Processing FM-sweeps: psychophysical sensitivity to acoustic spectrotemporal modulations of tone complexes in macaques and humans

Master thesis 2009

Anne M. M. Fransen[1]

Supervisors:
Robert F. Van der Willigen[1],
A. John Van Opstal[1]

In collaboration with:
Huib Versnel[1,2]

Second corrector:
Pascal Fries

[1]Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, Dept. Biophysics , PO Box 9101, 6500 HB Nijmegen, The Netherlands
[2]University Medical Center Utrecht, Rudolf Magnus Institute of Neuroscience, Dept. Otorhinolaryngology, PO Box 85500, 3508 GA Utrecht, The Netherlands

Contents

Processing FM-sweeps: psychophysical sensitivity to acoustic spectrotemporal modulations of tone complexes in macaques and humans

Anne M. M. Fransen[1], Robert F. Van der Willigen[1], A. John Van Opstal[1], and Huib Versnel[1,2]

[1]*Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, Dept. Biophysics , PO Box 9101, 6500 HB Nijmegen, The Netherlands*

[2]*University Medical Center Utrecht, Rudolf Magnus Institute of Neuroscience, Dept. Otorhinolaryngology, PO Box 85500, 3508 GA Utrecht, The Netherlands*

Abstract

Frequency-modulated (FM) sweeps are important components of natural sounds. However, to analyse these, the auditory system must perform combined spectral-temporal analysis. Two models on spectral-temporal analysis in the auditory system exist, one of which is based on combined spectrotemporal feature detectors, whereas the other model assumes separate spectral and temporal filter banks. To discriminate between these models we determined the psychometrical response curves to a large set of broadband spectral-temporal rippled noise stimuli (N=968). From the perceptual thresholds we derived the spectral-temporal transfer characteristic for each subject (5 rhesus monkeys, 5 humans). Our results confirmed all predictions that follow from separate spectral and temporal processing: i) Temporal modulation and spectral spacing do not interact (i.e. the transfer characteristic is separable) ii) The auditory system is equally sensitive to upward and downward moving ripples. iii) The transfer functions are comparable between humans and monkeys. We propose a model to unite separable processing with the ability of the auditory system to encode inseparable FM-sweeps.

KEYWORDS: cocktail party problem (CPP), auditory streaming, free-field audiogram, spectral-temporal separability, primate hearing, modulation filter bank, frequency-modulated sweeps (FM-sweeps)

Introduction

One of the most fundamental questions in auditory research is how the auditory system deals with combined spectral and temporal information. In this paper we studied the interaction between the spectral and temporal domain behaviourally. The hypotheses and set-up will be described at the end of this chapter; first we will shortly review auditory streaming and models on combined spectral-temporal auditory processing.

FM-sweeps and auditory streaming

Time and spectrum cannot be seen separately in many vocalizations and natural sounds, as these are characterised by combined modulations in frequency and time (FM-sweeps). This is illustrated by an example of human speech in Fig. 1, showing the spectrogram of the phrase "Your test starts now". The vowels in this phrase can be easily recognised by the regularly spaced high amplitude (red) bands that change in frequency over time (i.e. regularly spaced FM-sweeps). The direction (upward or downward) and speed of these sweeps allow one to discriminate between different vowels.

Acoustic features, like FM-sweeps, are not only important to discriminate between sounds, but also for perception in noise. Perception in noise is an ill-posed problem[1], since each frequency channel may contain contributions from both the source of interest and noise. This makes perception in noise a key problem for both the auditory system and those who study it, as noise is always present in natural environments and since both the sounds of interest and the background noise can be highly variable, ranging from white-noise to multiple interfering speakers, like at a cocktail party (Aubin and Jouventin, 1998;

[1] A problem is *ill-posed* if it contains many different solutions that cannot be constrained. E.g. a+b=7 has infinitely many sets [a,b] that obey the equation. To reconstruct the original set [a,b] constraints are needed. It is thought that the auditory system uses acoustic features to constrain the solutions.

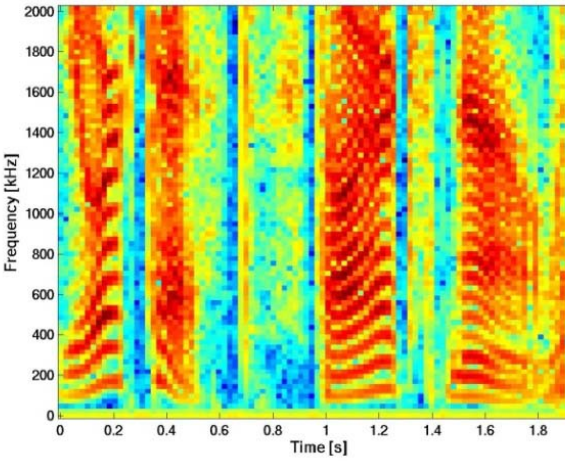

Fig. 1. Sonogram of a male human voice uttering the sentence: "Your test starts now". Note the harmonic structure of the vowels ['ou', 'e', 'a', 'ow'], as well as the dynamic upward and downward modulations across the different frequency bands.

Darwin, 2008). Much study has been devoted to perception in noise, and the cocktail-party problem (CPP) in particular. Yet, half a century after Cherry (1957) introduced the CPP, it is still not understood how the human auditory system is able to specifically attend to one of many sound sources in a noisy environment.

It is thought the auditory system deals with environmental noise via streaming. It uses acoustic features to group acoustic elements that belong to only one sound source into one perceptually coherent representation, or stream (Bregman, 1990) (for review, see Haykin and Chen (2005)).

Especially three acoustic features are thought to function as important criteria to group or segregate sound elements into streams. These are: 1) regular spectral spacing (harmonic spectra), which is typically caused by natural vibrations (Roberts and Bregman, 1991; Roberts and Bailey, 1993, 1996; Brunstrom and Roberts, 2000; Roberts, 2005); 2) across-frequency patterns of coherent amplitude modulations (Bregman et al., 1982; Langemann et al., 2005; Nassiri and Escabi, 2008) and 3) simultaneous onsets and offsets in different frequency channels. For example, simultaneous offset can be observed for the letter "t" in Fig. 1, which starts with a simultaneous offset in all frequencies (the onset of a blue vertical bar at about 0.3, 0.6, 0.9 and 1.3 s) (Darwin, 1981; Darwin and Ciocca, 1992; Ciocca and Darwin, 1993; Hukin and Darwin, 1995; Darwin and Hukin, 1998). In addition to these cues spatial location has been suggested to aid in source separation (Bregman, 1990; Eramudugolla et al., 2008).

Although streaming based on the cues described above (harmonics, across-frequency patterns of coherent amplitude modulations, simultaneous onsets and offsets and spatial location), can partly solve the CPP and can explain some experimental data on speech perception in noise, none of these cues – nor a combination of them – allow the grouping of an FM-sweep into one stream. This means that some cue or concept must still be missing, since FM-sweeps are believed to be important and omnipresent features of natural acoustic stimuli.

Separability

An important characteristic of FM-sweeps is that they are spectral-temporally inseparable, meaning that the normalised (at t=0) amplitude function over time (or frequency) is different for different frequencies (or time points) (see Fig. 2B). In mathematical terms this means that the entire spectrogram cannot be described by the outer product of one temporal and one spectral function. For clarity, hereafter we refer to inseparable combined spectral-temporal objects (i.e. inseparable sounds, filters or spectrograms) as 'spectrotemporal' (Fig. 2B), and to separable combined spectral-temporal objects as 'spectral-temporal' (Fig. 2A).

