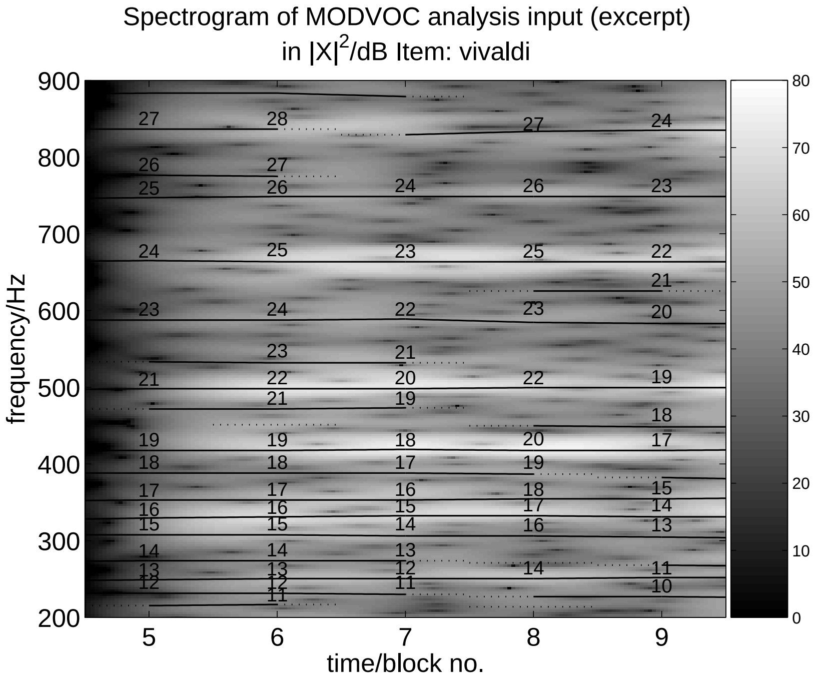


Figure 4.20: Orchestral music - Spectrogram of MODVOC analysis input superimposed by bonding data.

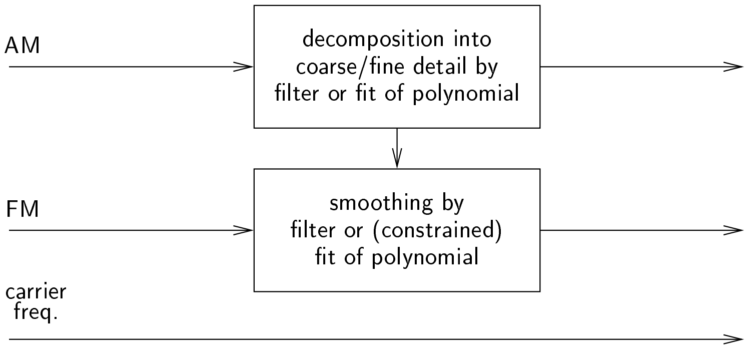

AM
decomposition into coarse/fine detail by filter or fit of polynomial
FM
smoothing by filter or (constrained) fit of polynomial
carrier freq.

Figure 4.21: Auditory roughness manipulation - Modulation processing.

auditory roughness can conversely be modified by removing the fine structure and maintaining the coarse structure. A processing scenario for roughness manipulation is shown in Figure 4.21. In order to decompose the AM into coarse and fine structure, linear filters or, alternatively, nonlinear methods can be utilized. For example, to capture the coarse AM one can apply a piecewise fit of a low order polynomial. The fine structure is represented in the residual signal and can be obtained as the difference of original AM and the coarse AM. Note that if any modifications are applied to the AM signal, it is advisable to restrict the FM signal to only be slowly varying, since the unprocessed FM may contain sudden peaks, due to beating effects inside one bandpass region [63][119]. These peaks appear in FM at the proximity of zero [60] of the related AM signal. These undesired peaks can be removed by e.g. constrained polynomial fitting on the FM, where the original AM signal acts as weights for the desired goodness of the fit. Thus, spikes in the FM can be removed without introducing an undesired bias. This approach may be additionally justified by the same considerations as the EWAIF method [33][1], introduced in Section 3.6. Both methods apply weights derived from the AM envelope to the FM signal in order to obtain a physically interpretable frequency variation measure.

A prominent audio effect is the *transposition* [74][58][124] of audio signals, while maintaining original playback speed. The necessary MODVOC processing is depicted in Figure 4.22. The effect is achieved through the multiplication of the carriers with a constant transposition factor. By also multiplying the FM with the same factor it is ensured that, for each component, the relative FM modulation depth is preserved. Since the temporal structure of the input signal is solely captured by the AM signals, it is unaffected by the processing. In Chapter 5, it will be shown that - beyond the possibilities of existing pitch transposition schemes - the MODVOC can perform a frequency selective transposition of music signals.

Another well-known effect is *time stretching* [32][41][93][112][57], which can be seen as the dual operation to transposition. Time stretching denotes the effect of temporal dilatation or compression of a signal while preserving its original pitch. A transposition of a signal by a certain factor can always be converted into a time stretching effect by subsequent resampling of the processed signal [124].

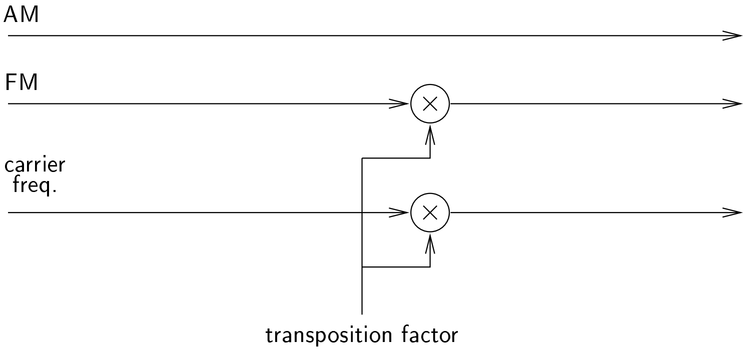


Figure 4.22: Transposition - Modulation processing.

4.6 Envelope shaping

As stated in Subsection 4.2.1, the MODVOC processing preserves spectral coherence in the passband area surrounding the carrier locations. However, the broadband global spectral coherence is not preserved. For quasi-stationary signals this has only minor impact on the perceptual quality of the synthesized signal. If the signal contains prominent transients like drum beats or castanets, the preservation of global coherence by temporal *envelope shaping* (ES) can greatly improve the reproduction quality of these signals [22].

The preservation of global coherence can be addressed by linear prediction in the spectral domain. For quite a while, similar approaches have been utilized in audio codecs, for instance by the *temporal noise shaping* (TNS) tool [46][48][47][45] in MPEG 2/4 *advanced audio coding* (AAC). Figure 4.23 outlines the integration of this technique into the MODVOC processing scheme. In the analysis, subsequent to the initial DFT of the input signal x, *linear prediction coefficients* (LPC) of a forward predictor along frequency direction having the impulse response $h(\omega)$ are derived by e.g. the autocorrelation method minimizing the prediction error in a least squares sense. Subsequently, the filter is applied to the spectral values, and the residual signal is further processed by the MODVOC algorithm. The filter coefficients, representing the global envelope, are conveyed to the synthesis stage. In the synthesis, the global envelope, derived by evaluation of the prediction filter on the unit circle $|H(e^{jt})|$, is restored by a multiplicative application of the same to the sum signal yielding the envelope shaped output signal y, as illustrated in Figure 4.24. The engagement of envelope shaping can be switched on or off signal adaptively depending on the *prediction gain*, which is defined to be the energy ratio of the signal and prediction error.

4.7 Summary

In this chapter, the MODVOC is proposed, denoting a multiband modulation analysis, processing and synthesis system for arbitrary audio signals. The MODVOC operates on successive, overlapping time blocks of an input signal. The multiband decomposition

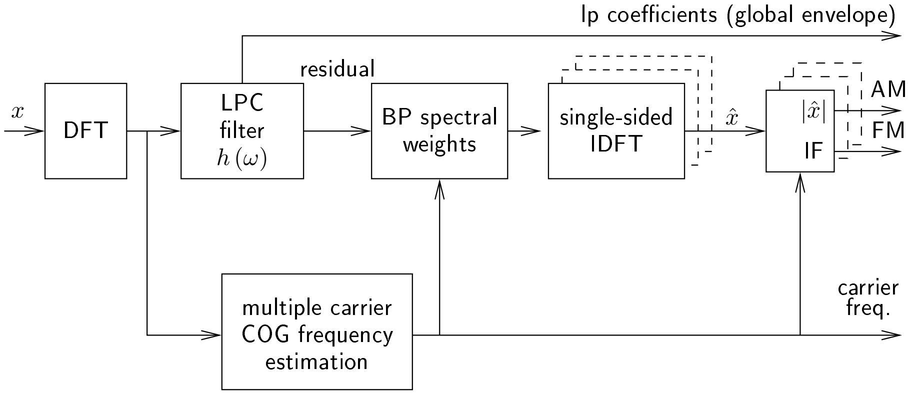


Figure 4.23: Linear prediction - Modulation analysis.

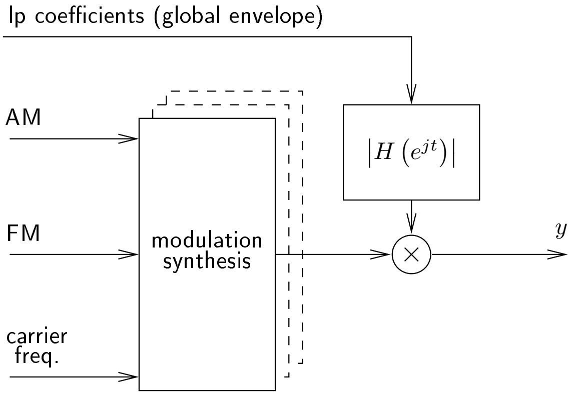


Figure 4.24: Linear prediction - Modulation synthesis.

front-end of the MODVOC analysis is based on the estimation of spectral local COG. A design scheme for a set of bandpass filters aligned to the estimated COG positions has been outlined. These filters are subsequently utilized to separate the broadband signal into signal dependent perceptually adapted multiband components. Due to the alignment with spectral local COG on the perceptually adapted ERB scale, it is claimed that meaningful and intuitively interpretable AM and FM parameters can be derived from the subband signals. This claim will be verified through listening tests in Chapter 6.

The MODVOC synthesis generates the output signal on an additive basis of all components. Successive synthesis blocks are linked by a component bonding mechanism using a parameter matching of the associated component carrier frequencies. In order to improve the perceptual quality of transients in audio signals processed by the proposed system, *envelope shaping* (ES) by linear prediction in the spectral domain can be incorporated in the MODVOC scheme.

Lastly, two exemplary modulation processing methods have been presented, one targeting auditory roughness manipulation and the other implementing means for global or even frequency selective pitch transposition of audio signals.

5 Application of MODVOC to frequency selective pitch transposition

In this chapter, the application of the proposed MODVOC to frequency selective pitch transposition of polyphonic audio signals is presented, offering the possibility to alter the key and the musical scale mode of chords, arpeggios or even complex sound mixes. Firstly, a brief introduction into music theory is given. Subsequently, suitable MODVOC operation parameter settings are derived from music theory and psychoacoustics. Lastly, an adapted MODVOC processing scheme is described, which implements the task at hand.

5.1 Music theory primer

5.1.1 Scales, tones and intervals

Almost every musical composition is based on certain *scales*. A scale is an ordered series of *intervals* between *tones* that provide the material for composing *melodies* and *chord* sequences. An interval denotes the distance between two tones. The smallest interval spans a semitone. Common names for musical intervals are listed in Table 5.1. The tones of the basic scale are aligned in ascending order and have been named by the first seven Latin letters («A», «B», «C», «D», «E», «F» and «G»). The intervals between the notes amount to full tones, except for two semi-tone intervals located between the tones B and C, and E and F. The first note of a scale (corresponding to the interval «perfect unison») specifies the *key* of a piece of music.

Tones can be be modified by *accidentals*. A sharp sign «♯» raises a note by a semitone, and a flat sign «♭» lowers a note by the same amount. Thus, in summary, twelve tones form the basis of western music which is called *chromatic scale*. Above the interval of an *octave*, the tones of a scale are repeated in a higher *pitch class*, which is denoted by Arabic numerals in the modern scientific system. Another commonly used notation is the traditional («von Helmholtz») system, which is centered on the great octave denoted by capital letters and the small octave having lower case letters. Lower octaves are labeled by primes before the letter, while higher octaves are marked with primes after the letter. The *standard pitch* is defined by the tone «A0» (a') and is usually tuned to a frequency of 440 Hz. A semitone is further subdivided into 100 cents, which cannot be expressed in

Table 5.1: Musical interval names.

Semitones	Interval name
0	perfect unison (P1)
1	minor second (m2)
2	major second (M2)
3	minor third (m3)
4	major third (M3)
5	perfect fourth (P4)
6	tritone
7	perfect fifth (P5)
8	minor sixth (m6)
9	major sixth (M6)
10	minor seventh (m7)
11	major seventh (M7)
12	perfect octave (P8)

the music score directly, but, nevertheless, this subdivision is important for exact tuning to e.g. standard pitch. An octave interval equals a pitch ratio of two.

The *Musical Instrument Digital Interface* (MIDI) system, which is based on a proposal by Salani and Smith in 1981 [94] and which is standardized since 1993 [44], is designed for the usage in electronic musical instruments and computers. For the description of musical pitch, it assigns numbers to the notes starting with note number 0 for «C-1» or (”C) at 8.1758 Hz up to note 127 for «G9» (or G””””) at 12544 Hz. The MIDI note numbers are given in Table 5.2.

5.1.2 Scale mode

For every key, a multitude of different scales exist that represent musical *modes.* A commonly used modal system is the *church mode* (or *ecclesiastical mode*) system. A rough classification of the different modes is the division by the mode attributes *major* and *minor*. Two scales of the same key but different modes only differ in selected notes. For example, in Figure 5.1, a major scale (*Ionian* mode in the ecclesiastical mode system) in the key C is displayed. In Figure 5.2, the natural minor scale (*Aeolian* mode in the ecclesiastical mode system) of the same key C is pictured.

A direct comparison reveals their difference in the three tones *E*♭, *A*♭, *B*♭. The generalized difference between major and natural minor, expressed in terms of musical intervals, is hence the substitution of the major third, the major sixth and the major seventh by a minor third, a minor sixth and a minor seventh, respectively. For any musical key, the tones to be mapped can be derived from the *circle of fifth*, as depicted in Figure 5.3. A major to natural minor conversion is obtained by a leap of three steps counterclockwise, whereas a minor to major change is accomplished by three steps clockwise.

Table 5.2: Musical tones and their MIDI note numbers.

	MIDI Note Number											
Octave	C	C#/ Db	D	D#/ Eb	E	F	F#/ Gb	G	G#/ Ab	A	A#/ Bb	B
-1	0	1	2	3	4	5	6	7	8	9	10	11
0	12	13	14	15	16	17	18	19	20	21	22	23
1	24	25	26	27	28	29	30	31	32	33	34	35
2	36	37	38	39	40	41	42	43	44	45	46	47
3	48	49	50	51	52	53	54	55	56	57	58	59
4	60	61	62	63	64	65	66	67	68	69	70	71
5	72	73	74	75	76	77	78	79	80	81	82	83
6	84	85	86	87	88	89	90	91	92	93	94	95
7	96	97	98	99	100	101	102	103	104	105	106	107
8	108	109	110	111	112	113	114	115	116	117	118	119
9	120	121	122	123	124	125	126	127				

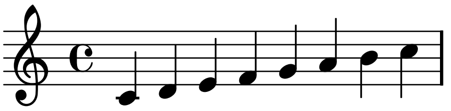


Figure 5.1: C major scale (Ionian mode).

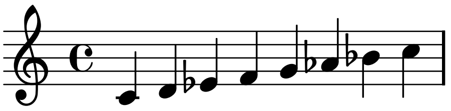


Figure 5.2: C natural minor scale (Aeolian mode).