A problem in studying responses to inseparable stimuli is that they are much more complex than (amplitude modulated) pure tones or Gaussian white noise, since they are modulated in two dimensions rather than in one. An efficient way to describe arbitrary sounds is by a representation that uses mutually independent basis functions. Any arbitrary spectrogram, however complex, can uniquely be represented by a superposition of so called *rippled noise stimuli* (ripples). These consist of a superposition of a large number of tones that are amplitude-modulated in a specific way. This is similar to the way the Fourier series (a weighted superposition of elementary harmonic functions i.e. sines and cosines) can uniquely represent any stationary periodic signal. Notably, the spectrograms of rippled noise stimuli are reminiscent to the well-known spatial gratings that are used in vision

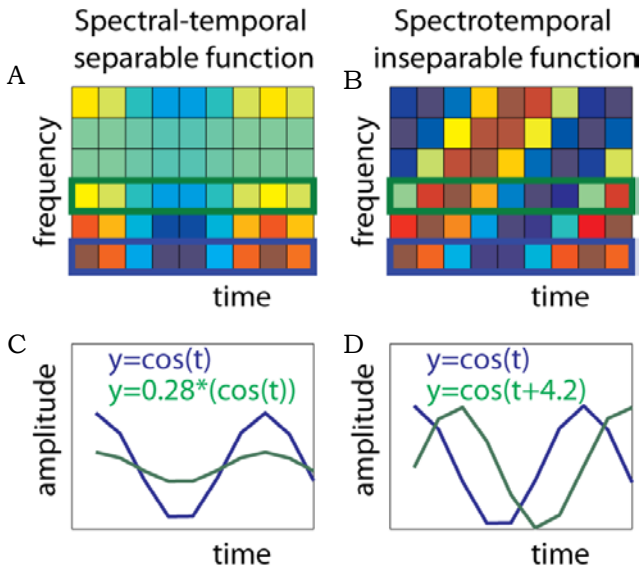

Fig. 2. Spectral-temporal separable (*A*) and spectrotemporal inseparable (*B*) spectrograms. Note that *A* can be written as: $\psi(f,t) = A(f) \cdot \cos(\alpha \cdot t)$. Thus, the temporal behavior is identical across frequencies, except for a scaling factor *A(f)* (C). Panel *B* shows an FM-sweep: $\psi(f,t) = \cos(\alpha \cdot t + \beta(f))$. This function is inseparable because the frequency-dependent phase alters the function describing amplitude over time for each frequency channel (*D*), such that the spectrogram cannot be described by the outer product of a spectral and a temporal function.

research (see examples in Fig. 4).

Since rippled noise stimuli can be parameterised and used to decompose any arbitrary spectrogram into a unique set of independent spectral-temporal components, ripple stimuli have sometimes been used in electrophysiological animal studies to determine so-called spectral-temporal receptive fields (STRFs) of auditory cells (Kowalski et al., 1996b, a; Shamma, 1996; Chi et al., 1999; Klein et al., 2000; Versnel et al., 2009). Analysis of such receptive fields showed that they are not always separable. Furthermore, it has been reported that the percentage of inseparable neurons tends to increase from the midbrain *colliculus inferior* to auditory cortex, A1 (Felsheim and Ostwald, 1996; Depireux et al., 2001; Linden et al., 2003; Versnel et al., 2009), for review, see (Joris et al., 2004).

An auditory filter bank: two models

To deal with spectrotemporal sounds, like FM-sweeps, the auditory system needs to be able to take into account phase differences between frequency channels. Also, it needs to be specifically sensitive to those patterns of phase differences that are behaviourally relevant. This can be achieved via spectrotemporal filters:

Inseparable spectrotemporal filters

Neurons with rippled STRFs (as in Fig. 2B) can not only represent FM-sweeps with diverse speeds and spectral resolutions, but also directly represent any other sound feature involved in streaming, with the exception of spatial location. Thus such a filter bank can encode any acoustic feature that is used in perception. However, this would mean that frequency and time are not represented separately throughout the auditory system, as is often believed. Instead, to represent sounds, a large 'filter bank' of FM-sweeps (elementary acoustic features) would be required within the audible range, analogous to the representation of oriented line segments in the visual system. A computational model that uses a filter bank of FM-sweeps was recently described by Shamma and colleagues (Chi et al., 2005; Elhilali and Shamma, 2008).

Separable spectral-temporal filters

In contrast to Shamma's spectrotemporal model, Dau and colleagues (Dau et al., 1997; Jepsen et al., 2008) proposed auditory filters to describe the sensitivity of the auditory system to different temporal modulations, with the important difference that their filter bank has no spectral dimension. Hence, Dau's model is not able to deal with FM-sweeps, because information on phase differences between spectral channels is lost.

Important to our study are the subsequent spectral (blue) and temporal (green) amplitude modulation filter banks (Fig. 3). As Dau's model has separate temporal and spectral filter banks, frequency and time are treated as independent variables, and sounds are represented along mutually orthogonal frequency-time dimensions. This view is supported by the finding that sensitivity to temporal modulations is independent of the frequency of the pure tone carrier used (Kohlrausch et al., 2000). However, some neurons with inseparable spectrotemporal receptive fields have been found, whose

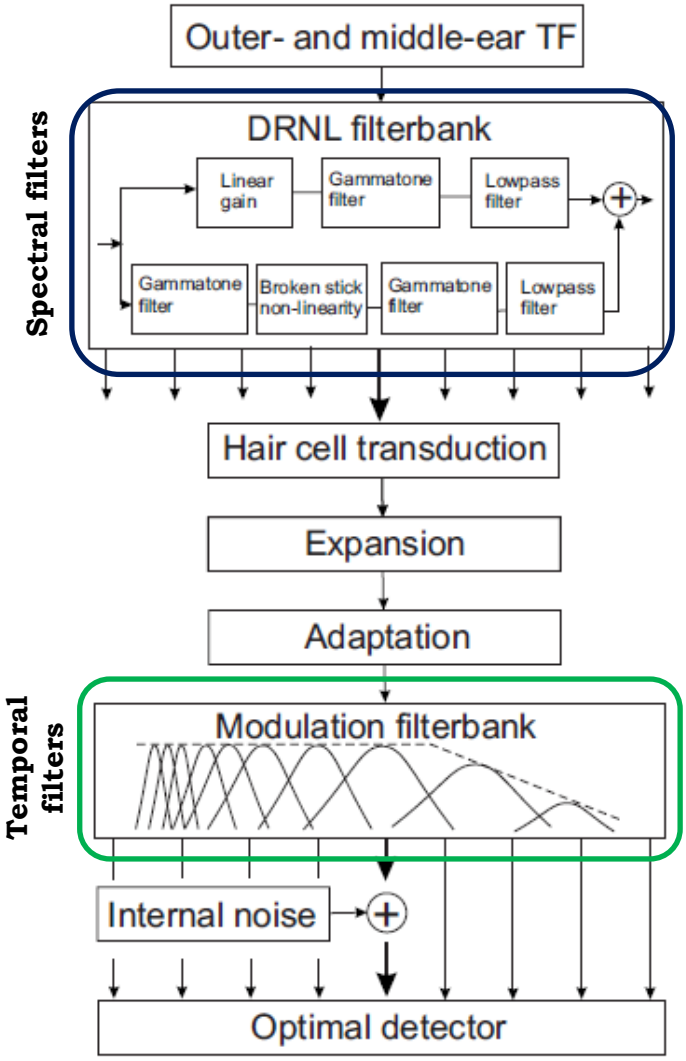

Figure 3. Auditory processing modeled by Dau and colleagues (Jepsen et al., 2008). Note that one set of temporal modulation filters is applied on all frequency channels. The separate spectral (blue) and temporal (green) filters predict that perception depends independently on the spectral and temporal properties of sound.

existence and function remain unexplained by Dau's model. In addition it is unclear how a separable auditory system could be able to deal with FM-sweeps, as it cannot detect across-channel phase differences.

An important advantage compared to Shamma's model, however, is that less filters are required and that it allows spectral topography throughout the auditory system.

Ideally, to differentiate between these models one would need a complete characterisation of all the neurons and networks involved. In practice, however, this is not feasible, and discrimination at a behavioural level is preferable, if possible.

From neurons to percept: outline of the present study

Shamma's and Dau's model differ critically in their predictions regarding sensitivity to inseparable stimuli at a behavioural level. Shamma's model predicts that sensitivity to an inseparable sound depends on the spectrotemporal properties of the sound. In contrast, Dau's model predicts that sensitivity is dependent on the independent spectral and temporal properties of an inseparable sound.

It makes sense to test these behavioural spectrotemporal sensitivities using the same inseparable stimuli that were used to assess the spectral-temporal separability in neurons: spectrotemporal ripples. Although some behavioural data on static (i.e. pure spectral or temporal) ripples (O'Connor et al., 2000) and specific dynamic (i.e. combined spectral-temporal) ripples (Chi et al., 1999; Osmanski et al., 2009) exists, so far the psychophysical detection of rippled stimuli has not been examined systematically.

We used 968 ripple stimuli to measure the behavioural spectral-temporal separability of responses to envelope amplitude modulations in both humans and rhesus monkeys. The acoustic ripples are fully described by three independent parameters: ripple velocity (ω, in Hz) determines amplitude modulations in the temporal dimension; ripple density (Ω, in cycles per octave) specifies the amplitude modulation in the spectral dimension; and modulation depth (ΔM, in %) describes the peak-to-peak amplitude difference of the ripple (ω, Ω, ΔM) as expressed in percentage of the average stimulus loudness.

ΔM was the independent variable in the experiments: for each of the 88 different ripples (ω, Ω) we determined the perceptual detection threshold by measuring the psychophysical curve as a function of ΔM. These thresholds were collected in a behavioural spectral-temporal modulation transfer function (ST-MTF).

Dau's separable model makes four predictions: first, it predicts that the threshold for detecting ripples in noise depends on the independent spectral and temporal modulation rates, and thus the ST-MTF is separable. Second, the ST-MTF is expected to be symmetrical around zero

velocity, i.e. the auditory system is expected to be equally sensitive to upward and downward moving ripples, since direction (i.e. the sign of the ripple) affects neither spectral nor temporal modulation rates. Third, the ST-MTF is expected to show a low-pass filtered shape, as temporal sensitivity increases if more subsequent channels are in phase and vice versa spectral sensitivity increases if spectral patterns are stable over longer time spans Finally, we expect a qualitatively similar ST-MTF across species, as the distribution of behaviourally relevant spectrotemporal modulations does not determine sensitivity. An exception should be made for the overall decay of temporal and spectral sensitivity, respectively, as the underlying physiological causes for these may differ across species (Kohlrausch et al., 2000; Jepsen et al., 2008).

These predictions contrast with those of Shamma's inseparable model: if the filter bank has mixed spectrotemporal dimensions, one would not expect sensitivity to depend only on independent spectral and temporal parameters, and thus the ST-MTF is expected to be inseparable. Also no symmetry between upward and downward ripples, and no low-pass shape are expected, as in principle any ST-MTF could be possible. Instead, the ST-MTF depends on the spectrotemporal tuning distribution of the filters. Finally, ST-MTFs are expected to differ qualitatively across species, as behavioural relevance of the spectrotemporal modulations would differ.

In summary, to distinguish between these two models, we asked the following questions: i) Are ST-MTFs spectral-temporally inseparable?, ii) does the direction of the dynamic modulation (i.e. upward or downward) affect sensitivity? And iii) Are ST-MTFs comparable between humans and monkeys?

Methods

Subjects

Five adult male rhesus monkeys (Macaca mulatta; referred to as M1-M5; weights between 6,5-9,5 kg) and five adult human subjects, ages between 23 and 43 years, participated in the experiments. Two human subjects were naive volunteers (MH and MM) and the remaining three were authors of this paper (RW, HV and AF). Monkeys were trained to detect the onset of an arbitrary ripple in a noisy stimulus. In each session a monkey earned small water rewards per

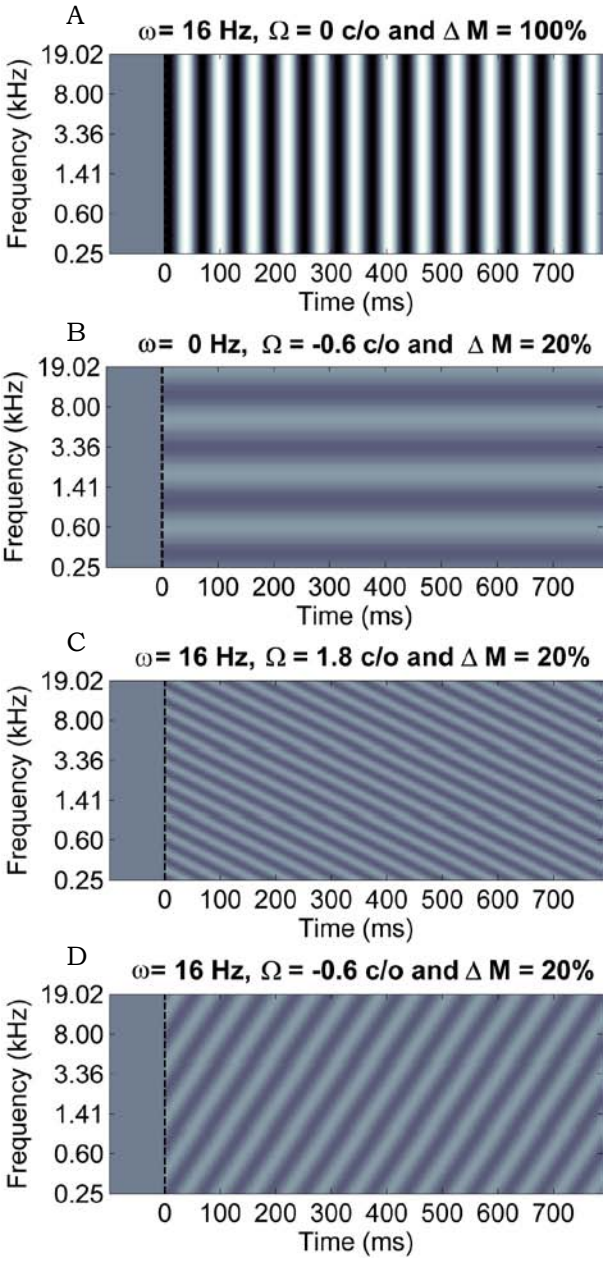

Figure 4. Examples of static (A,B) and dynamic (C,D) rippled noise stimuli used in the experiments. Various spectral-temporal modulations are shown as a function of time. A. A purely temporal modulated ripple with a ripple **velocity (*ω*)** of 16Hz and a **modulation depth (ΔM)** of 100%. B. A purely spectrally modulated stimulus with a **density** of -0.6 cycles per octave **(c/o)** at ΔM=20%.

Panels C and D show two dynamic ripples at *ω*=16Hz and ΔM=20%, with a downward (C; +1.8 c/o) and upward (D) density (-0.6c/o) respectively. Note how the negative sign flips the direction of the slope from downward to upward.

successful trial until he was satiated. Shortly before (~24 hours) recording sessions, daily water intake was limited to about 20 ml/kg. Records were kept of the monkeys' weight and health status, and supplemental fruit was provided after each recording session. During the experiments monkeys were seated in a primate chair with the head unrestrained. All experiments were conducted in accordance with the guidelines for the use of laboratory animals provided by the Society for Neuroscience, and the European Communities Council Directive of November 24, 1986 (86/609/EEC). All experimental procedures were approved by the local ethics committee (DEC) for the use of laboratory animals at the Radboud University Nijmegen.

Stimuli

Static/dynamic ripple stimuli were generated according to Depireux et al. (2001). The ripples consisted of a broadband complex of 126 components equally distributed (20 tones per octave) from f_0=250 Hz to nearly 20 kHz:

$$(1) \quad s(t) = \sum_{n=1}^{126} RIP(t,x) \cdot \sin(2\pi f_n t + \Phi_n)$$

$$\text{with } f_n = f_0 \cdot 2^{(n-1)/20} \text{ Hz}$$

All components had random phase (Φ_n between $-\pi$ and $+\pi$), apart from the f_0=250 Hz component, which had its sine phase fixed to $\Phi=\pi/2$, i.e. at maximum amplitude. The envelopes (described by *RIP*) were modulated in the spectral-temporal domain in the following way.

$$(2) \quad RIP(t,x) = \begin{cases} 1 & \text{for } 0 \le t < D \\ 1 + \Delta M \cdot \sin(2\pi \cdot (\omega t + \Omega x)) & \text{for } t \ge D \end{cases}$$

with t time; x=(n-1)/20, with n=1:126 the position of the spectral component in octaves above f_0; ω the ripple velocity (in Hz); Ω the ripple density (in cycles/octave, or c/o); and ΔM the modulation depth on a linear scale between 0 and 1. D is the duration of the static ripple complex. The sound intensity (RMS) was fixed at 56 dB SPL for both the static noise and the rippled noise.

The stimuli were selected from 11 ripple densities between -3.0 and +3.0 c/o in steps of 0.6 c/o and from 8 ripple velocities [0, 4, 8, 16, 32, 64, 128 and 256 Hz], for up to 11 different modulation depths (ΔM = 2.5, 5, 7.5, 10, 15, 20, 30, 40, 50, 70 and 100%). Subjects AF and RW were also presented with stimuli that had intermediate ripple velocities, which doubled the temporal resolution of their resulting ST-MTFs. Note that $\Omega < 0$ corresponds to an upward direction of the spectral envelope (Fig. 4D); $\Omega > 0$ to a downward direction (Fig. 4C); and that $\Omega = 0$ is equivalent to a pure amplitude modulation (AM) (Fig. 4A), and $\omega=0$ to a pure spectral modulation (Fig. 4A). One human subject (HV) and two monkeys (M4, M5) had been tested in an earlier version of the experiment, in which they received stimuli with slightly different densities (-3.2:.8:3.2 c/o).