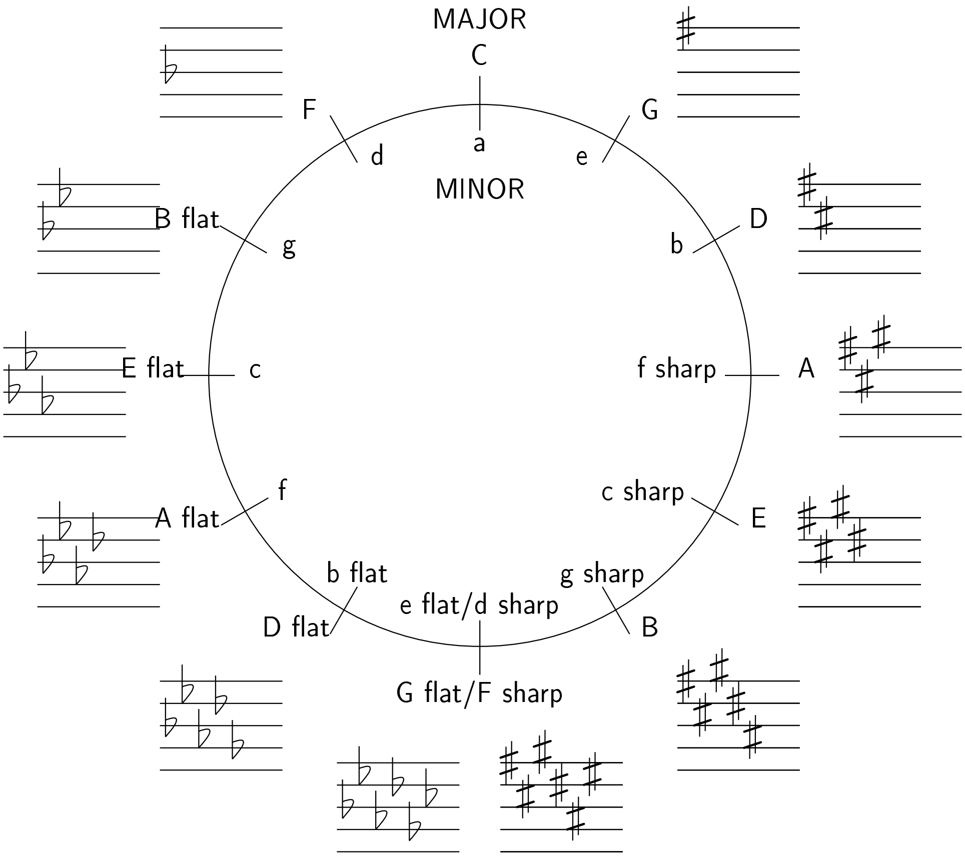


Figure 5.3: Circle of fifth illustrating the relationship between the twelve tones of western music and their corresponding major and minor keys.

5.2 MODVOC operation parameters

5.2.1 Global spectral resolution

For a frequency selective transposition of audio signals by the MODVOC, suitable parameter settings must be chosen. Primarily, the spectral resolution of the underlying DFT transform is set, such that the global frequency resolution for the COG estimation is still appropriate in a «worst case scenario». Since DFT spectra, as utilized in the MODVOC analysis implementation, described in Subsection 4.2.3, have a linear frequency scale, the worst case scenario relates to the low audio frequencies. These low frequencies are required to be resolved within an interval spanned by the thresholds of human spectral pitch discrimination, as introduced in Subsection 2.3.1, and the emergence of an «out-of-tune» sensation which manifests itself above an offset of approximately a quarter-tone. For a 100 Hz tone, the *just noticeable difference* for frequency discrimination (JNDF), as explained in Subsection 2.13, is 1 Hz and «out-of-tune» sensation corresponds to a shift of approximately 1.4 Hz. Hence, considering Equation 5.1, and assuming the sample frequency to be $F_s = 48$ kHz, a DFT length of $2^{15} = 32768$ is a well justified choice, in terms of global frequency resolution [102].

$$\begin{aligned} f_{res} &= \frac{F_s}{N_{DFT}} \\ N_{DFT} &= \frac{F_s}{f_{res}} \\ &= \frac{48000 \text{ Hz}}{1.2 \text{ Hz}} \\ &\approx 2^{15} \end{aligned} \tag{5.1}$$

5.2.2 Perceptual scale

Also, the mapping of the frequency scale towards a perceptual scale prior to COG estimation and thus the definition of the local scope within the COG estimation have to be adjusted. In this work, the ERB scale has been adopted for the mapping, since it provides a finer resolution in the lower bands than e.g. the Bark scale, as discussed in Subsection 2.2.5.

5.2.3 Local subband component resolution

For defining the local scope of the COG estimation, a value of approximately 1/3 ERB has been chosen. Figure 5.4 details the underlying reasoning of this choice based on psychoacoustic and music theoretical considerations. In this graph, the 1/3 ERB curve is depicted (solid black), along with two straight lines that mark the intervals of a semitone, which corresponds to a musical interval of a minor second, and a quarter-tone for a given center frequency (solid and dash-dotted gray, respectively). Additionally, the JNDF is plotted (dashed gray).

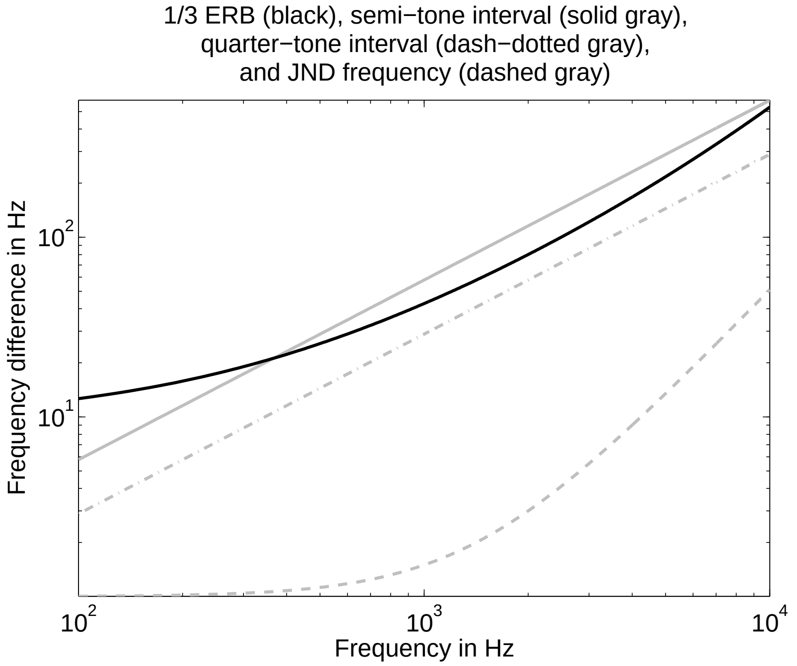
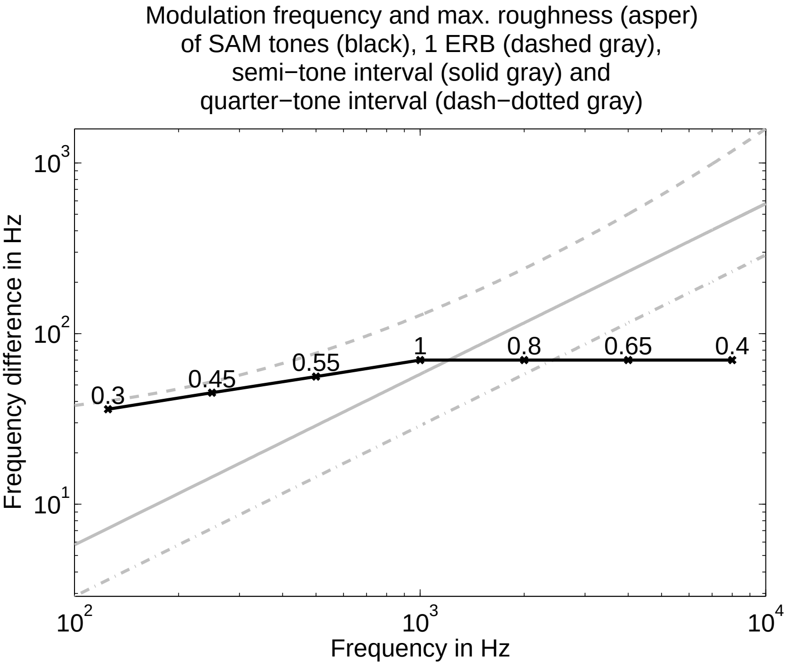


Figure 5.4: 1/3 ERB (solid black) graph and intervals of a semi-tone (solid gray) and a quarter-tone (dash-dotted gray) as a function of frequency. Additionally the JNDF is plotted (dashed gray).

For the task of selective music transposition, the COG local scope is required to separate simultaneous tones which are spaced by a semi-tone interval, since, for western music, this is the smallest interval quantity existent in musical scales. In addition, closely adjacent tones are fused in human perception and are perceived as one tone having an associated envelope fluctuation, as discussed in Section 2.4. Fluctuation frequencies below 20 Hz are perceived as temporal level variations (Subsection 2.4.3), fluctuation frequencies above 20 Hz are perceived as auditory roughness (Subsection 2.4.4). A schematic plot of the phenomenon was given in Figure 2.19.

The highest fluctuation fundamental frequency originated by a band limited signal of bandwidth Δf amounts to $|\Delta f|$ and stems from the beatings inherent in a two-tone complex, as explained in Subsection 2.4.2. Below approx. 300 Hz, the frequency of envelope fluctuation associated with a two-tone complex spaced in semi-tone distance falls below 20 Hz, and is thus perceived as a repeated temporal event. Consequently, the local scope for COG is required to fuse these intervals notwithstanding the violation of the semi-tone interval border. Moreover, in musical compositions, at lower pitches even musical intervals larger than a semi-tone like e.g. three or four semitones (corresponding to musical intervals of minor and major thirds) are usually avoided due to their associated roughness and hence, a fine separation as mandatory at medium pitches is not needed. Above 300 Hz, the 1/3 ERB line is located between quarter-tone and semi-tone interval line. This indicates the feasibility of a separation of intervals equal or greater than a



Modulation frequency and max. roughness (asper)
of SAM tones (black), 1 ERB (dashed gray),
semi−tone interval (solid gray) and
quarter−tone interval (dash−dotted gray)

Figure 5.5: Modulation frequency (black) of sinusoidally modulated tones (SAM) tones and degree of roughness (numbers) for a maximum roughness sensation at a given center frequency. Additionally, the 1 ERB (dashed gray) graph and intervals of a semi-tone (solid gray) and a quarter-tone (dash-dotted gray) as a function of center frequency are pictured. Roughness data according to [125] based on Figure 2.19.

semi-tone distance and the fusion of intervals equal or below a quarter-tone distance.

For a pleasant sounding pitch variation in natural tones (e.g. singing voice, finger vibrato of violin or guitar tones) a mean tonal variation of approx. 50 cents can be assumed [85]. This amounts to a quarter tone interval and is thus also captured well in one component of the MODVOC. Moreover, for tones exhibiting vibrato, the COG represents the effective spectral pitch perceived by listeners, as has been verified for stringed instruments in [12].

Figure 5.5 further illustrates the relation of perceptual scale, musical intervals and auditory roughness sensation. The ERB scale is plotted (dashed gray) together with two straight lines that mark the intervals of a semi-tone and a quarter-tone for a given center frequency (solid and dash-dotted gray, respectively). Additionally, data points indicate the modulation frequency (black) and the degree of roughness (numbers) for a maximum roughness sensation that can be obtained for a given center frequency by excitation with *sinusoidally amplitude modulated* (SAM) tones.

On the double logarithmic plot illustrated in Figure 5.5, it can be seen that for center frequencies below 1 kHz the modulation frequency that evokes maximum roughness increases linearly, starting from 30 Hz, a value close to the ERB interval associated with this center frequency, towards frequencies up to 70 Hz, which, however, correspond to increasingly smaller musical intervals. At 1 kHz, the modulation frequency that maximizes roughness amounts to 70 Hz resembling a semi-tone interval. The maximum roughness level that can be elicited also increases. For 1 kHz, the degree of roughness is globally maximum. Above 1 kHz, the frequency exciting maximum roughness remains static at 70 Hz, while the level decreases rapidly.

In the MODVOC, in the spectral region of fundamental tones ($<$ 1 kHz), spectral segments that contain signal parts corresponding to low modulation frequencies that are perceived as temporal level variations are fused, while segments that elicit roughness are still separated according to western-scale musical requirements. In the overtone spectral regions ($>$ 1 kHz), segments are again fused into a rough sounding tonal compound.

5.3 MODVOC selective pitch transformation

5.3.1 Objective

In Section 4.5, the application of the MODVOC to pitch transposition has been shown. This global transposition changes the original key of a music signal towards a target key (e.g. from C major to G major), while preserving the original tempo. However, due to the signal adaptive nature of the proposed modulation analysis, the MODVOC has the potential to go beyond this task. Even the transposition of selected components of polyphonic music becomes feasible, enabling a novel audio effect which retroactively alters the key mode (e.g. from C major to C minor) of a given music signal [21]. This is possible due to the fact that each component carrier closely corresponds to the perceived pitch in its spectral region. If only carriers that relate to certain original pitches are mapped towards new target values, the overall musical character that is determined by the key mode can be manipulated.

5.3.2 Processing scheme

The necessary processing on the MODVOC components is depicted in Figure 5.6. Within the MODVOC decomposition domain, the carrier frequencies are quantized to MIDI notes which are subsequently mapped onto appropriate corresponding target MIDI notes. For a meaningful reassignment of MIDI notes/pitches, a-priori knowledge of mode and key of the original music item is required. The AM of all components is not acted upon, since these contain no pitch information.

Specifically, the component carrier frequencies f, which represent the component pitch, are converted to MIDI pitch values m, according to Equation 5.2, where f_{std} denotes the standard pitch which corresponds to MIDI pitch 69, the note «A0».

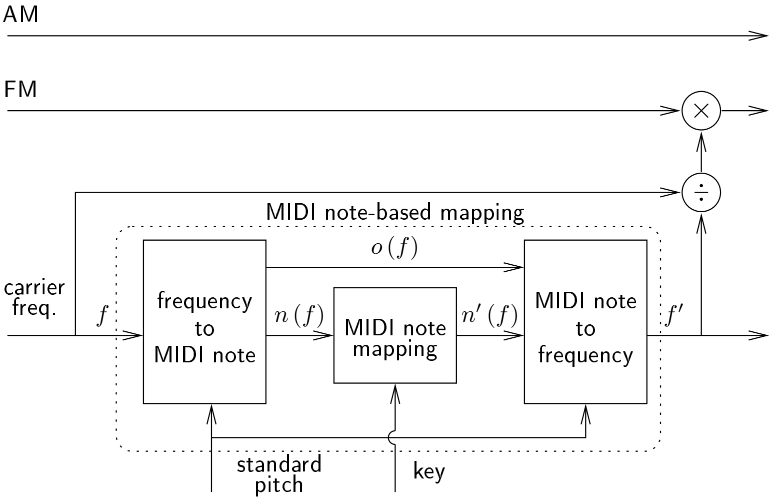


Figure 5.6: Selective transposition on MODVOC components. Carrier frequencies are quantized to MIDI notes which are mapped onto appropriate corresponding MIDI notes. Preservation of relative FM modulation depth is achieved via multiplication of the mapped components by the ratio of original and modified carrier frequency.

$$\begin{aligned} m(f) &= 69 + 12 \cdot \log_2 \frac{|f|}{f_{std}} \\ n(f) &= \operatorname{round}(m(f)) \\ o(f) &= m(f) - n(f) \end{aligned} \tag{5.2}$$

$$\begin{aligned} n &\rightarrow n' \\ f' &= f_{std} \cdot 2^{(n'+o(f)-69)/12} \end{aligned} \tag{5.3}$$

Subsequently, MIDI pitches are quantized to MIDI notes $n(f)$ and, additionally, the pitch offset $o(f)$ of each note is determined. Through the utilization of a MIDI note mapping table which is dependent on key, original mode and target mode, these MIDI notes are transformed to appropriate target values n'. In Table 5.3, an exemplary mapping is given for key of C from major to natural minor. Lastly, the mapped MIDI notes including their pitch offsets are converted back to frequency f' in order to obtain the modified carrier frequencies that are used for synthesis (Equation 5.3). Additionally, in order to preserve the relative FM modulation depth, the FM of a mapped component is multiplied by the individual pitch transposition factor, which is obtained as the ratio of original and modified carrier frequency. A dedicated MIDI note onset/offset detection is not required, since the temporal characteristics are predominantly represented by the unmodified AM and thus are preserved.

Table 5.3: MIDI note mapping table for a scale mode transformation from C major to C natural minor. The mapping applies for the notes of all octaves.

Original note	Target note
C	C
D	D
E	Eb
F	F
G	G
A	Ab
B	Bb

5.3.3 Fundamentals and harmonics

Most instruments excite *harmonic* sounds consisting of a *fundamental* frequency (f0) part and its *harmonics* being approximately integer multiples of the fundamental frequency. As the ratio between harmonics and fundamental can deviate from integer values by an *inharmonicity factor* [37][7], it is sometimes preferred to use the term *overtones.* Since musical intervals obey a logarithmic scale, each harmonic or overtone resembles a different musical interval, with respect to the fundamental (and its octaves). Table 5.4 lists the correspondence of harmonic numbers and musical intervals for the first seven harmonics.

Consequently, in the task of selective transposition of polyphonic music content, an inherent ambiguity with respect to the musical function of a MODVOC component exists [22]. If the component originates from a fundamental it has to be transposed according to the desired scale mapping; if it is dominated by a harmonic to be attributed to another fundamental it has to be transposed together with this fundamental in order to best preserve the original timbre of the tone. This applies especially for single instrument solo parts, since the human auditory system tends to perceive prominent tones as a musical event having a certain timbre which is closely related to the overtone structure. Thus, there emerges the need for an assignment of each MODVOC component, whether it is independent or a harmonic, in order to select the most appropriate transposition. To achieve this, the simple processing scheme, introduced in Subsection 5.3.2, has to be extended by a *harmonic locking* (HL) functionality [22].

5.3.4 Harmonic locking

The harmonic locking examines all MODVOC components prior to transposition whether a component is to be attributed to a fundamental or is to be regarded as an independent entity. This is performed by an iterative algorithm. The flowchart of this algorithm is depicted in Figure 5.7. The algorithm evaluates frequency ratios, energy ratios and envelope cross correlations of a test component with respect to all other components. The succession of test components is determined by their A-weighted energy, such that

Table 5.4: Harmonic numbers and related musical intervals with respect to the fundamental and its octaves for the first ten harmonics.

Harmonic number	Interval name
1, 2, 4, 8	perfect unison (P1)
	minor second (m2)
9	major second (M2)
	minor third (m3)
5, 10	major third (M3)
	perfect fourth (P4)
	tritone
3, 6	perfect fifth (P5)
	minor sixth (m6)
	major sixth (M6)
7	minor seventh (m7)
	major seventh (M7)

the evaluation order is in sequence of decreasing energy. The A-weighting [2][3] is applied to model the perceptual prominence of each component in terms of its loudness [38].

The following features are examined in a comparison to thresholds

- harmonic carrier frequency match
- harmonic carrier frequency missmatch
- component energy
- normalized amplitude envelope correlation at zero-lag

The frequency match and missmatch are defined according to Equations 5.4, with f_t being the test component carrier frequency, and f_i being the component with index i. For the frequency match, all multiples greater than 1 are potential harmonics. A suitable threshold value for the frequency missmatch allowable for a potential harmonic is e.g. 22 Hz.

$$\begin{aligned} match_i &= \text{round}\left(\frac{f_i}{f_t}\right) \\ missmatch_i &= |f_i - (match_i \cdot f_t)| \end{aligned} \tag{5.4}$$

The A-weighted component energy ratio (Equation 5.5) of harmonics versus fundamental is required to be smaller than a predefined threshold, reflecting the fact that for the vast majority of instruments the harmonics exhibit lower energy than the fundamental. A suitable threshold value, for instance, is the ratio of 0.6.

$$nrgRatio_i = \frac{nrg_i}{nrg_t} \tag{5.5}$$

The normalized zero-lag cross correlation of the envelope of the test component env_t, and the envelope env_i of the component with index i, is defined by Equation 5.6. This measure exploits the fact that a fundamental and its harmonics share a rather similar temporal envelope within the block length M. A suitable threshold value was determined to be 0.4 through informal experiments.

$$xcorr_i = \frac{\sum\limits_{m=0}^{M-1} env_i(m) \cdot env_t(m)}{\sqrt{\sum\limits_{m=0}^{M-1} env_i^2(m) \sum\limits_{m=0}^{M-1} env_t^2(m)}} \tag{5.6}$$

After being examined, all components i that meet all of the four threshold conditions are labeled as harmonics to be locked with respect to the test component and are subsequently removed from the search. Next, the test component is also excluded from further iterations by setting its energy to zero. The algorithm is repeated until all components have been assigned, indicated by the maximum component energy being zero.

Figure 5.8 shows the enhanced processing scheme of selective transposition by the MODVOC, incorporating harmonic locking. As opposed to the processing depicted in Figure 5.6, only non-locked components enter the MIDI note-based transposition stage, while locked components are directly modified by the same transposition factor that has been applied to their attributed fundamentals.

5.4 Summary

Beyond the possibilities of existing pitch transposition schemes, the MODVOC is capable of performing a frequency selective pitch transposition. For this task, the MODVOC must be configured with suitable parameters. The global frequency resolution is set such that it complies with the just noticeable difference for frequency discrimination and operates within the limits of «in-tune» sound perception. The spectral local scope of the COG computation is adjusted in accordance with the fusion of spectrally adjacent tones into modulated sounds in the human auditory system, and also with regard to the separation requirement set by an interval of one semi-tone. One semi-tone is the smallest interval quantity commonly used in western musical scales.

The MODVOC, if configured accordingly, can be utilized to transform the original key mode of a pre-recorded piece of music into a different key mode. The original MODVOC components can be individually mapped to appropriate target components using dedicated transposition factors obtained by conversion of their carrier frequencies to *Musical Instrument Digital Interface* (MIDI) notes and a subsequent table based MIDI note mapping. The table content is dependent on the original key, original mode and target mode.

To obtain a musically satisfying result especially for single instruments, such as a solo piano or violin, where overtones are explicitly linked to their fundamental by the

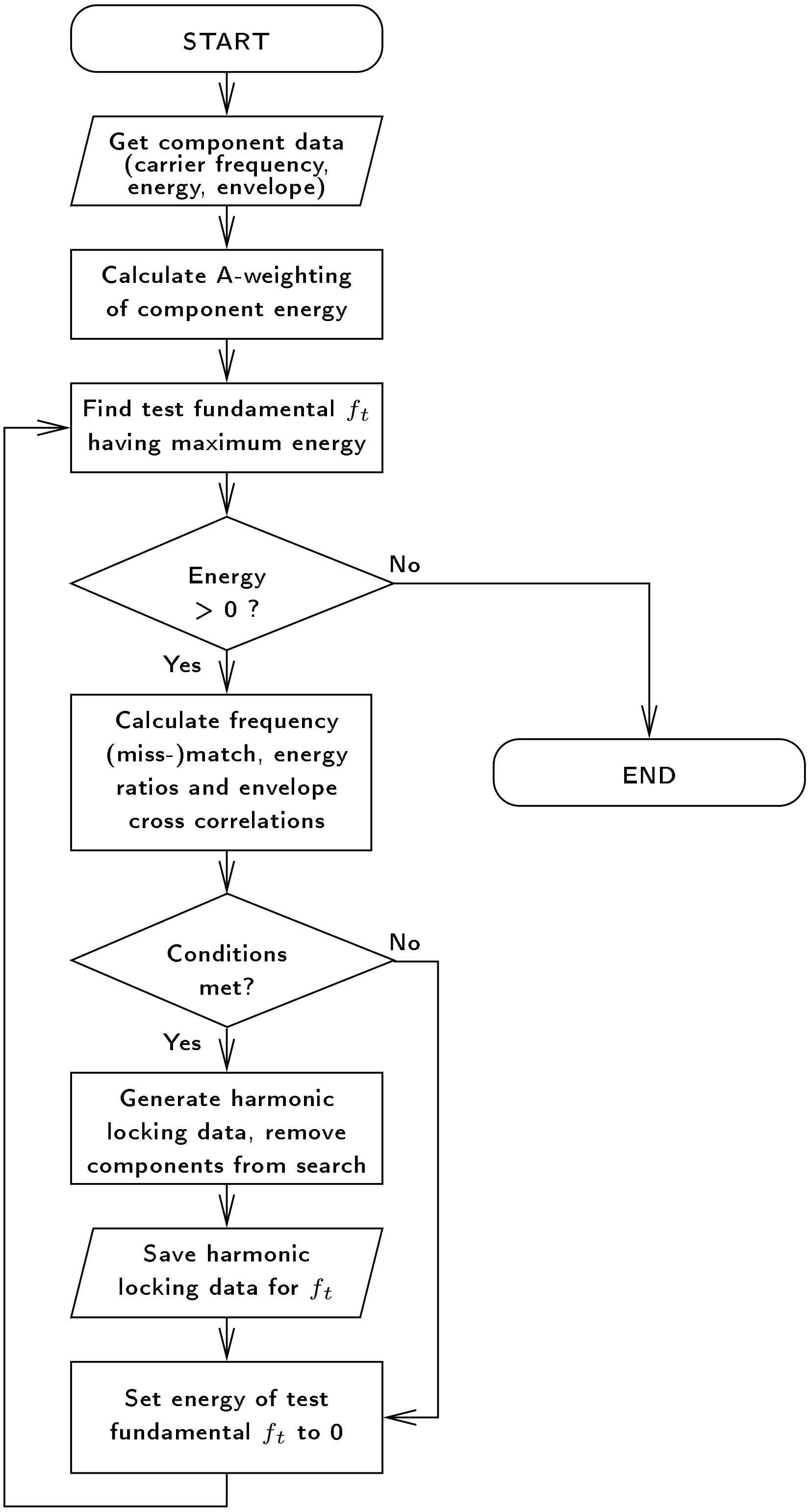

START
Get component data (carrier frequency, energy, envelope)
Calculate A-weighting of component energy
Find test fundamental f_t having maximum energy
Energy > 0 ?
No
Yes
END
Calculate frequency (miss-)match, energy ratios and envelope cross correlations
Conditions met?
No
Yes
Generate harmonic locking data, remove components from search
Save harmonic locking data for f_t
Set energy of test fundamental f_t to 0

Figure 5.7: Flowchart of harmonic locking. Estimated harmonic components with respect to the test fundamental f_t are iteratively labeled by harmonic locking data and are removed from the search space of further iterations.

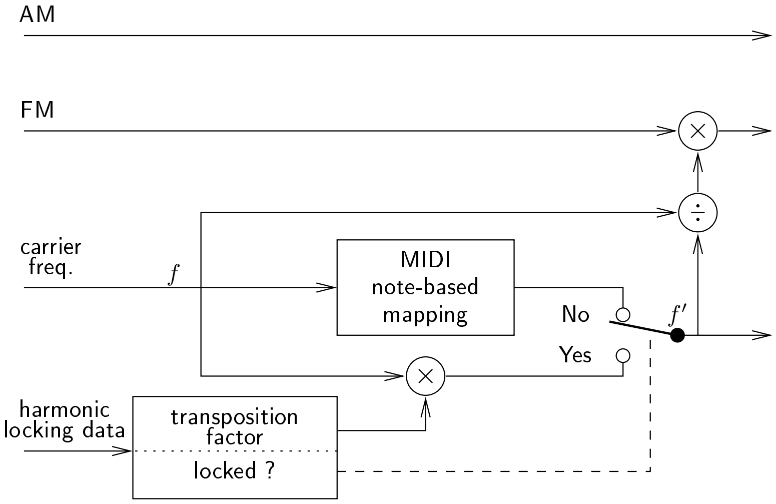


Figure 5.8: Enhanced selective transposition on MODVOC components using harmonic locking. Locked carrier frequencies are transposed via multiplication by the ratio of original and modified carrier frequency of their attributed fundamental. Non-locked carrier frequencies are quantized to MIDI notes which are mapped onto appropriate corresponding MIDI notes.

human auditory system, this thesis further proposes *harmonic locking* (HL), which ties the individual transposition factors of detected overtones to the factor applied to their estimated fundamentals.

Albeit, being a dedicated novel audio effect, this application also demonstrates the meaningfulness of the MODVOC decomposition components and hence the ability of the MODVOC to provide a general basis for new powerful music modification tools.

6 Results

In this chapter, the audio quality that can be obtained by application of the MODVOC and its enhancements to the task of frequency selective pitch transposition of polyphonic audio material is evaluated through listening tests. A test methodology capable to assess the subjective perceptual quality of such extreme manipulations of the original audio stimuli is proposed. Results obtained by said subjective perceptual quality assessment and further results originating from preference tests on selected aspects of perceptual quality are presented. Therefore, test items have been converted between minor and major key mode by the MODVOC and by a commercially available software configured to handle this task. Moreover, test results on perceptual quality of the MODVOC synthesis on unaltered components are presented.

6.1 Frequency selective pitch transposition

6.1.1 Scope

To evaluate the overall subjective audio quality of the MODVOC for frequency selective pitch transposition application and, moreover, the merit of the proposed enhancements to the basic MODVOC principle, a set of exemplary synthetic audio files has been assembled and processed accordingly. Additionally, the MODVOC is compared to a commercial audio software for polyphonic audio manipulation, «Melodyne editor» by «Celemony». Lastly, two main perceptual aspects - *melody and chords transposition* and *timbre preservation* - of the total subjective quality rating are assessed separately in detail, using preference tests for both synthetic and natural sound recordings.

6.1.2 Commercial reference

Melodyne editor became newly available on the market at a point of time close to finalization of this thesis (autumn 2009). Therefore, it was included in the evaluation of the MODVOC perceptual quality, since it is to be regarded as a pioneering and yet commercially unrivaled application with respect to its capability to manipulate polyphonic sound material.

Melodyne editor contains a technology which has been branded and marketed by Celemony by the term *direct note access* (DNA). Unfortunately, there have been no scientific publications by Celemony related to the underlying technology of DNA processing. However, a patent has been filed, presumably covering and thus disclosing the essential functionality of DNA [79]. It can be assumed from this patent application text, that, most of all, the DNA processing included in Melodyne editor relies on heuristic object based classification algorithms. The patent states[1] that firstly «an identification of event objects and note objects» is performed in a Short Time Fourier Transform (STFT) domain environment and a subsequent «linking of event objects to note objects» takes place. Hereby, also the «plausibility of such a mapping» is considered. Next, the different (weighted) spectral proportions to be attributed to each note object are estimated and subtracted from the residual signal by an iterative algorithm. Again, these proportions are subject to a «plausibility check» or, optionally, to a «template matching procedure», in order to «avoid unmotivated jumps in overtone characteristics». To calculate these plausibility checks, the «temporal context spanning several seconds of a note object» must be considered. Note objects usually comprise several time blocks in the STFT domain. Again, the assembly of note objects spanning several time blocks is computed by an iterative algorithm, and is subject to further «plausibility checks». Finally, a post-correction step is mentioned, that relies on the «musical probability» of a detected note event. For example, the sudden occurrence of a single isolated high-pitched note object is very unusual in common musical compositions and can therefore be discarded or fused with another note object.