Sound stimuli were created in Matlab (R2008a) at 49 kHz sampling rate, and delivered to a speaker by use of TDT hardware. Stimuli were presented in the free field at the frontal central position by a speaker (Philips AD-44725; Blaupunkt pcxg352; or Visaton GmbH SC5.9) with a flat frequency characteristic within 5 dB between 0.2 and 20 kHz after equalization (Behringer Ultra-Curve). The sound intensity (RMS) was measured at the position of the subjects' head with a calibrated sound amplifier and microphone (Brüel and Kjær; BK2610/BK4134). Ambient background noise level was 50-55 dB SPL and any reflections above 500 Hz were effectively absorbed by acoustic foam that was mounted on the walls, floor, ceiling, and every large object present.

Ripple detection paradigm

We used the method of constant stimuli to estimate the psychometric curve for each ripple. Subjects were trained (monkeys) or instructed (humans) to press a handle until they detected a change (i.e. the onset of a ripple) in the stimulus, upon which they had to release the handle as fast as possible. The static noise duration D was randomly varied between 1000 and 4000 ms (1000 to 3000 ms for human subjects) and the ripple lasted for 800 (1000) ms. Catch trials (i.e. trails in which the ripple has zero velocity and zero density, and thus without modulation) were presented to estimate guess rates. Monkeys received a water reward of 5 ml for every correct detection (i.e. response latencies between 220-800 ms after ripple onset) and a smaller one (2 ml) for misses (i.e. response latencies >800 ms) to maintain the monkey's motivation.

The different ripples and modulation depths were randomly interleaved, such that the specific target ripple was not known in

advance. Given the large variety in spectral-temporal properties (11*8), this strongly reduced the advantage of attending to a specific (temporal or spectral) cue, and thus motivated a more bottom-up tactic.

Each ripple was presented at least 16 times (8 times for humans), and a total of >16.000 (monkey; human: >6.000) trials were collected for each subject in about 20 sessions.

Audiogram measurements

Standard free-field pure tone audiograms were determined for all subjects, for frequencies ranging from 250 to 32,000 Hz. We presented all stimuli in the free field at the frontal central position by a speaker (Pioneer TSE1702i) with a flat frequency characteristic within 5 dB between 0.1 and 50 kHz after equalization (Behringer Ultra-Curve). Tones were generated online, using TDT 3 hardware.

Loudness started at 65 dB and was adjusted according to a 2-up 3-down staircase paradigm, which converges on 60% hits / (hits + misses). This was done in steps of 10 dB until the fourth (for humans until the second) reversal, followed by steps of 2 dB. Each tracking ended after 11 reversals for humans, and for monkeys after at least 13 reversals, as soon as the average amplitude of the last 4 reversals and that of the previous 4 was stable within 2 dB. Frequencies were presented in a random order (see Fig. 5 for an example of an experimental block).

Subjects pressed a handle until they detected a tone. We varied the inter-stimulus delay between 500 and 2300 ms, and the tone lasted for 600 ms. Monkeys received 35% catch trials (trials with a delay of 3000 to 3100 ms, followed by a tone with clearly audible amplitude) to measure guess and lapse rates, for which they could also earn small water rewards. Humans received 5% catch trials. For monkeys all measurements were repeated at least twice for each threshold, to assess the reproducibility and reliability of the obtained thresholds.

Data analysis

For each ripple *(ω,Ω)* we fitted a psychophysical curve (Weibull distribution, with lapse rate as free parameter) using the constrained maximum-likelihood algorithm from Wichmann and Hill (Wichmann and Hill, 2001b); as presented in Fig. 6D,E. Guess

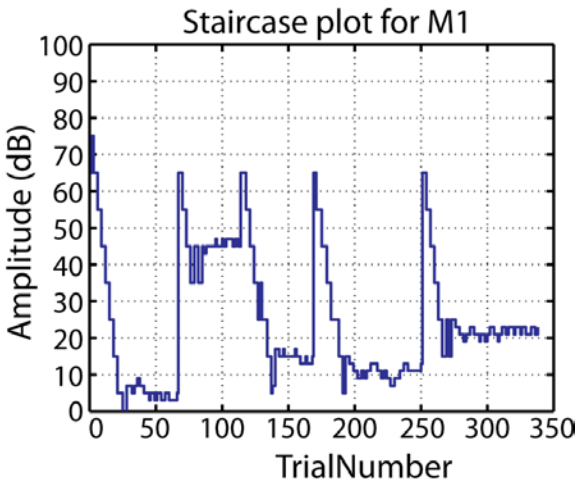

Figure 5. Staircase plot for a typical block of audiogram measurements. Amplitude of the tone (y-axis) is changed after every 3 correct detections or 2 misses. After a threshold for one tone is determined the next is presented at 65 dB. Note that thresholds are stable, indicating the monkey (M1 in this case) is under stimulus control.

rates were determined per subject from the hit/miss ratio in catch trials (undetectable trials with ω=0, Ω=0) and the guess rate (γ) was fixed at this level. The psychometric curve describes the percentage of correct responses (i.e. defined by a reaction time between 220 – 800 ms) as function of ΔM. From the psychophysical curve we determined the detection threshold (ΔM values at 50% correct after correction for the guess and lapse rate of each subject) and its confidence intervals (Wichmann and Hill, 2001b, a). In this way we obtained an 8x11 matrix that contained the thresholds of all different ripple velocities and densities. This matrix is hence on referred to as the psychophysical spectral-temporal modulation transfer function, ST-MTF. From these ST-MTFs we calculated several statistics that are described in the next paragraphs.

Spectral-temporal separability

Singular value decomposition (SVD) was performed to assess separability between the temporal and spectral dimensions of the ST-MTF. The ST-MTF can be written as:

(3) $\text{ST-MTF}(\omega,\Omega) = G(\omega)*K(\lambda)*H(\Omega)$

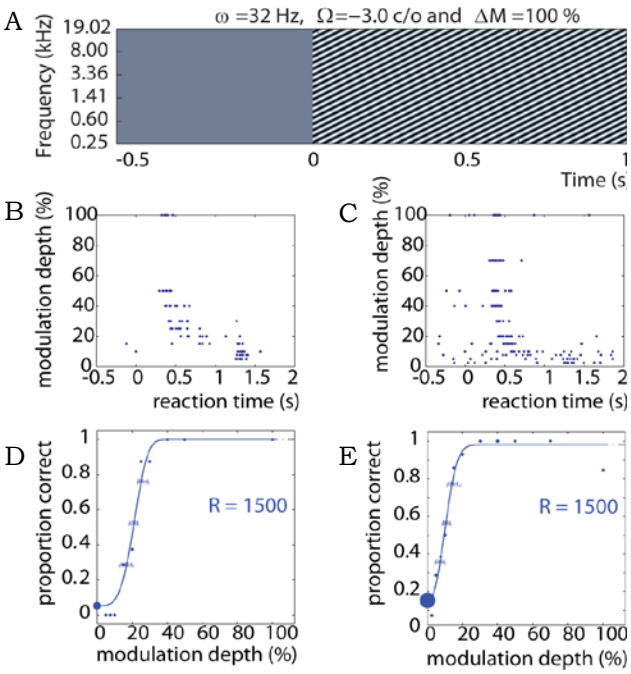

Figure 6. Threshold determination for each ripple (ω,Ω). A. Noise stimulus rippled at t=0 (after 1 to 4s).

B, C. Responses delay between ripple onset and release of a handle (dots) for a human (B, MH) and monkey (C, M1) subject. Reaction times (x axis) were recorded for several modulation depths (y axis). A response window (220-800 ms) was used to determine correct detections.

D,E. The ratio of correct detections over misses (slower or no reactions) was calculated per subject (D: MH, E: M1) for each modulation depth, and a psychometric curve was fitted. Threshold was determined as the modulation depth where the subject had 50% correct detections, after correction for guess rate ('correct' detections when no ripple was present) and lapse rate (proportion of misses for very simple stimuli (ΔM=100%).

with K(λ) a matrix with eigenvalues λ_i on its diagonal, and zeros elsewhere. G(ω) is the time-dependent part of the ST-MTF, while H(Ω) is the frequency-dependent part.

If the ST-MTF is separable, only the first eigenvalue differs from zero (λ_1 / λ_{ALL} = 1). This is quantified by the so-called inseparability index, α, which is defined as:

(4) $\alpha = 1 - \lambda_1^2 / \Sigma \lambda_n^2$

with summation over n=1 to n=8 (the number of velocities).

If α is zero, the power in the ST-MTF is only determined by the first eigenvalue and the function is separable; if it's significantly different from zero the ST-MTF is considered to be inseparable. In this case, the temporal and spectral representations interact. The critical a was determined by simulations and found to be α>0.05, which corresponds to p<0.01. In each iteration of the simulation a subject was chosen randomly and his thresholds were randomly redistributed over the ST-MTF using the permute function of Matlab. SVD analysis was performed on these scrambled ST-MTFs. This was repeated 10.000 times, and the resulting distribution of a values was used to determine significance levels.