6.1.3 Synthetic signals

Methodology

Since frequency selective pitch transposition drastically alters the audio content of a signal, a direct comparison of original and processed signal - usually an inherent part in standard listening tests - is apparently not expedient in this case. In order to measure the subjective audio quality in a meaningful way, a special listening test procedure has been applied [22]: the listening test set originates from symbolic MIDI data that is rendered into waveforms using a high quality MIDI expander. This approach enables a direct comparison of similarly altered audio files within the test and allows for the investigation of the effect of the selective pitch processing in isolation. The procedure for generating the test set is summarized in Figure 6.1. The original test signals are prepared in symbolic MIDI data representation (upper left). A second version of these signals is generated by a symbolic MIDI processing, which resembles the target processing under test on the waveform rendered original audio (upper right). Subsequently, these signal pairs are rendered by a high quality MIDI expander into waveform (WAV) files (lower left and right). In the listening test, the waveform rendered from the processed MIDI file and several MODVOC processed versions of the rendered original MIDI file are compared (lower right). Additionally, the output of the MODVOC is compared to the output of Melodyne editor. Melodyne editor initially performs an automatic analysis of the entire audio file. After the initialization phase, Melodyne suggests a decomposition of the audio file. Through user interaction, this decomposition can be further refined. For the sake of a fair comparison to the MODVOC processing results, we chose to base

[1] All citations have been translated into English from the original German text.

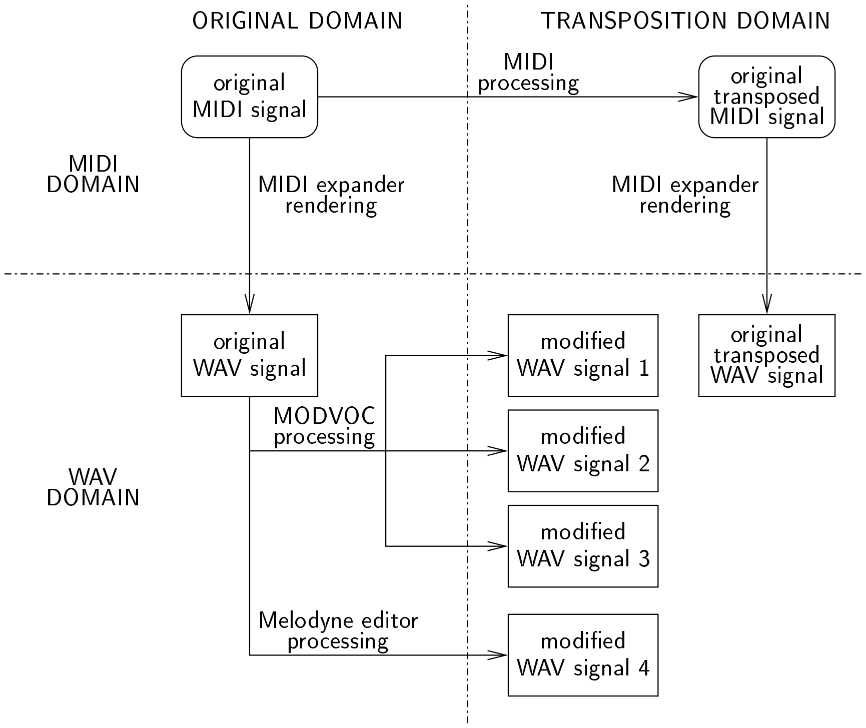


Figure 6.1: Procedure of generating the test set for evaluation of the subjective quality of MODVOC processing for selective pitch transposition.

our evaluation on the outcome of this automatic initial analysis, since, apart from the a-priori knownledge of key and standard pitch, the MODVOC decomposition is fully automatic as well.

The listening test setup was based on a standard *MUltiple Stimuli with Hidden Reference and Anchor* (MUSHRA) test according to the ITU recommendation BS.1534 [52]. First of all, MUSHRA is a blind listening test. For each item, the test presents the labeled reference and all test conditions, along with the hidden reference and a hidden lowpass filtered anchor to the listener in a time-aligned fashion. Hidden reference and lower anchor are included, in order to check the listeners reliability. Only one person at a time is subject to the test. Individual switching between conditions while listening is permitted and so is setting a loop on arbitrarily selected partitions of the item, as suggested in the BS.1116-1 [51] and applicable to MUSHRA tests as well. There is no limit of the number of repetitions the test subjects could listen to before rating the test item and proceeding to the next. This allows for a very close comparison and thorough examination of the different conditions. The perceptual quality of the items is rated on a scale ranging from «excellent» (100 points) via «good», «fair», «bad» and down to «poor» (0 points). The sequence of test items is randomly ordered and the order of the conditions of each item is randomized as well.

Table 6.1: MIDI items for MUSHRA test on synthetic signals.

name	description	instruments	key mode
A	Violin Concerto, J. S. Bach, BWV1041	Orchestra	Amin
B	Eine kleine Nachtmusik, W. A. Mozart, KV525 Mv1	String Quartet	Gmaj
C	Berceuse, G. Fauré, Op56	Flute and Guitar	Emaj
D	Nocturno, F. Strauss, Op7	Horn and Piano	Dbmaj
E	Waltz, F. Carulli, Op241 No1	Guitar	Cmaj
F	Ein Musikalischer Spass, W. A. Mozart, KV522 Mv1	Horns, Violin, Viola, Cello	Fmaj
G	Ode an die Freude, L. V. Beethoven	Piano	Gmaj
H	Piano Trio, L. V. Beethoven, Op11 Mv3	Clarinet, Cello and Piano	Bbmaj

Test items and conditions

The eight test items listed in Table 6.1 have been sourced from the MUTOPIA project[2], which provides free sheet music for public use. Suitable excerpts having an approximate duration of twenty seconds at maximum have been extracted from various pieces of classical music, containing both single instruments (e.g. G, E) and dense full orchestra parts (e.g. F). Moreover, dominant instrumental solo melodies accompanied by other instruments (for example C) are included in the test set. Besides the short-term quasi-stationary tonal parts, percussive elements are also contained in several items (onsets of plucked guitar in C and piano in G), which pose a special challenge on the transient response of the system. The items were further chosen to result in a sufficiently musically pleasant outcome if subjected to the intended key mode change by selective transposition.

The MIDI processing for obtaining the original transposed signals has been done in «Sonar 8» manufactured by Cakewalk. The high quality waveform rendering has been performed using «Bandstand» from Native Instruments in sound library version 1.0.1 R3. The MODVOC processing was evaluated in three different combinations with the two enhancement processing steps being *harmonic locking* (see Subsection 5.3.4) and *envelope shaping*, as introduced in Section 4.6. The MODVOC parameters were set as described in detail in Section 5.2, the envelope shaping was signal dependently switched on and off as a function of the prediction gain. For comparison to Celemony's Melodyne editor, version 1.0.11 was utilized. All conditions are listed in Table 6.2.

[2] http://www.mutopiaproject.org/

Table 6.2: Conditions for MUSHRA test on synthetic signals.

condition	name	description
1	*_reference	MIDI transposed original
2	*_3k5Hz_reference	3.5 kHz lowpass filtered original (anchor)
3	*_MODVOC	MODVOC
4	*_MODVOC_hl	MODVOC with harmonic locking
5	*_MODVOC_hl_es	MODVOC with harmonic locking and envelope sharping
6	*_dna	Melodyne editor (DNA) fully automatic mode

Test setup

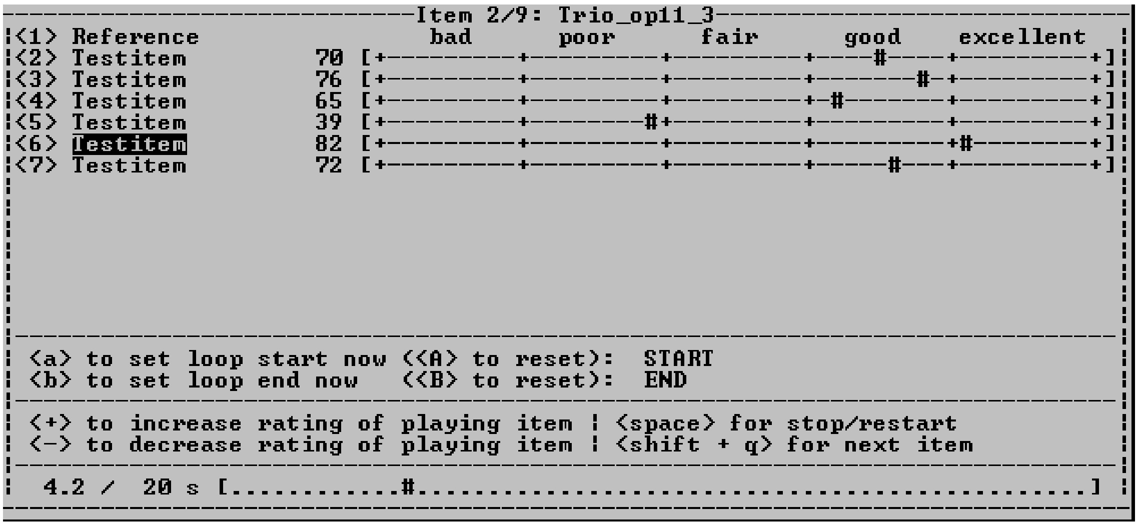
The subjective listening tests were conducted at the Fraunhofer IIS facility in an acoustically isolated listening lab that is designed to permit high-quality listening tests in an environment similar to an «ideal» living room. The listeners were equipped with STAX electrostatic headphones that were driven from an Edirol USB sound interface connected to an Apple MAC mini. The listening test software was «wavswitch» by Fraunhofer IIS, operated in MUSHRA mode. A snapshot of the graphical user interface (GUI) of the test software is depicted in Figure 6.2. The listeners could switch between the reference (1) and the different conditions (2-7) during playout. Each listener could decide individually how long to listen to each item and condition. During the actual switching, the sound playout was muted. The horizontal bars visualize the rating attributed to each condition.

Only experienced listeners that are familiar with audio coding but as well have a musical background were invited for the test in order to obtain an educated judgment on typical processing artifacts (e.g. pre- and post-echoes or dispersion of transients), and on musical parameters (e.g. pitch, melody, chords and timbre).

Absolute scores

In total, fifteen subjects contributed to the test result, where one listener had to be post-screened due to obviously failing to successfully identify the hidden original (by giving it a grade of 64 points). Figure 6.3 summarizes the results of the listening test. The perceptual quality for the items processed by selective pitch transposition ranges from «fair» to «good». The lower anchor was rated between «poor» and «bad» so that the distance from the processed items and the anchor amounts to approx. 40 points.

Absolute scores provide information quantifying the perceptual quality of each item (in each of the test conditions) and thereby implicitly rate the quality difference between



```
----------------------------Item 2/9: Trio_op11_3-----------------------------
¦<1> Reference                  bad       poor      fair      good    excellent  ¦
¦<2> Testitem          70 [+---------+---------+---------+----#----+---------+]¦
¦<3> Testitem          76 [+---------+---------+---------+-------#-+---------+]¦
¦<4> Testitem          65 [+---------+---------+---------+-#-------+---------+]¦
¦<5> Testitem          39 [+---------+--------#+---------+---------+---------+]¦
¦<6> Testitem          82 [+---------+---------+---------+---------+#--------+]¦
¦<7> Testitem          72 [+---------+---------+---------+-----#---+---------+]¦
¦                                                                               ¦
¦                                                                               ¦
¦                                                                               ¦
¦                                                                               ¦
¦                                                                               ¦
¦                                                                               ¦
¦-------------------------------------------------------------------------------¦
¦ <a> to set loop start now (<A> to reset):  START                              ¦
¦ <b> to set loop end now   (<B> to reset):  END                                ¦
¦-------------------------------------------------------------------------------¦
¦ <+> to increase rating of playing item ¦ <space> for stop/restart             ¦
¦ <-> to decrease rating of playing item ¦ <shift + q> for next item            ¦
¦-------------------------------------------------------------------------------¦
¦  4.2 /  20 s [............#............................................]     ¦
--------------------------------------------------------------------------------
```

Figure 6.2: GUI of wavswitch MUSHRA. The test conditions are organized in lines, the rating is visualized by stylized horizontal bars.

the items in the testset, but are unsuitable to compare the different conditions within the listening test, since the ratings of these conditions are not independent [103][77]. As suggested in [77], for a direct comparison of the conditions originating from the different selective transposition processing schemes, score differences are considered in the following.

Difference scores for MODVOC enhancements

Figure 6.4 depicts the outcome based on score differences of the enhanced MODVOC variants (conditions 4 and 5), with respect to the plain MODVOC (condition 3) results. Here, all enhanced MODVOC variants score considerably better than the plain MODVOC processing (all mean scores are well located above zero). There is significance in the 95% confidence sense for all items and conditions, except for the application of harmonic locking only in item A and C.

Difference scores for comparison MODVOC vs. Melodyne editor

Figure 6.5 displays the test scores as score differences with respect to condition 6 (Melodyne editor). For item C, the MODVOC in condition 5 scores significantly better than Melodyne editor, while condition 4, albeit being slightly positive, and condition 3 are inconclusive in a 95% confidence interval sense (confidence intervals overlap with 0). Also, no significant conclusion can be drawn for items B (condition 2), F, and G (condition 5), but rather a tendency for better performance of the MODVOC can also be seen for item C in condition 4 and item F in conditions 4 and 5. In all other cases, the MODVOC scores significantly worse than Melodyne editor.

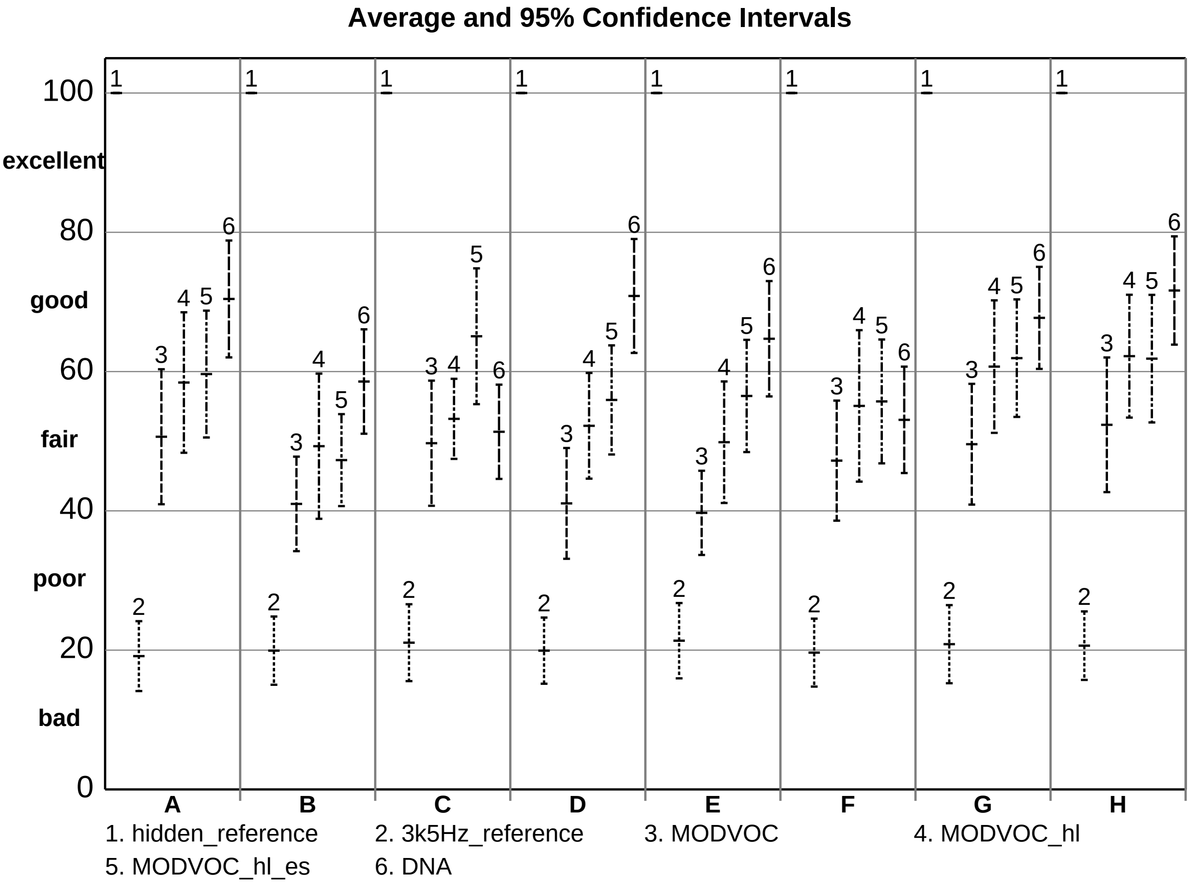


Figure 6.3: Absolute MUSHRA scores and 95% confidence intervals of listening test addressing selective pitch transposition. Condition 3 was processed by the MODVOC, condition 4 with additional harmonic locking (MODVOC_hl), condition 5 with additional harmonic locking and envelope shaping (MODVOC_hl_es) and condition 6 by manipulation using Melodyne editor (DNA).

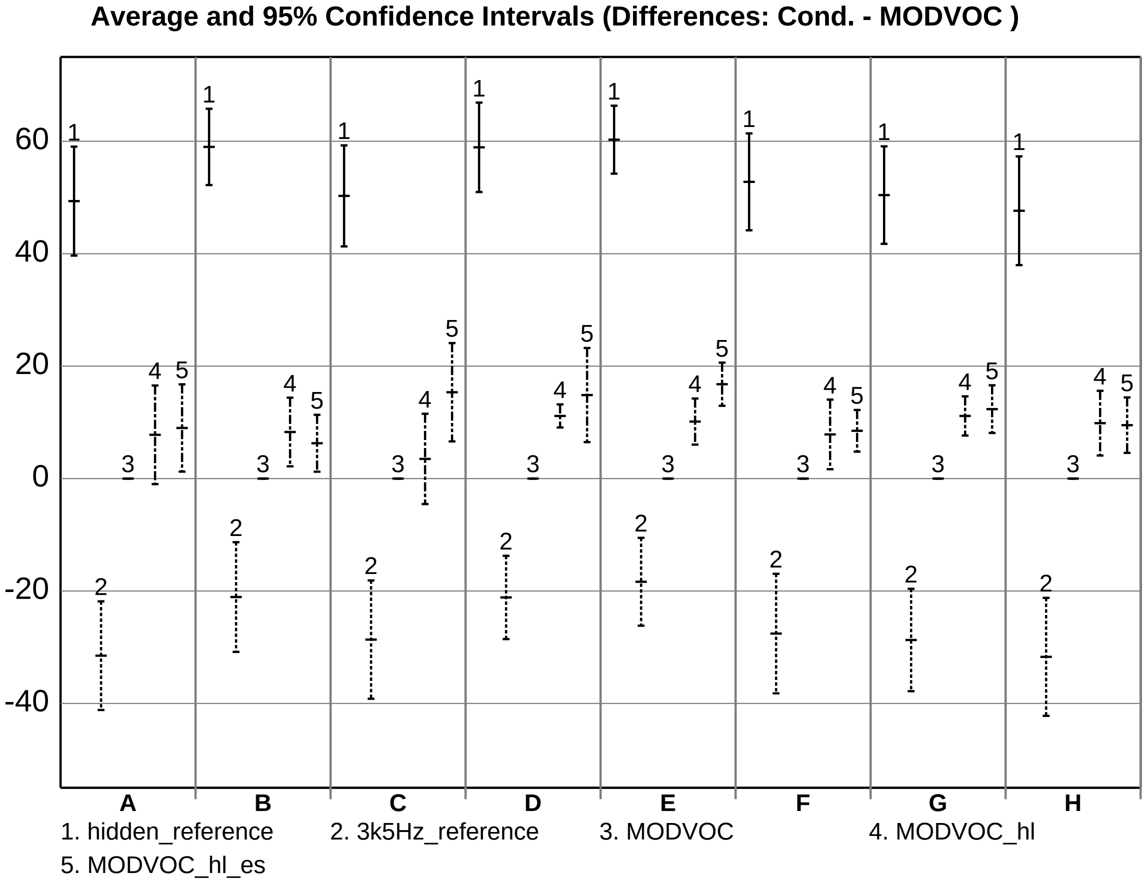


Figure 6.4: Difference MUSHRA scores with respect to condition 3 (MODVOC) and 95% confidence intervals of listening test addressing selective pitch transposition.

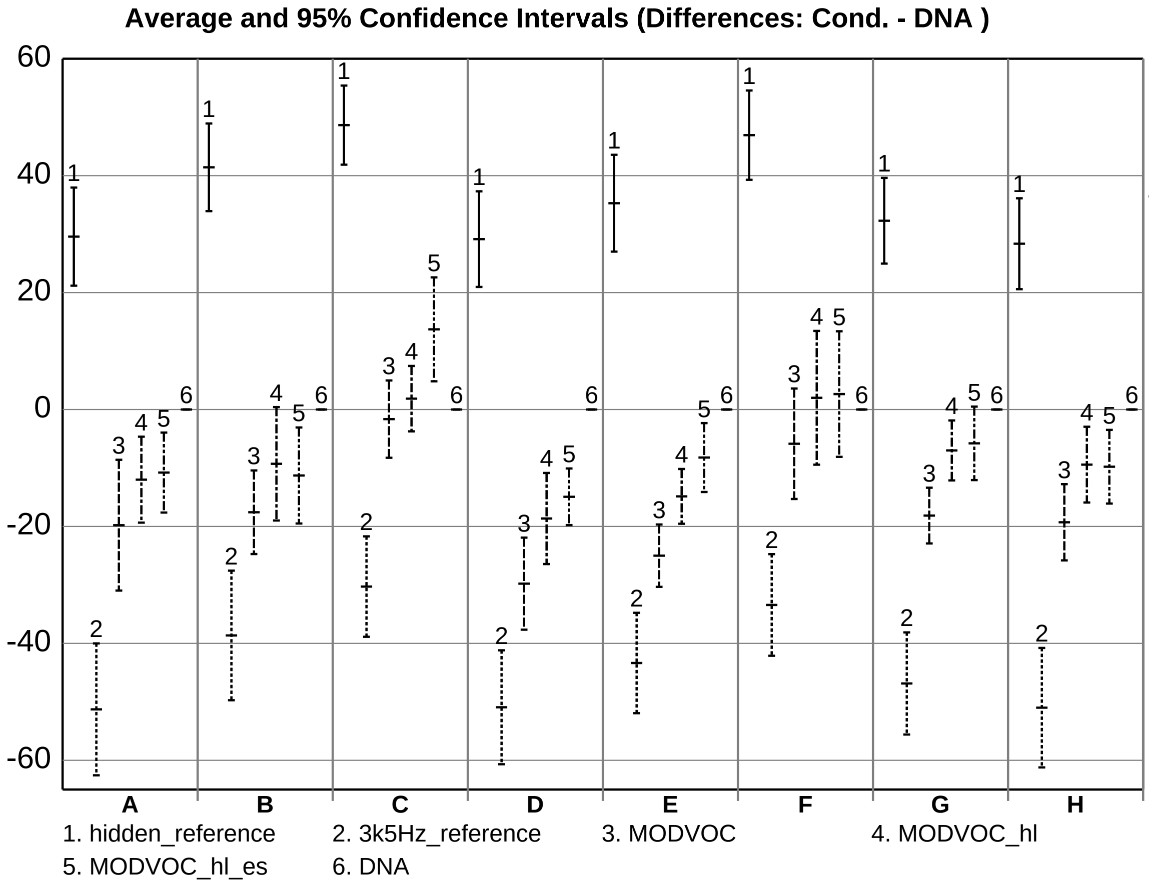


Figure 6.5: Difference MUSHRA scores with respect to condition 6 (DNA) and 95% confidence intervals of listening test addressing selective pitch transposition.

Discussion

The score reflects overall quality judgment comprising aspects, like unnaturally sounding artifacts, such as the degradation of transients by pre- or post-echoes, pitch accuracy, correctness of melody and chords, and the preservation of timbre. In order to interpret the results in more detail, the listeners were asked to note their informal observations alongside with noting the actual score. From these observations, it can be concluded that the preservation of the timbre and absence of unnatural sounding artifacts contributed to the overall score to a higher degree than the performance in terms of melody and chord transposition. If a certain melody and chord progression was previously unknown to the listeners, it seemed that the test persons were not able to memorize the reference melody and chord progression on short notice during the test and thus were unsure about the true melody and underlying chords. This can be an explanation of the higher overall rating of the Melodyne editor processed items, since these usually have a higher fidelity with respect to preservation of timbre (see Subsection 6.1.4), especially of sounds originating from single instruments that exhibit many strong overtones. However, this comes at the price of severe accidentally occurring melody and chord errors that happen presumably due to misclassification. In contrast, the MODVOC is less prone to the occurrence of such errors, since it does not predominantly rely on feature based classification techniques.

6.1.4 Perceptual quality aspects

Test items and conditions

To investigate how specific perceptual quality aspects may influence the overall rating of listeners, two main aspects denoted by «melody and chords transposition» and «timbre preservation» were evaluated separately in a preference test [23]. This test considers items which have already been transposed in the MIDI domain and subsequently rendered into waveforms. These waveformes were finally transposed back into their original key mode by both MODVOC and Melodyne editor. This procedure has been chosen to exclude any effect of listener failure to memorize melodies and chord progressions which otherwise would have sounded unfamiliar due to their changed key mode (see Subsection 6.1.3). The items of this test are listed in Table 6.3. Table 6.4 specifies the two alternative processing methods (conditions) that were tested.

Test setup

For conducting the preference test, a dedicated GUI has been prepared, which is depicted in Figure 6.6. In the left box, the original items were also available for supporting the actual decision process. Since synthetic items were tested, these non-transposed originals could be used as a ground truth with respect to both aspects, melody and chords transposition and timbre preservation. The randomized test conditions are located in the middle box. The listeners were asked to indicate their exclusive preference for one of these conditions, A or B, by setting tickmarks in the appropriate boxes on the right. Listeners were allowed to listen as often as desired to each of the test items. To force

Table 6.3: MIDI items for preference test on synthetic signals.

name	description	instruments	key mode
1	Violin Concerto, J. S. Bach, BWV1041	Orchestra	Amin
2	Berceuse, G. Fauré, Op56	Flute and Guitar	Emaj
3	Ode an die Freude, L. V. Beethoven	Piano	Gmaj
4	Concerto for Violin and Orchestra L. V. Beethoven, Op61	Violin and Orchestra	Dmaj

Table 6.4: Preference test conditions.

condition	description
MODVOC	MODVOC with harmonic locking and envelope sharping
DNA	Melodyne editor (DNA) fully automatic mode

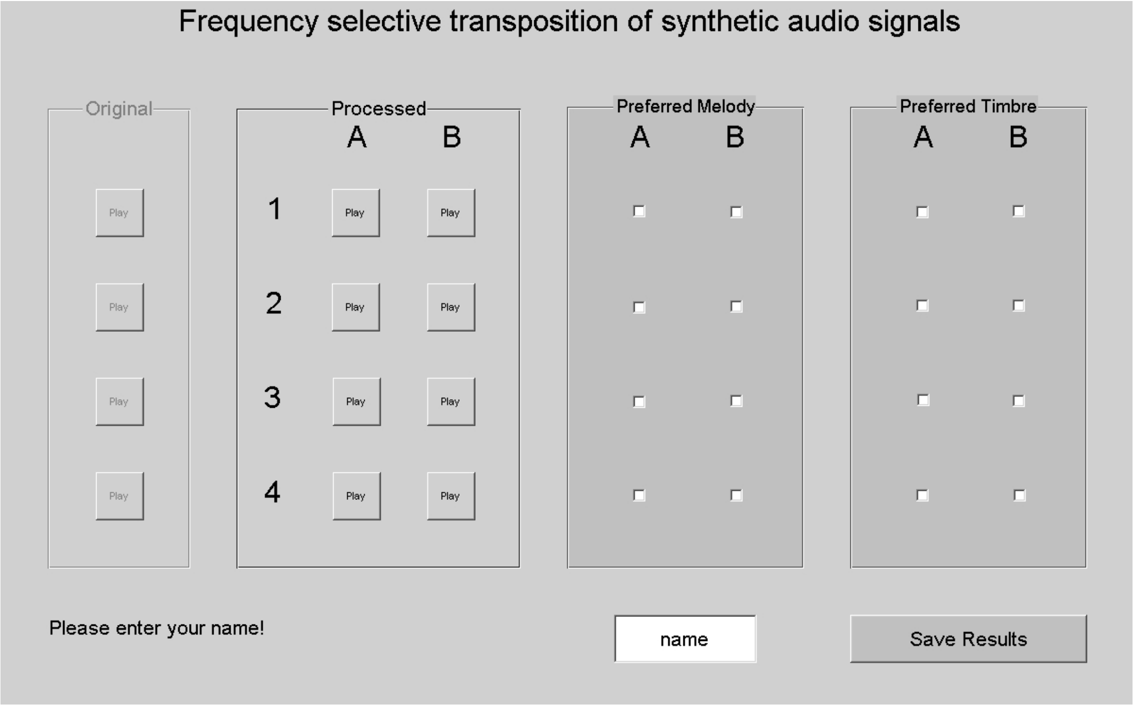
listeners into listening to the entire item before making their choice, no functionality for setting loops or interrupting playout was provided and listening was restriced to full playout of the entire length of the items. The listeners were instructed to focus on the following phenomena and associated questions

- quality of melody and chords transposition; do the melody and chords consistently sound as if originally played in that key throughout the item or are there “wrong notes” or bad intonations audible?
- believability of timbre; is the timbre consistent throughout the item and is it plausible for every instrument contained the mix?

Results

Twelve subjects participated in this test. All subjects were expert listeners and, at the same time, musicians capable of playing at least one instrument. Playout was from a notebook via Edirol UA-25 USB audio interface into Beyerdynamic DT-770 closed headphones. Figure 6.7 and Figure 6.8 illustrate the listeners preference choice with respect to melody and chords transposition, and timbre preservation, respectively.

For three out of four items, the MODVOC was preferred by the majority of listeners over Melodyne editor in terms of melody and chords transposition, for one item the preferences were on par. In contrast, in three out of four items Melodyne editor was preferred with regard to timbre preservation and for one item (solo piano), the MODVOC was clearly preferred.



Frequency selective transposition of synthetic audio signals

Please enter your name!

Figure 6.6: GUI of preference test. The original is provided as informal reference (left box). The test conditions (A-B) and items (1-4) are organized in a matrix (middle box). The preference choice is indicated by setting tickmarks (right boxes).

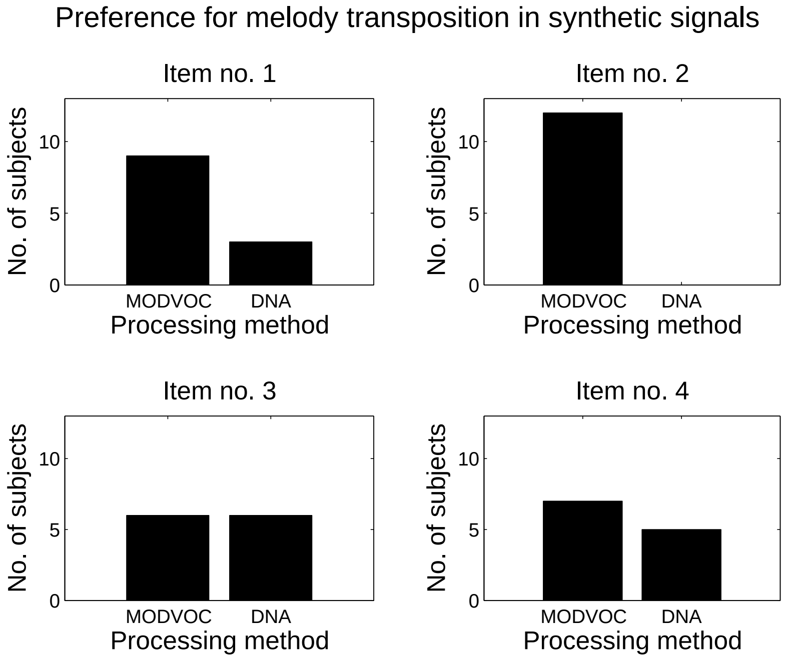


Figure 6.7: Preferences with respect to melody and chords transposition for synthetic signals.