In addition, we used the function

(5) G(ω)*K(λ)*H(Ω) = ST-MTF(ω,Ω),

to reconstruct the ST-MTF using only the first eigenvalue in K(λ) non-zero (i.e. under the assumption of full separability). Pearson's correlation between the estimated and the experimentally measured ST-MTF was calculated, and used as an additional measure of separability.

Sensitivity to upward and downward ripples

To test whether the direction of the dynamic modulation (i.e. upward or downward) affects sensitivity, we calculated Pearson's correlation between thresholds obtained for upward (-Ω) and downward (+Ω) ripples. In addition, we calculated the difference between upward and downward movement per ripple (ω,Ω), to see if there were any systematic asymmetries.

Comparison between human and monkey ST-MTFs

To compare human and monkey sensitivity profiles, we first normalised the ST-MTF of each subject based on its best and worst performance. To compare differences between the human and macaque ST-MTF, we adapted a method for a comparable analysis in EEG and MEG data (i.e. local differences between test conditions) (Maris and Oostenveld, 2007). As this analysis is not standard for psycho-acoustics, we will describe this method in the results section.

Results

To distinguish whether the auditory

system analyses sound with a temporal filter bank, or via a combined spectral-temporal filter bank, we assessed the behavioural thresholds for detection in noise of 88 spectral-temporal amplitude modulated stimuli (ripples). For each subject we fitted a psychometrical curve to each ripple (ω,Ω), with modulation depth (ΔM) as independent variable, and collected the 50%-thresholds in a spectral-temporal modulation transfer function (ST-MTF) (Fig. 7).

Spectral-temporal separability

Separability between the spectral and temporal axes of the ST-MTFs was assessed using SVD analysis, and is quantified by the separability index α (threshold at α>0.05, corresponding to p<0.01). α indicates the fraction of the power in the ST-MTF which is accounted for given the assumption of separability (see methods). The critical α values were obtained via simulations. None of the α-values (reported in Table 1) reached significance, meaning that all measured ST-MTFs were separable. Additionally, ST-MTFs were reconstructed under the assumption of full separability and correlated with the experimentally determined ST-MTFs. In all subjects Pearson's correlation *r* was high.

The SVD also returns the pure temporal, $G_1(\omega)$, and pure spectral, $H_1(\Omega)$, modulation functions. In Fig. 8A (human) and B (monkey) we plotted the temporal modulation function that was simulated under conditions of full

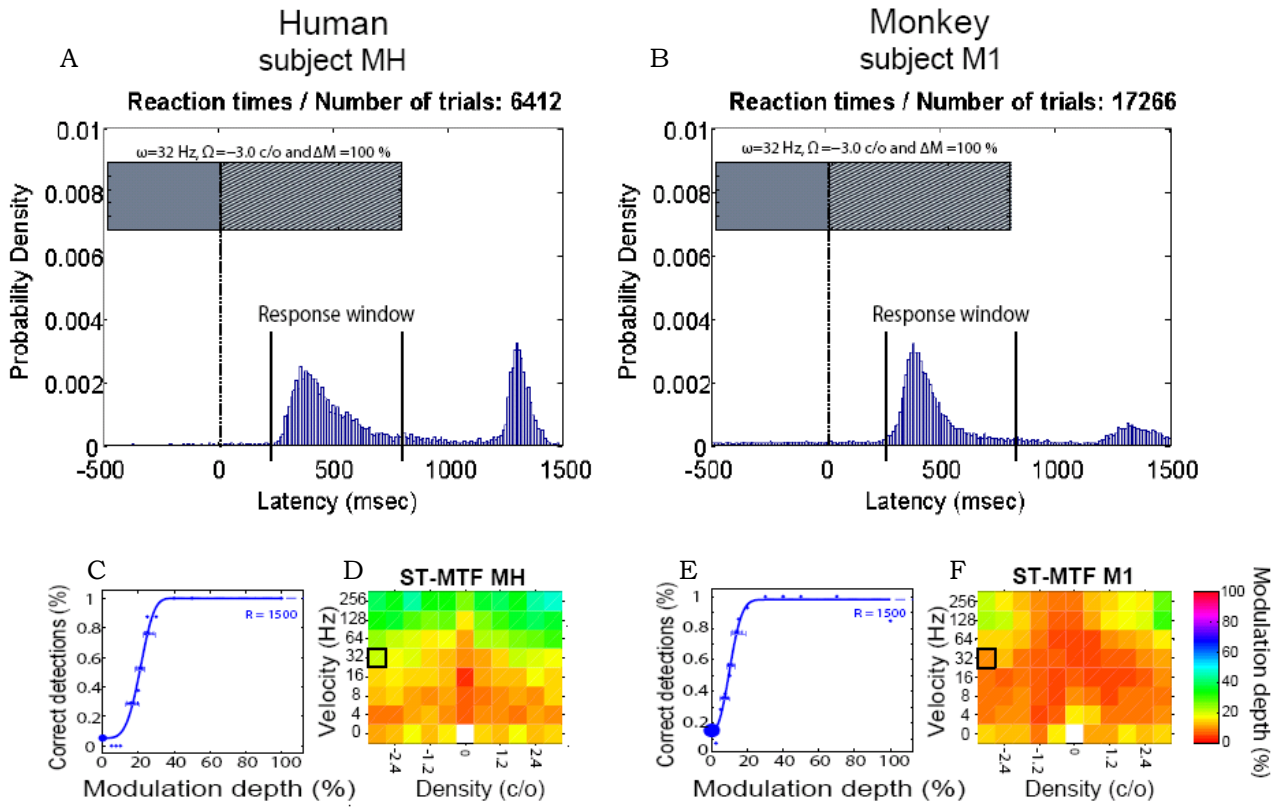

Figure 7. Illustration of the analysis of raw response data to determine the Psychophysical Spectra-Temporal Modulation Transfer Function (MTF) for one human subject (left) and one monkey (right). *A.* Distribution of all reaction times relative to the onset of the stimuli (pooled for all ripples tested). Note that the first reaction peak falls in the designated response window and corresponds to correct detections (hits); later responses are classified as misses and earlier responses are discarded. Note that the second peak reflects the responses to the end of the sound stimulus. *B.* Psychophysical curve fitted to the proportion of hits plotted against modulation depth for a ripple stimulus with ω=32 Hz and Ω=-3.0 c/o. The guess rate is estimated from the percentage "hits" in unmodulated catch trials. The resulting threshold value for this ripple is colour-coded and plotted in *C.* (the marked box), together with threshold values obtained for all other 87 ripples. White box: stimulus at (0,0), used to determine guess rates.

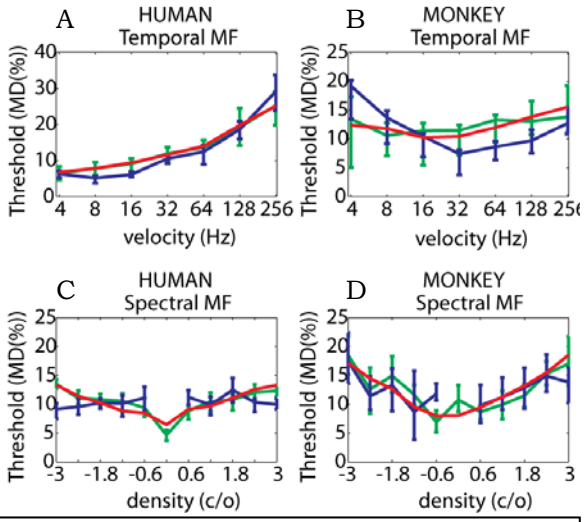

Figure 8. Pure temporal (A,B) and pure spectral (C,D) modulation functions for all subjects (humans: A,C; monkeys: B,D). The red line depicts the pure modulation function as predicted by the SVD analysis, whereas the blue line depicts the measured pure modulation functions for zero density (A,B) or zero velocity (C,D) respectively. The green line depicts the measured temporal (spectral) modulation function for a nonzero density (velocity). Error bars depict 95% confidence intervals.