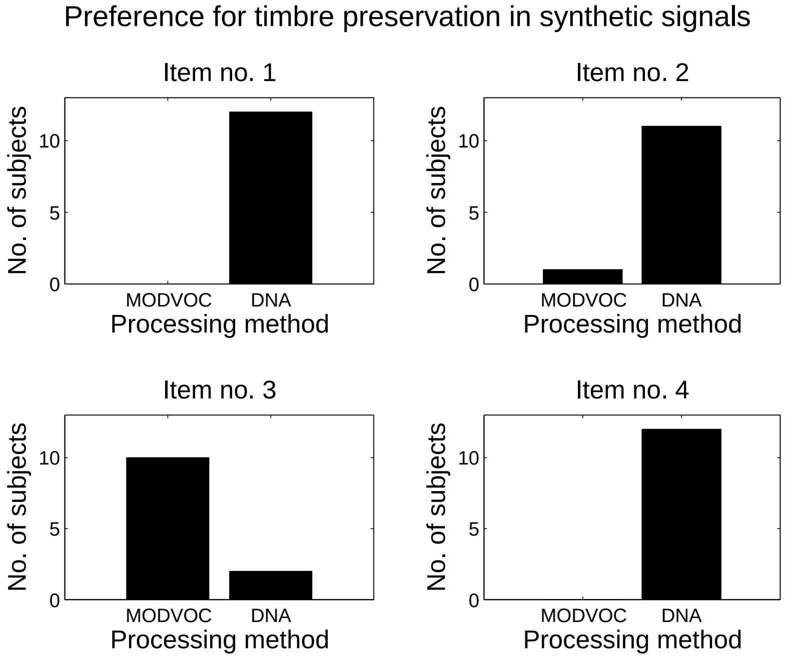


Figure 6.8: Preferences with respect to timbre preservation for synthetic signals.

Table 6.5: Items for preference test on natural signals.

name	description	instruments	key mode
1	Violin Concerto, J. S. Bach, BWV1041	Harpsichord Orchestra	Amin
2	Quintetto 2 (maestoso assai) L. Boccherini	Guitar Orchestra	Emaj
3	Etude, F. Chopin, Op10, No3	Piano	Emaj
4	Four seasons, spring allegro A. Vivaldi	Orchestra	Emaj

Discussion

The outcome of the preference test strongly supports the viewpoint already stated in the discussion of the MUSHRA test, based on the informal comments of the listeners (see Subsection 6.1.3). Since for item no.3 the MODVOC processing was preferred in terms of both melody/chords and timbre, it can be concluded that the harmonic locking performs well for sparse polyphonic mixtures of harmonic instruments, such as the piano.

6.1.5 Natural recording signals

Test items, conditions and setup

Additionally, the perceptual quality of the MODVOC and Melodyne editor for the frequency selective pitch transposition application has been investigated for natural audio recordings. The aim of this test is to compare the stability and performance of both algorithms when applied to real-world signals. Such signals may exhibit natural inaccuracies in tuning or intonation due to the «human factor», and disturbing components originating from room reverb, ambience or tape hiss.

Since no direct ground truth is available in this test, a preference test, as described in Subsection 6.1.4, was conducted and the listeners were instructed to consider the originals as an informal reference only for the timbre preservation property. A similar test setup as outlined in detail in Subsection 6.1.4 was employed. Table 6.5 displays the set of test items.

Results

The same twelve persons that participated in the previously described synthetic items test also attended the natural items test. The results are depicted in Figures 6.9 and 6.10. In terms of melody and chords transposition, three of the MODVOC processed items are preferred, while item no.2 (Guitar/Orchestra) is equally preferred in both processing versions. The timbre preservation properties are consistently preferred in the Melodyne editor processed items.

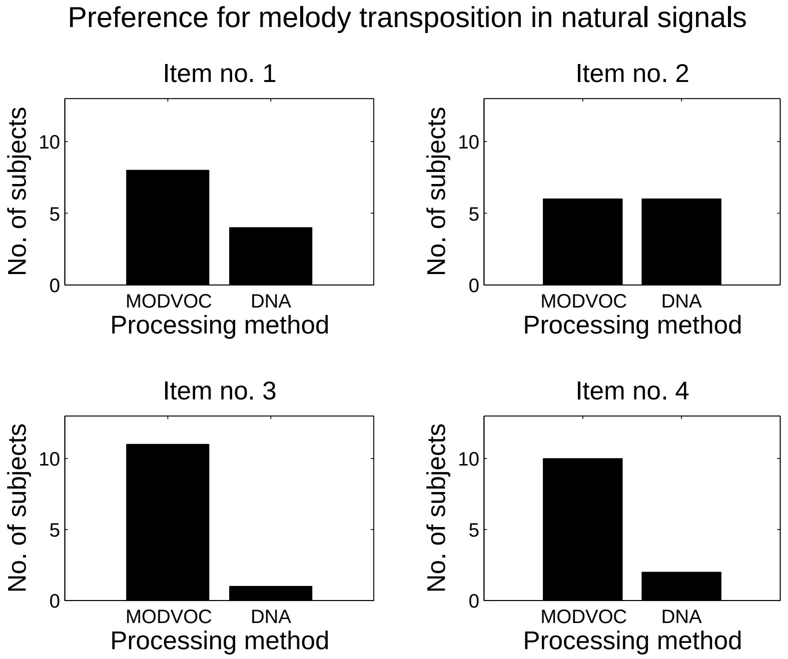


Figure 6.9: Preferences with respect to melody and chords transposition for natural signals.

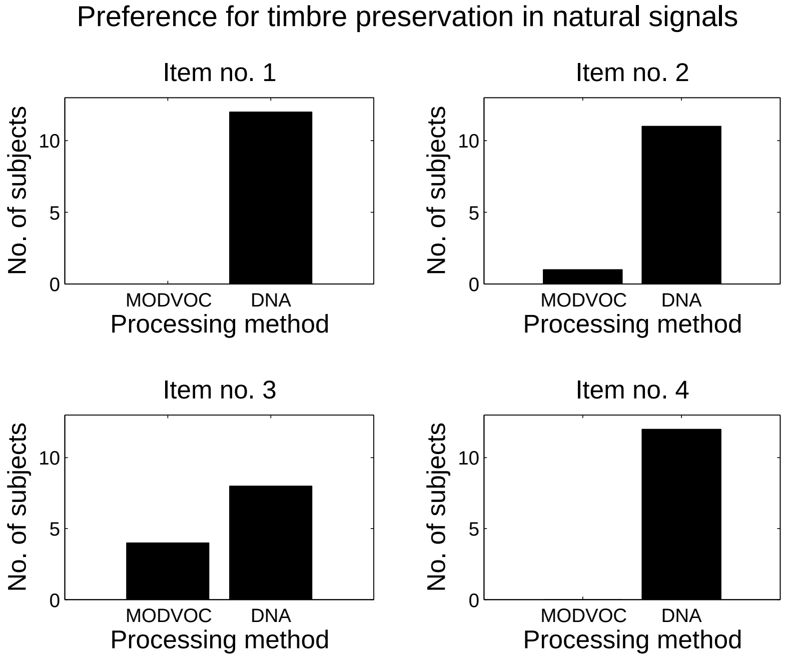


Figure 6.10: Preferences with respect to timbre preservation for natural signals.

Table 6.6: Test set of critical items for MUSHRA transparency test on natural signals.

name	description
si01	harpsichord
si02	castanets
si03	pitch pipe
sm01	bag pipe
sm02	glockenspiel
sm03	plucked strings

6.2 Transparency of MODVOC analysis and synthesis

6.2.1 Scope

The *perceptual transparency* (indistinguishability from the original) of the MODVOC output in case of unaltered components has been assessed using a standard MUSHRA test, in order to be able to judge the basic quality of the analysis-synthesis processing chain. This gives an indication on the maximum subjective quality that can be obtained by any intermediate MODVOC processing on the modulation components. Moreover, a condition comprising the application of envelope shaping has been included in order to additionally evaluate the benefit of this technique separate from a selective transposition application.

6.2.2 Test items and conditions

The six test items have been selected from a well known set originating from MPEG perceptual audio coding standardization. The set consists of the most critical music material and is thus well suited for the evaluation of perceptual transparency. Table 6.6 lists all items of the set.

Apart from the original and lowpass filtered anchor, a MODVOC processed version of the stimuli and a MODVOC processed version applying *envelope shaping* (Section 4.6) has been included in the test. The MODVOC parameters were set as described in detail in Section 5.2. Naturally, no intermediate modulation processing has been included between analysis and synthesis. All test conditions are listed in Table 6.7.

6.2.3 Absolute scores

In total, twelve listeners contributed to the test and eleven subjects produced valid results (one subject graded two hidden references with 87 and 75 points, respectively). In Figure 6.11, the absolute MUSHRA scores for the test items are depicted. For the test set consisting of most critical items, the gradings on five out of six items ranged from a perceptual quality labled «good» to «excellent». One item, however, containing a solo castanet («si02») only scored «fair», which indicates that pure transient signals

Table 6.7: Test conditions for MUSHRA transparency test on natural signals.

condition	name	description
1	*_reference	original
2	*_3k5Hz_reference	3.5 kHz lowpass filtered original (anchor)
3	*_MODVOC	MODVOC
4	*_MODVOC_es	MODVOC with envelope sharping

do not fit the MODVOC signal model. Nevertheless, predominantly tonal signals even if they contain transient onsets are handled very well by the MODVOC processing.

6.2.4 Difference scores

To assess the effect of an added envelope shaping functionality, difference scores with respect to plain MODVOC processing have been evaluated and are depicted in Figure 6.12. Five out of a total of six items indicate an improvement, while items «si02» «sm02» and «sm03» are significantly better in the 95% sense. Item «sm01», which was rated «excellent» in absolute scores (see Subsection 6.2.3), shows no change in subjective quality. The highest improvement, amounting to more than 10 points, is obtained for «si02» which, in absolute scores, has a «fair» quality (see Subsection 6.2.3).

6.2.5 Discussion

The above results show that a «good» to «excellent» perceptual quality can be expected from MODVOC processed critical items containing mostly tonal audio material, while a purely transient item was only rated «fair». Nonetheless, the usage of envelope shaping proved to be beneficial, especially for this transient item, and also for other items containing pronounced onsets of tones (e.g. plucking of strings, hits on a glockenspiel plates). This further substantiates the statement of Section 4.6 that the envelope shaping improves exactly on the limits of the MODVOC signal model, specifically the loss of global spectral coherence.

6.3 Summary

The listening test results obtained for frequency selective transposition of pitch of synthetic audio signals lead to the conclusion that the plain MODVOC is indeed enhanced by e*nvelope shaping* (ES) and *harmonic locking* (HL). Further, a comparison of the MODVOC with frequency selective transposition results achieved by a commercially newly available software (Melodyne editor) revealed that for the majority of items the processing by Melodyne editor is rated with higher scores than the MODVOC processing. This is to be attributed to a better preservation of timbre by Melodyne editor.

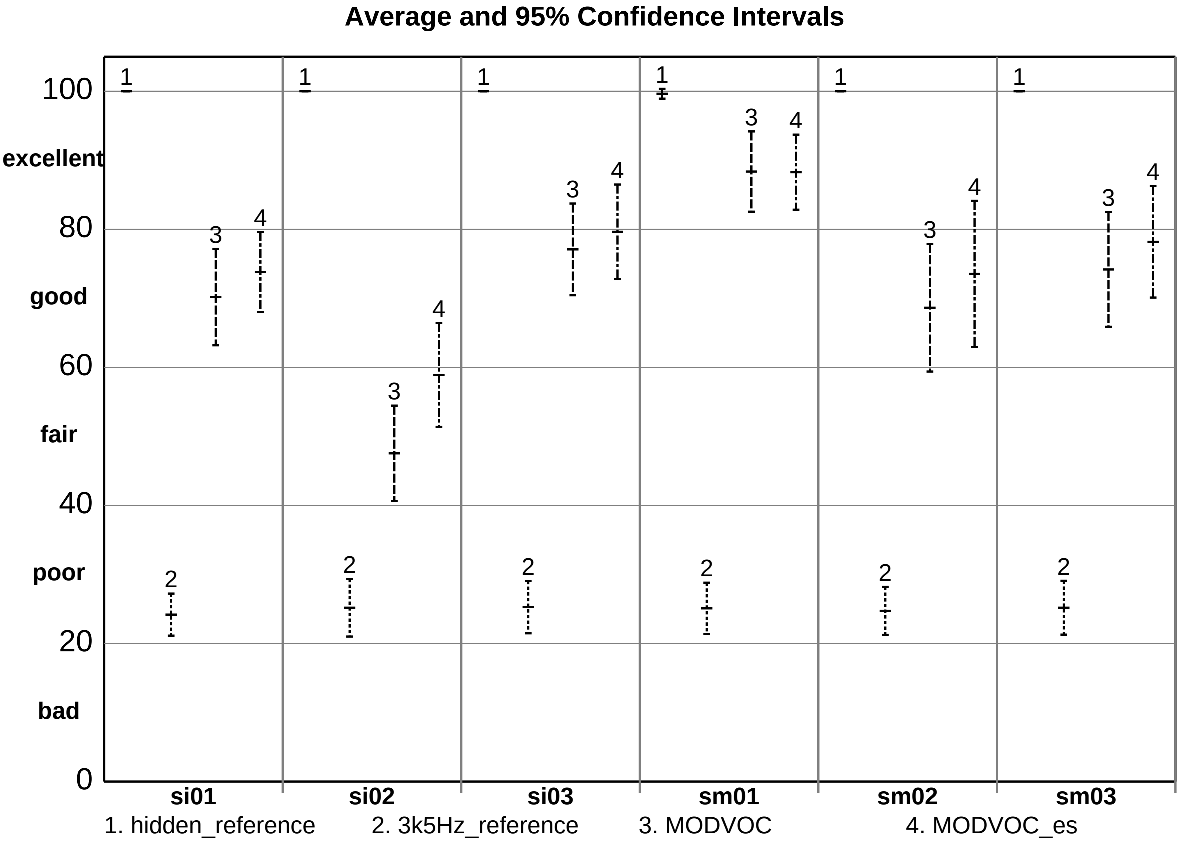


Figure 6.11: Absolute MUSHRA scores and 95% confidence intervals of listening test addressing transparency.

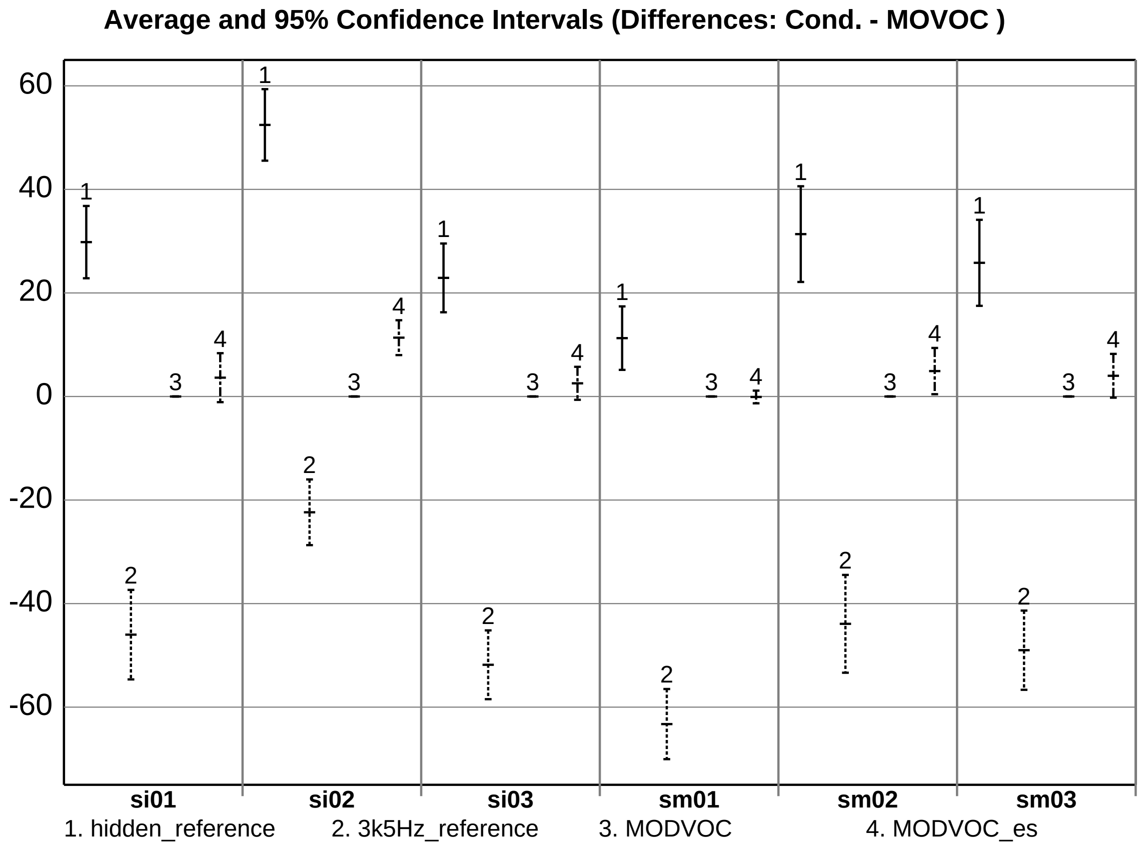


Figure 6.12: Difference MUSHRA scores with respect to condition 3 (MODVOC) and 95% confidence intervals of listening test addressing transparency.

Nonetheless, the MODVOC proved to be more robust to missinterpretation of melody and chord progressions since unlike the commercial system it essentially does not rely on classification decisions.

Additional preference tests on the detailed quality aspects, «melody and chords transposition», and «timbre preservation» confirmed that for the majority of test items the MODVOC was preferred in terms of melody and chords transposition whereas Melodyne editor was chosen most often as the preference in terms of timbre preservation.

It can be further speculated that the quality of melody and chords transposition, and the accuracy of timbre preservation possibly constitute opposing quality tradeoff aspects of any selective transposition scheme, at least for audio material containing complex polyphonic mixtures.

Melodyne editor essentially performs a multi-pass analysis on the entire audio file, whereas the MODVOC is based on a single-pass blockwise processing. Thus, in contrast to Melodyne editor, the application of the MODVOC allows for a streaming or realtime operation scenario. The restrictions posed on the MODOVC hereby can be seen as another reason for the quality difference of both methods.

The MODVOC transparency test results, assessing an analysis and synthesis scenario without intermediate processing, showed that the MODVOC can provide near-transparent «good» to «excellent» perceptual quality. This applies for predominantly tonal items, and especially, if the MODVOC is enhanced by ES.

7 Conclusions

There is an increasing demand for digital signal processing techniques which enable extreme retroactive signal manipulations. Such manipulations are a prerequisite to fit already existing audio recordings, e.g. so-called «samples» taken from a database, into a new musical context of other pre-recorded audio material. Therefore, original high level semantic signal properties like tempo, pitch, musical key and scale mode need to be adapted to different target values. This requires signal processing methods that are broadly applicable to different classes of signals, including polyphonic mixed music content. In the past, the selective transposition of pitch, being the underlying processing for musical key and scale mode change, had been restricted to monophonic signal content. Thus, in this thesis, for the first time, a system is proposed, capable of selective pitch transposition of polyphonic signals. In contrast to other approaches, this method relies on psychoacoustic findings, with regard to modulation perception rather than source separation techniques motivated by auditory scene analysis.

More precisely, the general approach of this thesis to audio signal manipulation is based on a signal adaptive decomposition of these signals into perceptually adapted subband components and associated parameters motivated by psychophysical findings. These parameters being subband carrier, *amplitude modulation* (AM) and *frequency modulation* (FM) can thus be directly interpreted in a perceptual sense. The carrier frequency describes the pitch sensation elicited by the signal contained in the component. Amplitude modulations by frequencies lower than approximately 20 Hz capture the temporal level variation of the carrier perceived as *tremolo*, while frequency modulations by low frequencies represent pitch variations of the carrier perceived as *vibrato*. Both amplitude and frequency modulation by higher frequencies introduce auditory roughness into the signal.

The newly proposed method is termed *modulation vocoder* (MODVOC) and it is applied for analysis, manipulation and synthesis of audio signals. The fundamental idea of this approach is to decompose polyphonic mixtures into subband components that are perceived by humans as sonic entities. All signal elements contained in each component can then be jointly manipulated in subsequent processing stages. Therefore, the signal is analyzed in successive temporal blocks. In each time block, subbands having bandpass characteristics are signal adaptively aligned with spectral local *centers of gravity* (COG). In their bandwidth, these subband filters follow the perceptual *equivalent rectangular bandwidth* (ERB) scale. Finally, a Hilbert based AM/FM decomposition is applied on each subband component. To ensure global spectral coherence for high quality transient reproduction, it is proposed to apply *envelope shaping* (ES) by linear prediction in the spectral domain. Additionally, a synthesis method is introduced that renders a smooth and perceptually pleasant, yet - depending on the type of manipulation applied - drastically modified output signal from the AM/FM decomposed subband signals. In detail, the synthesis method performs a parameter matching on successive time blocks, based on the spectral distance of the carrier frequencies and interpolates between matched pairs of carrier frequencies and their associated AM and FM contours applying overlap-add. The output signal is obtained as the sum of all components.

An application of the MODVOC to frequency selective pitch transposition for polyphonic music signals is proposed, and a suitable parametrization of the MODVOC is provided, based on psychoacoustic and musical criteria for this dedicated application. It is pointed out that a musically meaningful selective transposition operation is inevitably tied to the ambiguity, whether a MODVOC component is dominated by a fundamental tone or an overtone of a certain musical instrument contained in the signal mix. In the first case, the component can be manipulated appropriately on its own, while in the latter case, the component must be manipulated according to its estimated fundamental. Therefore, an additional technique is introduced, called *harmonic locking* (HL), which locks components dominated by overtones to their estimated respective fundamental component by applying a common transposition factor. HL can improve the subjective quality of the frequency selective transposition result. It is demonstrated that a selective pitch change opens up possibilities for advanced audio effects, such as musical key and scale mode change of readily recorded and mixed audio tracks.

To assess the perceptual quality of the MODVOC application to selective pitch transposition, a dedicated listening test methodology is introduced, which is adapted to accommodate rather incisive changes of the original audio items. The test items have been rendered from MIDI files and were subsequently subjected to selective pitch transposition by the MODVOC in several configurations. These are compared to a target *ground truth*. A sufficiently well-defined ground truth can be obtained from the original MIDI file by performing an equivalent symbolic MIDI based manipulation prior to rendering. Following this methodology, a MUSHRA based listening test evaluating the quality of items which have been altered in their key mode by MODVOC processing has been carried out.

During the term of the thesis work, a first commercial software named «Melodyne editor» by «Celemony» had become available in late 2009, which also supports a selective transposition of pitch for arbitrary signals. However, due to the strictly commercial background of this software, no scientific papers have been published related to its functional principle. For a comparison to the MODVOC, a similar manipulation has been performed within the Melodyne editor program environment on the test set and included in the test.

In listening tests, the subjective perceptual quality of the test items is usually rated by a single value, albeit many quality aspects contribute to the overall listening impression. For instance, in the case of selective transposition of pitch, the aspects «artifact insertion», «pitch stability», «melody and chords transposition» or «timbre preservation» are implicitly summed up with individually unknown weights to arrive at a global rating number.

The listening test clearly proved the benefits of HL for MODVOC processing in terms of an improved timbre. Moreover, listeners informally reported a musically convincing transposition result in terms of melody and chords transposition by the MODVOC processing. In contrast, since Melodyne editor supposedly applies many heuristic classifications that can fail at some selected spots of the test signal, it was prone to severe melody and chord errors, albeit at a constantly good preservation of timbre. However, from the listeners informal comments, it was concluded that the preservation of timbre had a greater influence on their overall rating than the correct interpretation of melody and chords. In their overall rating, the listeners therefore mostly preferred the processing by Melodyne editor over the MODVOC. To further investigate this assumption, preference tests assessing the two main perceptual quality aspects «melody and chords transposition» and «timbre preservation» have been conducted. The outcome of these tests confirmed that the application of MODOVOC processing indeed yields a melody and chords transposition that is clearly preferred by the majority of listeners for the majority of test items while, in turn, Melodyne editor was preferred in terms of timbre preservation for most test items.

To improve the timbre of the MODVOC processed items, further work might be necessary on HL functionality. The potential for improvement is motivated by the observation that in the preference test results for synthetic solo piano MODVOC processing was clearly preferred in terms of timbre preservation, while scoring on par with regard to melody and chords transposition. This already indicates a good fit of the current HL functionality for selected classes of signals. It is conceived that a more advanced psychoacoustic model within HL, e .g. rating the perceptual importance of overtones, would enhance the timbre preservation characteristics of the MODVOC based selective pitch transposition for a broader range of signal classes.

Additionally, a listening test assessing perceptual quality in the case of synthesis of unaltered MODVOC components has been performed on so-called *critical test items*, which are known to pose special challenges on any audio processing system. This test can be regarded as an upper bound for the perceptual quality that can be obtained by applying the MODVOC to arbitrary audio signals. The test results confirm that for predominantly tonal signals near-transparent audio quality ranging between «good» and «excellent» is achieved.

In summary, the newly introduced MODVOC processing scheme has been shown to take the human auditory perception sufficiently well into account, in order to deliver modification and reproduction results that have been rated by listeners in a range spanning from «fair» to «good», even for well-known critical test items.

Apart from the application of selective pitch transposition presented in this thesis, modulation processing can also be a promising basis for other application fields, e.g. efficient perceptual coding of audio signals. In audio coding, it is common practice to apply waveform coding only to a lowpass filtered base band signal and synthesize an approximation of the original high band content by *transposition* of the baseband signal and subsequent application of a parameter driven post-processing for shaping the transposed signal [18][59][50]. Contemporary implementations of this so-called *band-*

width extension (BWE) often suffer from unpleasant roughness artifacts [75][76]. Since modulation based audio processing provides a handle to both pitch transposition via manipulation of the estimated carrier frequencies, and, at the same time, to auditory roughness, which is related to the modulation fine structure, future bandwidth extension techniques could also benefit from the use of the technology developed and discussed in this thesis.

Another field of application is audio watermarking. Audio watermarking denotes a technique for inaudibly embedding an additional information signal into an audio signal. Audio watermarks are used for copyright protection purposes, automatic broadcast monitoring or conveyance of side information over existing (analog) channels, where a dedicated side information channel is not available. Typically, watermark embedding technologies exploit *auditory masking* or *threshold-in-quiet* effects of human auditory perception. Since, in the modulation domain, such masking effects also exist [49], these could be very well utilized [24] for data hiding purposes through the application of the MODVOC.

Bibliography

[1] J. Anantharaman, A. Krishnamurthy, and L. Feth. Intensity-weighted average of instantaneous frequency as a model for frequency discrimination. *Journal of the Acoustical Society of America (JASA)*, 94:723–729, 1993.

[2] ANSI. Ansi standard s1.4-1983, 1983.

[3] ANSI. Ansi standard s1.42-2001, 2001.

[4] L. Atlas and J. Fang. Quadratic detectors for general nonlinear analysis of speech. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2:9–12, 1992.

[5] L. Atlas and C. Janssen. Coherent modulation spectral filtering for single-channel music source separation. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 461–464, 2005.

[6] W. Aures. Ein berechnungsverfahren der rauhigkeit. *Acustica*, 58:268 28, 1985.

[7] P. Barbieri. The inharmonicity of musical string instruments (1543-1993). with an unpublished memoir by j.-b. mercadier (1784). *Studi Musicali*, XXVII-2:385–419, 1998.

[8] J. P. Bello. *Towards the automated analysis of simple polyphonic music: A knowledge-based approach 2003.* PhD thesis, University of London, London, UK, 2003.

[9] H. Bode. History of electronic sound modification. *Journal of the Audio Engineering Society (JAES)*, 32/10:730–739, 1984.

[10] A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound.* MIT Press, 1990.

[11] G. J. Brown. *Computational auditory scene analysis: a representational approach.* PhD thesis, University of Sheffield, Sheffield, UK, 1992.

[12] J. Brown and K. Vaughn. Pitch center of stringed instrument vibrato tones. *Journal of the Acoustical Society of America (JASA)*, 100 (3):1728–1735, 1996.

[13] A. Christensen, M. G and. Jakobsson, S. V. Andersen, and S. H. Jensen. Amplitude modulated sinusoidal signal decomposition for audio coding. *IEEE Signal Processing Letters*, 13(7):389–392, 2006.

[14] M. G. Christensen and S. H. Jensen. Computationally efficient amplitude modulated sinusoidal audio coding using frequency-domain linear prediction. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 5:61–64, May 2006.

[15] M. G. Christensen, S. van de Par, S. H. Jensen, and S. V. Andersen. Multiband amplitude modulated sinusoidal audio modeling. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 4:169–172, 2004.

[16] P. Dallos, N. B. Slepecky, P. Wangemann, J. Schacht, R. Patuzzi, E. de Boer, C. J. Kros, M. C. Holley, J. J. Guinan, and W. F. Sewell. *The Cochlea.* Springer Handbook of Auditory Research. Springer, 1996.

[17] P. Daniel and R. Weber. Psychoacoustical roughness: Implementation of an optimized model. *Acustica*, 83:113–123, 1997.

[18] M. Dietz, L. Liljeryd, K. Kjörling, and O. Kunz. Spectral band replication, a novel approach in audio coding. *Proceedings of the 112th AES Convention, Munich*, 2002.

[19] S. Disch and B. Edler. An amplitude- and frequency modulation vocoder for audio signal processing. *11th International Conference on Digital Audio Effects (DAFX-08)*, 2008.

[20] S. Disch and B. Edler. An iterative segmentation algorithm for audio signal spectra depending on estimated local centers of gravity. *12th International Conference on Digital Audio Effects (DAFX-09)*, 2009.

[21] S. Disch and B. Edler. Multiband perceptual modulation analysis, processing and synthesis of audio signals. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.

[22] S. Disch and B. Edler. An enhanced modulation vocoder for selective transposition of pitch. *13th International Conference on Digital Audio Effects (DAFX-10)*, 2010.

[23] S. Disch and B. Edler. Frequency selective pitch transposition of audio signals. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

[24] S. Disch, J. Herre, and J. Kammerl. Audio watermarking using subband modulation spectra. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 245–248, April 2007.

[25] H. Dudley. The vocoder. *Bell Labs Record*, 17:122–126, 1939.

[26] H. Dudley. The carrier nature of speech. *Bell System Technical Journal*, 19:495–515, 1940.

[27] C. Duxbury, M. Davies, and M. Sandler. Improved time-scaling of musical audio using phase locking at transients. *Proceedings of the 112th AES Convention, Munich*, 2002.

[28] M. Elhilali. *Neural basis and computational strategies for auditory processing.* PhD thesis, University of Maryland, College Park, 2004.

[29] M. Elhilali, T. Chi, and S. A. Shamma. A spectro-temporal modulation index (stmi) for assessment of speech intelligibility. *Speech communication*, 41:331–348, 2003.

[30] M. Elhilali and S. Shamma. A biologically-inspired approach to the cocktail party problem. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, May 2006.

[31] D. P. W. Ellis. *Prediction-driven computational auditory scene analysis.* PhD thesis, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA, 1996.