HUMAN			MONKEY		
	a	*r*		*a*	*r*
RW	0.020	0.96	M1	0.038	0.88
AF	0.018	0.96	M2	0.030	0.90
MH	0.014	0.97	M3	0.030	0.90
MM	0.017	0.97	M4	0.014	0.93
HV	0.019	0.93	M5	0.013	0.94
All	**0.010**	**0.98**	**All**	**0.023**	**0.93**

Table 1. Separability measures. The separability index *a* indicates spectral-temporal inseparability above 0.05. Note that all *a* levels are well below this threshold. The correlation coefficient (*r*) shows the correlation between the simulated ST-MTF given full separability and the experimentally measured ST-MTF. Note the high correlation for all subjects, indicating a high degree of spectral-temporal separability in both humans and monkeys. The bold row gives *a* and *r* for the grouped data (per species).

separability (red) together with the measured data at zero density (blue), and the scaled version of measured data at a nonzero density (green), for comparison. Scaling was performed such that given full separability, all functions would be identical. In Fig. 8C (human) and D (monkey) we plotted the spectral modulation function that was simulated under conditions of full separability (red) together with the measured spectral modulation functions for a zero (blue) and scaled nonzero velocity (green). We plotted both measured data at zero density (or velocity, respectively) and nonzero density (velocity), since the temporal modulation filter bank model of Dau and colleagues does not predict responses to stimuli with zero velocity (Dau et al., 1997; Jepsen et al., 2008).

As can be seen, the simulated functions predict all nonzero modulation functions, as the simulated function falls well within all error bars (95% confidence intervals). However, the modulation functions at zero velocity and density may deviate somewhat, especially for the temporal modulations in monkey and the spectral modulations in human data.

Symmetry in the ST-MTF

If the auditory system processes temporal and spectral properties separately, sensitivity is determined by independent spectral and temporal parameters. The sign of the ripple causes a 180° phase shift of the ripple in both spectral and temporal dimensions. Given separability the parameters that determine sensitivity are thought to be independent of any phase shift. This means that the auditory system is equally sensitive to upward and downward moving spectral-temporal modulations, if it processes sound in a separable way.

To test for this we correlated responses to upward moving ripples with those to downward moving ripples. The result pooled over all subjects is plotted in Fig. 9. Each pair of ripples had the same parameters, with the exception of the sign of the density. However, note that not only the direction of the slope varied with the sign of the density, but also that the amplitude as a function of spectrum is inversed at onset of the ripple. Correlation between upward and downward ripples was $r = 0.93$.

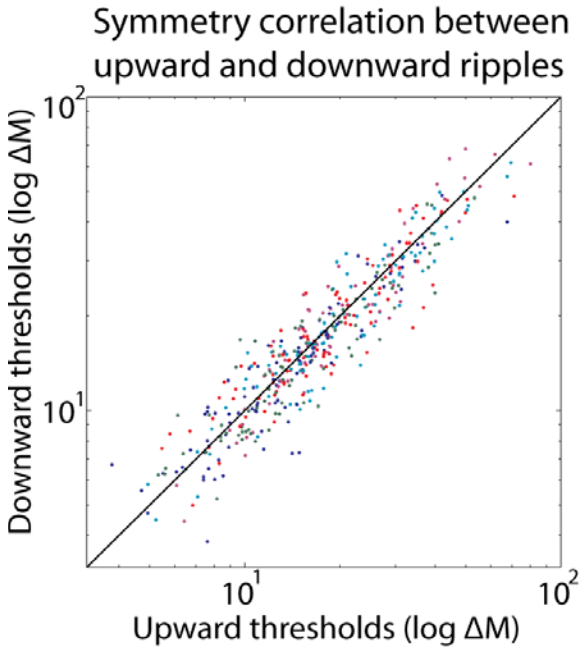

Figure 9. A. Correlation between upward and downward moving ripples for all pooled subjects. Note the good correspondence with the black y=x line, indicating that the auditory system is equally sensitive for upward and downward ripples. Dots of each velocity are colour coded. Note the random distribution of all colours around the black line, indicating that there are no systematic deviations

To elucidate any localised asymmetries we calculated the difference between each upward and downward moving ripple pair, normalised by dividing each difference by the average 95%-confidence interval size of both ripples. The results were averaged over all subjects per species. Any systematic differences can be seen as clusters of higher (red) or lower (blue) sensitivity to upward ripples (see supplemental Fig. S1). However, no such systematic differences were found.

Human versus macaque sensitivities

To qualitatively compare the ST-MTFs of our human subjects to the monkeys' ST-MTFs, each subject's ST-MTF was normalised to his minimal and maximal sensitivity. These normalised ST-MTFs were averaged across subjects to obtain a species-specific ST-MTF (Fig. 10). Pearson's correlation between these human and monkey ST-MTFs gave $r = 0.69$.

For a more localised comparison, we adapted a method to test for localised activity differences between different conditions in EEG and MEG data (Maris and Oostenveld, 2007). In short, this method consists of three

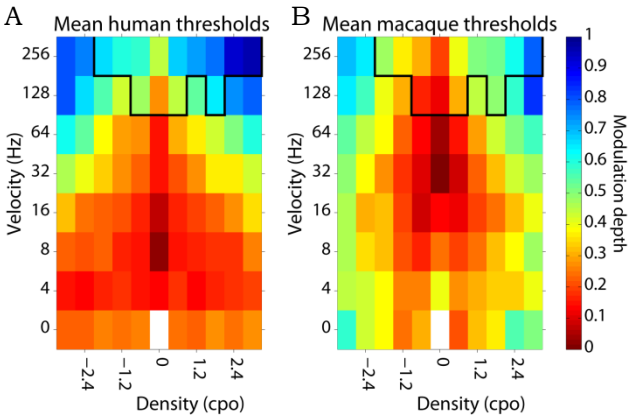

Figure 10. Normalised ST-MTFs for all humans (left) and all monkeys (right). Note the difference between humans and monkeys at high ripple velocities, and the upward shift in optimal performance in the monkeys. The demarked area marks a region where macaques performed significantly ($p<0.01$) better. Thresholds are colour coded, with dark red corresponding to best performance. Thresholds were normalised for each subject to their top and worst performance before averaging.

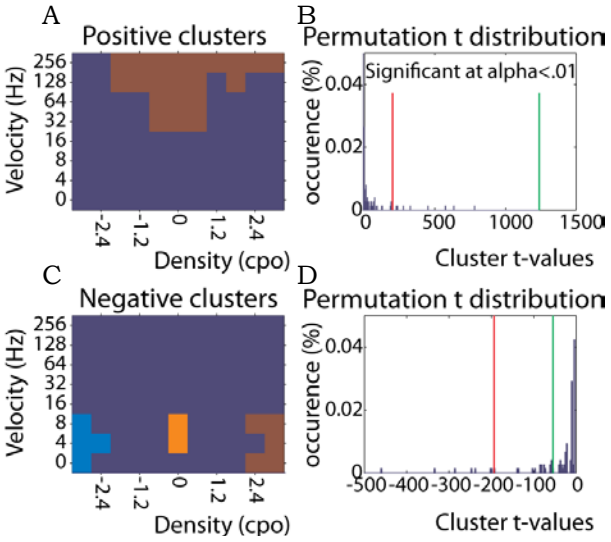

Fig 11. Clusters of thresholded t-values for qualitative comparison between monkey and human sensitivity. A. The clusters indicate regions where monkeys may be more sensitive; whereas clusters where humans may be more sensitive are shown in C. Significance was tested via permutation and bootstrapping. B and D show the resulting distributions of cluster-t-values. The red line indicates the critical cluster-t value at $p<0.05$. The green line indicates the cluster-t-value for the largest cluster in A and C, respectively.

steps: First, a t-test for each pixel (velocity-density pair) was performed, comparing the human (N=5) with the macaque (N=5) thresholds. Second we thresholded these t-values at t>2 (or t<-2) and clustered adjacent pixels that were above threshold (t-values >2.0 or <-2.0). Third, for each cluster we calculated a cluster-t-statistic, meaning that we added the individual t-values in the cluster. In the case that multiple clusters were present, we discarded all but the largest cluster-t-statistic.

We tested the significance of this cluster-t-statistic via a permutation test: we grouped all subjects, monkey and human, and assigned each subject randomly to either the human or macaque test group. For all possible subject combinations (10!/(5!*5!)=252), we calculated the cluster-t-statistic as described above. From the distribution of all the obtained cluster-t-values (Fig. 11B, D), we could determine the critical cluster-t-value at p<0.05 and p<0.01, enabling us to assess the significance of the cluster-t-values measured in the data. This method allows strong control over the critical parameter, for details see (Maris and Oostenveld, 2007; Maris et al., 2007). The location and size of significant clusters provide information on the type of ripples to which macaques are more (or less) sensitive than humans (Fig. 11A, C).

We found one significant (p<0.01) cluster at ripple velocities above 128 Hz, where macaques perform better than humans. This area is demarked in Fig. 10. Note that although monkeys may appear to be less sensitive for ripples with low velocities and high densities, this was found to be non-significant (p=0.13 (left), and p=0.12 (right)).