[32] G. Fairbaks and R. P. Jaeger. Method for time or frequency compression-expansion of speed. *Institute of Radio Engineers Transactions on Audio*, AU-2:7–12, 1954.

[33] L. L. Feth. Frequency discrimination of complex periodic tones. *Attention, Perception, & Psychophysics*, 15:375–378, 1977.

[34] J. L. Flanagan. Parametric coding of speech spectra. *Journal of the Acoustical Society of America (JASA)*, 68 (2):412–419, 1980.

[35] J. L. Flanagan and R. M. Golden. Phase vocoder. *Bell System Technical Journal*, 45:1493–1509, 1966.

[36] H. Fletcher. Auditory patterns. *Reviews of Modern Physics*, 12:47–65, 1940.

[37] H. Fletcher, E. Blackham, and R. Stratton. Quality of piano tones. *Journal of the Acoustical Society of America (JASA)*, 6:749–761, 1962.

[38] H. Fletcher and W. Munson. Loudness, its definition, measurement and calculation. *Journal of the Acoustical Society of America (JASA)*, 5:82–108, 1933.

[39] E. B. George and M. J. T. Smith. Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones. *Journal of the Audio Engineering Society (JAES)*, pages 497–516, 1992.

[40] W.-B. Goh and K.-Y. Chan. Amplitude modulated sinusoidal modeling using least-square infinite series approximation with applications to timbre analysis. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 6:3561–3564, 1998.

[41] D. W. Griffin and J. S. Lim. Signal estimation from modified short-time fourier transform and its application to speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-32(2):236–243, April 1984.

[42] H. M. Hanson, P. Maragos, and A. Potamianos. Finding speech formants and modulations via energy separation: with application to a vocoder. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2:716–719, 1993.

[43] W. M. Hartmann. *Signals, Sound, and Sensation.* Modern Acoustics and Signal Processing. AIP Press, 5th printing 2004 edition, 1998.

[44] J. Heckroth. *Complete MIDI 1.0 Detailed Specification.* MIDI Manufacturers Association (MMA), 1996.

[45] J. Herre. Temporal noise shaping, quantization and coding methods in perceptual audio coding: A tutorial introduction. *Proceedings of the 17th AES International Conference on High Quality Audio Coding, Florence*, pages 312–325, 1999.

[46] J. Herre and J. D. Johnston. Enhancing the performance of perceptual audio coders by using temporal noise shaping (tns). *Proceedings of the 101st AES Convention, Los Angeles*, (Preprint 4384), 1996.

[47] J. Herre and J. D. Johnston. A continuously signal-adaptive filterbank for high-quality perceptual audio coding. *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk*, 1997.

[48] J. Herre and J. D. Johnston. Exploiting both time and frequency structure in a system that uses an analysis / synthesis filterbank with high frequency resolution. *Proceedings of the 103rd AES Convention, New York*, (Preprint 4519), 1997.

[49] T. Houtgast. Frequency selectivity in amplitude-modulation detection. *Journal of the Acoustical Society of America (JASA)*, 85(4):1676–1680, April 1989.

[50] ISO. 14496-3:2001/fpdam 1: Bandwidth extension, 2001.

[51] ITU-R. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, 1994-1997.

[52] ITU-R. Method for the subjective assessment of intermediate sound quality (mushra), 2001.

[53] J. F. Kaiser. On a simple algorithm to calculate the energy of a signal. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 381–384, 1990.

[54] A. Klapuri. *Signal Processing Methods For the Automatic Transcription of Music.* PhD thesis, Tampere University of Technology, 2004.

[55] R. Kumaresan and A. Rao. Minimum/maximum phase decomposition of signals inspired by the auditory periphery. In *Signals, Systems and Computers, 1995. 1995 Conference Record of the Twenty-Ninth Asilomar Conference on*, volume 2, pages 1239–1243, Oct./Nov. 1995.

[56] R. Kumaresan and A. Rao. Algorithm for decomposing an analytic signal into am and positive fm components. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 1561–1564, May 1998.

[57] J. Laroche and M. Dolson. Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio Processing*, 7(3):323–332, 1999.

[58] J. Laroche and M. Dolson. New phase-vocoder techniques for real-time pitch-shifting, chorusing, harmonizing and other exotic audio modifications. *Journal of the Audio Engineering Society (JAES)*, 11(47):928–936, 1999.

[59] E. Larsen and R. Aarts. *Audio Bandwidth Extension - Application to psychoacoustics, Signal Processing and Loudspeaker Design.* John Wiley & Sons, Ltd, 2004.

[60] Q. Li and L. Atlas. Over-modulated am-fm decomposition. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 5559, pages 172–183, Oct. 2004.

[61] Q. Li and L. Les Atlas. Coherent modulation filtering for speech. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4481–4484, 2008.

[62] P. J. Loughlin and B. Tacer. On the amplitude- and frequency-modulation decomposition of signals. *Journal of the Acoustical Society of America (JASA)*, 100(3):1594–1601, 1996.

[63] P. J. Loughlin and B. Tacer. Comments on the interpretation of instantaneous frequency. *Signal Processing Letters, IEEE*, 4(5):123–125, May 1997.

[64] D. Malah. Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):121–133, April 1979.

[65] P. Maragos, J. F. Kaiser, and T. F. Quatieri. Energy separation in signal modulations with application to speech analysis. *IEEE Transactions on Signal Processing*, 41 (10):3024–3051, 1993.

[66] P. Maragos, T. F. Quatieri, and J. F. Kaiser. Speech nonlinearities, modulations, and energy operators. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1:421–424, 1991.

[67] K. Martin. A blackboard system for automatic transcription of simple polyphonic music. Technical report, MIT Media Lab, 1995.

[68] K. D. Martin. *Sound-Source Recognition: A theory and Computational Model.* PhD thesis, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA, 1999.

[69] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinussoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):744–754, August 1986.

[70] N. Mesgarani and S. Shamma. Speech enhancement based on filtering the spectrotemporal modulations. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 1105–1108, March 18-23, 2005.

[71] N. Mesgarani, S. Shamma, and M. Slaney. Speech discrimination based on multiscale spectro-temporal modulations. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, May 2004.

[72] B. Moore and B. Glasberg. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 4:103–138, 1990.

[73] B. C. J. Moore and B. R. Glasberg. A revision of zwicker's loudness model. *Acta Acustica*, 82:335–345, 1996.

[74] E. Moulines and J. Laroche. Non-parametric techniques for pitch-scale and timescale modification of speech. *Speech Communication*, 16:175–205, 1995.

[75] F. Nagel and S. Disch. A harmonic bandwidth extension method for audio codecs. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 0, pages 145–148, April 2009.

[76] F. Nagel, S. Disch, and S. Wilde. A continuous modulated single sideband bandwidth extension. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 0, pages 357–360, April 2010.

[77] F. Nagel, T. Sporer, and P. Sedlmeier. Towards a statistically well-grounded evaluation of listening tests - avoiding pitfalls, misuse, and misconceptions. *Proceedings of the 128th AES Convention, London*, 2010.

[78] S. H. Nawab, T. F. Quatieri, and J. S. Lim. Signal reconstruction from short-time fourier transform magnitude. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 31(4):986–998, August 1983.

[79] P. Neubäcker. Method for acoustic object-oriented analysis and note object-oriented processing of polyphonic sound recordings (ep2099024), September 2009.

[80] K. Nie and F.-G. Zeng. A perception-based processing strategy for cochlear implants and speech coding. *Proceedings of the 26th IEEE Conference of the Engineering in Medicine and Biology Society*, 6:4205–4208, 2004.

[81] M. D. Plumbley, S. A. Abdallah, J. P. Bello, M. E. Davies, G. Monti, and M. B. Sandler. Automatic music transcription and audio source separation. *Cybernetics and Systems*, 33 (6):603–627, 2002.

[82] A. Potamianos and P. Maragos. A comparison of the energy operator and the hilbert transform approach to signal and speech demodulation. *Signal Processing*, 37:95–120, 1994.

[83] A. Potamianos and P. Maragos. Speech formant frequency and bandwidth tracking using multiband energy demodulation. *Journal of the Acoustical Society of America (JASA)*, 99:3795–3806, 1996.

[84] A. Potamianos and P. Maragos. Speech analysis and synthesis using an am-fm modulation model. *Speech Communication*, 28(3):195–209, 1999.

[85] E. Prame. Vibrato extent and intonation in professional western lyric singing. *Journal of the Acoustical Society of America (JASA)*, 102(1):616–621, 1997.

[86] H. Purnhagen. *Very Low Bit Rate Parametric Audio Coding.* PhD thesis, Gottfried Wilhelm Leibniz Universität Hannover, Fakultät für Elektrotechnik und Informatik, 2008.

[87] H. Purnhagen, B. Edler, and C. Ferekidis. Object-based analysis/synthesis audio coder for very low-bit rates. *Proceedings of the 104th AES Convention, Amsterdam*, (Preprint 4747), 1998.

[88] T. F. Quatieri and A. V. Oppenheim. Iterative techniques for minimum phase signal reconstruction from phase or magnitude. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(6):1187–1193, December 1981.

[89] A. Rao and R. Kumaresan. On decomposing speech into modulated components. *IEEE Transactions on Speech and Audio Processing*, 8(3):240–254, May 2000.

[90] A. Röbel. A new approach to transient processing in the phase vocoder. *13th International Conference on Digital Audio Effects (DAFX-03)*, pages 344–349, 2003.

[91] A. Röbel. Transient detection and preservation in the phase vocoder. *International Computer Music Conference (ICMC'03)*, pages 247–250, 2003.

[92] J. Roederer. *Physikalische und psychoakustische Grundlagen der Musik.* Springer, 1977.

[93] S. Roucos and A. M. Wilgus. High quality time-scale modification for speech. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 493–496, 1985.

[94] S. Salani and D. Smith. A programmable signal processing system. *Proceedings of the 70th AES Convention, Anaheim*, October 1981.

[95] E. D. Scheirer. Sound scene segmentation by dynamic detection of correlogram comodulation. In *the International Joint Conference on AI Workshop on Computational Auditory Scene Analysis*, 1999.

[96] S. Schimmel and L. Atlas. Coherent envelope detection for modulation filtering of speech. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[97] S. M. Schimmel. *Theory of Modulation Frequency Analysis and Modulation Filtering, with Applications to Hearing Devices*. PhD thesis, University of Washington, Seattle (Washington), USA, 2007.

[98] S. M. Schimmel, L. Atlas, and K. Nie. Feasibility of single channel speaker separation based on modulation frequency analysis. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 4:605–608, 2007.

[99] S. M. Schimmel, K. R. Fitz, and L. Atlas. Frequency reassignment for coherent modulation filtering. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 5:261–264, 2006.

[100] X. Serra. *Musical Sound Modeling with Sinusoids plus Noise*. Swets & Zeitlinger, 1997.

[101] X. Serra and J. O. Smith. Spectral modeling and synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, pages 12–24, 1990.

[102] W. A. Sethares. *Tuning, Timbre, Spectrum, Scale*. Springer, second edition edition, 2004.

[103] T. Sporer, J. Liebetrau, and S. Schneider. Statistics of mushra revisited. *Proceedings of the 127th AES Convention, New York*, 2009.

[104] R. Sussman. Analysis and resynthesis of musical instrument sounds using energy separation. Master's thesis, The State University of New Jersey, Graduate School, New Brunswick, 1993.

[105] E. Terhardt. On the perception of periodic sound fluctuations (roughness). *Acustica*, 30:201–213, 1974.

[106] E. Terhardt. *Akustische Kommunikation.* Springer, 1998.

[107] J. Thiemann and P. Kabal. Reconstructing audio signals from modified non-coherent hilbert envelopes. *Proceedings of Interspeech (Antwerp, Belgium)*, pages 534–537, 2007.

[108] J. K. Thompson and L. Atlas. A non-uniform transform for audio-coding with increased time resolution. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 397–400, 2003.

[109] J. Timoney and T. Lysaght. Eps models of am-fm vocoder output for new sounds generations. *3rd International Conference on Digital Audio Effects (DAFX-01)*, December 2001.

[110] T. Tolonen. *Object-Based Sound Source Modeling.* PhD thesis, Helsinki Univ. of Technology, Lab. of Acoustics and Audio Signal Processing, Espoo, Finland, 2000.

[111] H. Traunmüller. Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America (JASA)*, 88(1):97–100, 1990.

[112] W. Verhelst and M. Roelands. An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 554–567, 1993.

[113] M. S. Vinton and L. Atlas. A scalable and progressive audio codec. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3277–3280, 2001.

[114] H. B. Voelcker. Toward a unified theory of modulation–part i: Phase-envelope relationships. *Proceedings of the IEEE*, 54(3):340–351, 1966.

[115] H. B. Voelcker. Toward a unified theory of modulation–part ii: Zero manipulation. *Proceedings of the IEEE*, 54(5):735–755, 1966.

[116] L. L. M. Vogten. *Facts and models in hearing. Pure tone masking: a new result from a new method.* Springer, 1974.

[117] H. von Helmholtz. *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik.* Vieweg, Braunschweig, 1863.

[118] Y. Wang and R. Kumaresan. Real time decomposition of speech into modulated components. *Journal of the Acoustical Society of America (JASA)*, 119 (6), 2006.

[119] D. Wei and A. Bovik. On the instantaneous frequencies of multicomponent am-fm signals. *IEEE Signal Processing Letters*, 5:84–86, 1998.

[120] Q. Xu, L. L. Feth, J. N. Anantharaman, and A. K. Krishnamurthy. Bandwidth of spectral resolution for the "c-o-g" effect in vowel-like complex sounds. *Journal of the Acoustical Society of America (JASA)*, 101, May 1997.

[121] X. Yang, K. Wang, and S. A. Shamma. Auditory representations of acoustic signals. *IEEE Transactions on Information Theory*, 38 (2):824–839, 1992.

[122] C. Yeh. *Multiple fundamental frequency estimation of polyphonic recordings.* PhD thesis, École doctorale edité, Université de Paris, 2008.

[123] W. A. Yost, D. M. Green, B. C. J. Moore, N. F. Viemeister, C. J. Plack, F. L. Wightman, D. J. Kistler, and S. Sheft. *Human Psychophysics.* Springer Handbook of Auditory Research. Springer, 1993.

[124] U. Zölzer, editor. *DAFX: Digital Audio Effects.* John Wiley & Sons, 2002.

[125] E. Zwicker and H. Fastl. *Psychoacoustics - Facts and Models.* Springer, Berlin - Heidelberg, 2. edition, 1999.

Lebenslauf

Sascha Illja **Disch**

- geboren am 09.11.1968 in Freiburg im Breisgau
- verheiratet, zwei Kinder
- Staatsangehörigkeit: Deutsch

Ausbildung

- 1976 - 1980 Johannes Schwartz Grundschule
- 1980 - 1989 Wentzinger Gymnasium Freiburg-West
- 1989 **Abschluss: Allgemeine Hochschulreife**
- 1989 - 1992 Ausbildung zum Kommunikationselektroniker am Fraunhofer Institut für angewandte Festkörperphysik (IAF), Freiburg
- 1992 **Abschluss: Kommunikationelektroniker, Fachrichtung Informationstechnik**
- 1993 - 1999 Studium der Elektrotechnik an der Universität Hamburg-Harburg (TUHH)
- 1999 **Abschluss: Diplomingenieur Elektrotechnik (Nachrichtentechnik)**

Beruf

- 1992 - 1993 Kommunikationselektroniker am Fraunhofer Institut für angewandte Festkörperphysik (IAF), Freiburg
- 1999 - 2007 Wissenschaftlicher Mitarbeiter am Fraunhofer Institut für integrierte Schaltungen (IIS), Erlangen
- 2007 - 2010 Wissenschaftlicher Mitarbeiter am Laboratorium für Informationstechnologie (LFI), Leibniz Universität Hannover
- ab 2010 Wissenschaftlicher Mitarbeiter am Fraunhofer Institut für integrierte Schaltungen (IIS), Erlangen