Pure-tone audiograms

Audiogram measurements showed that all subjects had hearing within the normal range for their species (see Fig. 12). This means that monkeys remained sensitive up to frequencies of 32 kHz and were somewhat less sensitive to low frequencies (<750Hz) than humans. No audiograms were measured for monkeys M4 and M5.

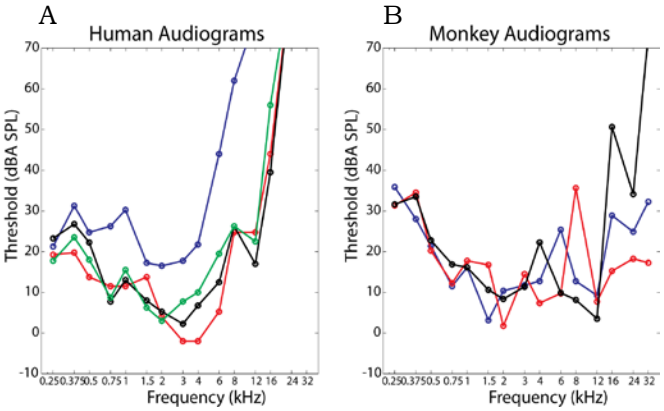

Figure 12. Free-field audiograms for human (left) and monkey (right) subjects. Note that monkeys remain sensitive up to much higher tone frequencies than humans. Frequency is plotted on a log scale.

Discussion

To distinguish whether the auditory system makes use of separate spectral and temporal filter banks, or uses combined spectrotemporal filters, we measured behavioural thresholds to spectrotemporal ripples. The model using separate filter banks makes strong predictions at a behavioural level: i) The spectral-temporal transfer function (ST-MTF) will be separable ii) The auditory system will be equally sensitive to upward and downward ripples. iii) The ST-MTF will be similar across species (with possible exception of temporal cut-off). iv) Last, the ST-MTF will have a low-pass shape for dynamic spectral-temporal ripples. None of these conditions are expected to be met if the auditory system processes sound in an inseparable way.

Our results show that all these predictions are met: we will discuss our findings in relation to these predictions and the literature, for each prediction separately.

Spectral-temporal separability

The small separability indices and the high correlation between the separated transfer function and the data indicate that the ST-MTF is separable in both species. This is consistent with human data from (Chi et al., 1999), who performed SVD analysis on grouped human sensitivity to ripples with a narrower temporal and broader spectral range. Their results too showed separability of the ST-MTF ($\alpha_2 = 1 - \lambda_1 / \lambda_{all} < .15$), and although they did not calculate the r directly, they found that simulated data, given the assumption of full separability, lead to a difference of 3.4% with their measured data (least mean squares). Thus, the correlation in their data was $r = 0.97$, which is comparable to our data.

Moreover, we found that the temporal and

spectral modulation functions extracted from the SVD analysis yielded a good correspondence for the entire data set (Fig. 8, red and green lines, respectively). However, there was less similarity between the retrieved modulation functions and the pure spectral (at velocity zero) and pure temporal (at density zero) amplitude modulations. This suggests that measuring only pure temporal and pure spectral modulations, respectively, is not the best approach to retrieve the full spectrotemporal sensitivity characteristic. This is not unexpected given Dau's model, as spectral interference on the temporal filters does not reflect the capacity of the auditory system to detect spectral patterns.

The deviations in the pure temporal and spectral thresholds were similar for all members in each species, suggesting the higher divergence of the human zero velocity curve in humans and the monkey zero density curve were systematic. These results lead us to suspect that humans and monkeys involuntarily focussed on different aspects (i.e. temporal and spectral, respectively) of the ripples, causing their performance for dynamic spectral-temporal ripples to primarily follow the sensitivity function for the preferred aspect.

Symmetry in the ST-MTF

If the auditory system processes time and spectrum separately, it should be equally sensitive to upward and downward moving ripples. The direction of the ripple is determined by its sign (+ for downward ripples, – for upward ripples). This sign causes an 180̊ phase shift in both spectrum and time. A phase shift at ripple onset is not thought to affect sensitivity, since the ripple is faded in. Also during the remainder of the stimulus it is thought to be undetectable, since a separable system cannot encode phase information. This loss of relative phase information is due to the temporal filters, which presumably convert temporal phase information into a firing rate code (Dicke et al., 2007; Wang et al., 2008).

The expectation of equal sensitivity to upward and downward ripples was confirmed by the high correlation (r = .93) between these ripples (Fig. 9). Moreover, no systematic differences between sensitivity to upward and downward ripples were found. This is in accordance with findings from Osmanski and colleagues, who tested a smaller set of ripples, and al found equal sensitivity to upward and downward ripples in humans (Osmanski et al., 2009).

In conclusion, the symmetry between upward and downward moving ripples in our human and monkey subjects is consistent with our previous finding of separability of the ST-MTF: both suggest that spectral and temporal modulations are processed independently.

Similarity of the ST-MTFs across species

We found a close correlation between the ST-MTFs of monkeys and humans (r = 0.69). Only one area was found to differ systematically ($p<0.01$, Fig. 10). This area approximately spanned the entire spectral range for velocities above 128 Hz. This difference thus reflects a difference in overall temporal sensitivity, which could be due to physiological differences.

In contrast some clusters at low velocities appeared different at first sight, although significance was not reached ($p=0.12$ and 0.13, Fig. 11A,C). The between-subject variability in those areas was rather large in both species, which could be an artefact of the normalisation. This is supported by the similarity between both species of the raw data in these areas (Fig. 7D,F).

As reported previously, we found that sensitivity of our human subjects followed their pure temporal sensitivity, whereas sensitivity of our monkeys followed their pure spectral sensitivity. The spectral sensitivity function of monkeys was comparable with that of our human subjects. This contrasts with results from O´Connor and colleagues (2000), who compared sensitivity to pure spectral and temporal modulations between human and macaque subjects. They found very poor sensitivity to all pure spectral modulations. Furthermore, they found comparable temporal sensitivity functions for human and macaque subjects, with monkeys performing slightly worse for modulation frequencies from 15 to 250 Hz and slightly better for very high modulation frequencies (up to 2 kHz). Although the performance of their monkey data is very similar to our results, our human subjects showed reduced sensitivity for velocities above about 128 Hz. This early temporal decay in human subjects confirms other human data, including a systematic study on temporal modulation detection by (Kohlrausch et al., 2000). Also see (Ewert and Dau, 2000).

The difference in the reported temporal

decay within our study and across studies may have been caused partly by the spectral range of the stimuli used. Macaques remain highly sensitive to well above the maximum frequency present in our stimuli (sensitivity was found to remain high up to 30 kHz), whereas humans show a decline in sensitivity starting from about 10 kHz. Because our stimuli consisted of carrier frequencies from 0.25 to 20 kHz, monkeys had more frequency channels available for signal analysis, which may have had a positive effect on their signal to noise ratio, which in turn may have made the analysis of higher temporal frequencies relatively easier. This is also supported by findings from Dau and colleagues (1997), who found that performance on temporal amplitude modulation detection increases with increasing carrier bandwidth. This suggests that information on modulation rates is integrated over frequencies, and that, since monkeys had more frequencies available, this caused an overall increase in their performance, most notably at high temporal modulation rates. This is also in accordance with a study from He and colleagues (He et al., 2008), who found that subjects with an overall reduced hearing are specifically worse at detecting temporal modulations between 40 and 200 Hz. Thus the difference in temporal cut-off between our human and macaque subjects was probably due at least partly to stimulus properties.

Our observation that no qualitative differences across species were present, is supported by recent work from Osmanski and colleagues (2009), who compared performance on a small selection of ripples in two bird species (zebrafinches and budgerigars) and humans. In accordance with our findings, they too found no differences in performance across species.

Low-pass filter shaped spectral sensitivity

For dynamic ripples, performance is expected to deteriorate with increasing density, given separable spectral-temporal processing, as suggested by our previous findings. This sensitivity decay with increasing density is caused by a reduction in the number of subsequent frequency channels that fire synchronously. Similarly, thresholds are expected to become elevated with increasing temporal frequency, resulting in an ST-MTF with an overall low-pass shape.

Our results show both optimal performance at zero density, and a decrease in performance with increasing density. This is in accordance with the predictions from a separate spectral and temporal filter bank and it supports the findings that the input to the temporal filter bank is indeed integrated over all frequencies (Dau et al., 1997).

Conclusions: separable encoding of FM-sweeps

In summary, our results strongly support independent analysis of spectral and temporal characteristics by the auditory system. Our results confirmed all predictions that ensue from this spectral-temporal model, shown in Fig. 3 and (Dicke et al., 2007; Jepsen et al., 2008): i) The spectral-temporal transfer function (ST-MTF) was found to be separable ii) The auditory system was found to be equally sensitive to most upward and downward ripples. iii) The ST-MTF was similar across species. iv) And last, sensitivity to dynamic ripples had a low-pass shape, as predicted by Dau's model.

However, Dau's separable spectral-temporal model leaves open the issue how the auditory system deals with FM-sweeps, which are inseparable. The lack of across-frequency phase information and the lack of any other differences between upward and downward moving ripples pose a paradox: it predicts that a separable system cannot distinguish between upward and downward moving ripples. This contrasts with our everyday experience, as we are quite capable to discriminate between upward and downward moving FM-sweeps. Furthermore, we seem quite able to process FM-sweeps in general.

The importance of FM-sweeps for speech processing was pointed out earlier, and is well illustrated by a study from Fu and colleagues (Fu et al., 1998). They found that in tonal languages, a frequency shift, most notable in the fundamental frequency (F0), may determine the meaning of a phoneme. For example, the syllable /ma/ can mean "mother", "linen", "scold" or "horse", dependent on whether the F0 is flat, rises, falls or falls and then rises, respectively (Fu et al., 1998).

Thus the question arises how a separable auditory system can deal with this inseparable acoustic feature?

Model of spectral-temporal processing

It is generally assumed that only inseparable systems can deal with inseparable acoustic features. However, if one

looks at the SVD of an FM-sweep, one sees that it can be fully reconstructed by the sum of two separable functions (Fig. 13)[2]. If the auditory system uses summation of two separate channels to detect FM-sweeps, the auditory system would be able to process FM-sweeps, and use them as cues for auditory streaming. At the same time, sensitivity to FM-sweeps would still be determined by the pure temporal and pure spectral sensitivity functions.

To achieve this, the model from Dau and colleagues needs two adjustments. First, the temporal filter bank can be simplified by applying temporal filters not to each frequency separately, but to band-pass shaped frequency filters. The adjusted filter bank consists of overlapping filters with different time constants and centre frequencies (example filters are given in Fig. 13B-E). Note that the difference in the time constants of the filters is essential for detecting sweeps and pure temporal modulations with different speeds. By the use of inhibitory connections, only neurons that are sensitive to positive amplitude at t=0 are needed.

Second, a subsequent layer should be added with neurons that are sensitive to certain combinations of active and suppressed separable neurons which together form a certain FM-sweep. For example, a neuron that would be sensitive to the sweep in Fig. 13A would need input from four neurons with overlapping band-pass filters, and regularly increasing centre frequencies (Fig 13B-E, respectively).

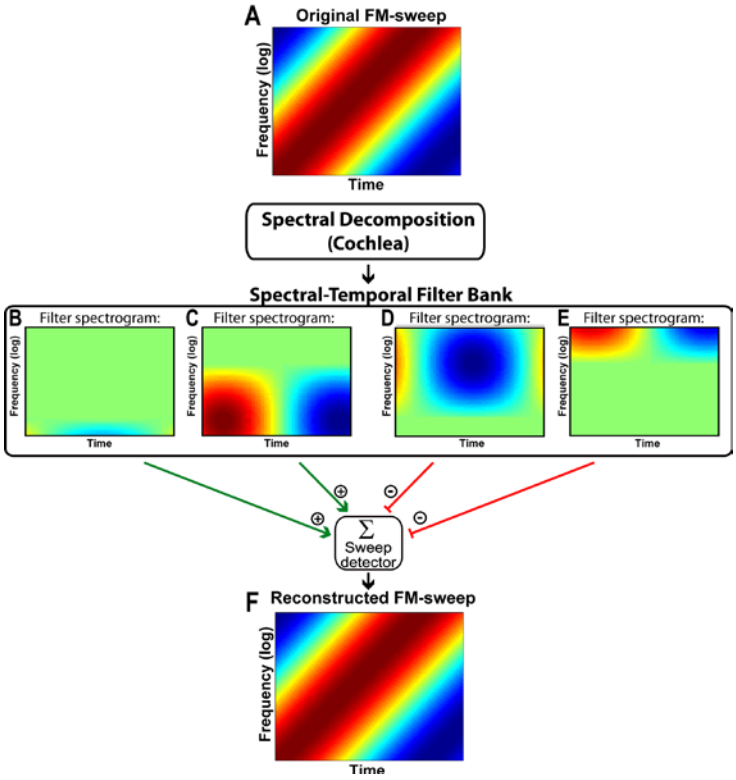

Figure 13. Proposed separable spectral-temporal filter bank model. Example simple input (A) and output (F) are shown to illustrate the ability of the auditory system to process FM-sweeps using a separable filter bank.

After spectral decomposition sound is processed in a filter bank with different temporal decay constants (to process different speeds), e.g. B and C, and different center frequencies (B-E). Note that the frequency bands overlap, such that they cover the entire frequency range twice with regularly spaced center frequencies (B-E).

Thus, although the auditory system processes sound in a separable way at an early level, spectrotemporal filters are still present in higher areas. These filters do not affect psychophysical detection thresholds, as detection of ripples is already determined by the separable spectral-temporal filters. This model is the only one that allows spectrotemporal modulations like FM-sweeps to be used for diverse tasks, such as speech perception in noise and vowel classification, despite of the loss of across-spectral phase information which is due to a separable system.

Although it may seem wasteful to have both separate spectral and temporal filter banks and subsequent spectrotemporal filters, it should be kept in mind that a very small collection of filters to FM-sweeps may suffice for everyday purposes. In principle only those detectors are needed that encode behaviourally relevant FM-sweeps. Results from Malayath and Hermansky (2003) suggest the number of behaviourally relevant FM-sweeps may be limited. Using data-driven feature extraction they derived optimal filters for automated speech recognition. They found four significant discriminants, among which two that focused on specific ripples in the

[2] This can also be deduced from the formula of FM-sweeps $A=\cos(x/\omega + y/\Omega)$, which can be rewritten as the sum of two separable functions using the cosine rule.

central part of the critical band spectrum. As this approach was data-driven using a large set of speech data, their results suggest that to use FM-sweeps in speech only a very limited number (i.e. 2 in their case) of filters is needed.

Another advantage of our model is that the separable filter banks allow the spectrotemporal filters to be highly selective and adaptive to behavioural needs, without interfering with overall spectral and temporal processing and sensitivity.

Moreover, covering the entire frequency range with overlapping filters maximises information and reliability, while minimising coding costs. A similar strategy has been found in the visual system: using synchronous spiking, the receptive field size of subsequent layers can have a higher resolution than the receptive field sizes in the filter bank (Pirenne and Denton, 1952). Thus signal to noise ratio is maximised, while information flow is highly compressed.

Comparison to the visual system

In the previous paragraph we compared the distribution of auditory spectral-temporal receptive fields to that of visual spatial receptive fields in the context of coding efficiency. The visual correlate of spectral-temporal modulations, however, has been argued to be spatiotemporal modulations (Chi et al., 1999). Measurements of visual spatiotemporal MTFs using sinusoidally modulated gratings with various orientations and drifting velocities showed that in contrast with the auditory system, the visual system does process time and space in an inseparable way (Kelly, 1961; Dong and Atick, 1995). However, spatiotemporal modulations differ from spectral-temporal modulations in one important aspect: space has at least one extra dimension compared to frequency. This higher dimensionality may introduce additional complexity, such that two or three separable functions are hardly ever enough to fully reconstruct a visual structure.

Acknowledgements

First of all, I would like to thank my supervisors, Rob and John, for their support and guidance. Thanks to them I have learned very much during the last year, which makes me very happy. Also I would like to thank all subjects for their perseverance and pleasant attitude, without which this study would not have been possible.

References

Aubin T, Jouventin P (1998) Cocktail-party effect in king penguin colonies. P Roy Soc Lond B Bio 265:1665-1673.

Bregman AS (1990) Auditory scene analysis : the perceptual organization of sound. Cambridge, Mass.: MIT Press.

Bregman AS, Darwin C, Abramson J (1982) Spectral Integration of Sounds Based on Common Amplitude-Modulation. Bulletin of the Psychonomic Society 20:133-133.

Brunstrom JM, Roberts B (2000) Separate mechanisms govern the selection of spectral components for perceptual fusion and for the computation of global pitch. Journal of the Acoustical Society of America 107:1566-1577.

Cherry C (1957) On human communication: a review, a survey, and a criticism. [Cambridge]: Technology Press of Massachusetts Institute of Technology.

Chi T, Ru P, Shamma SA (2005) Multiresolution spectrotemporal analysis of complex sounds. J Acoust Soc Am 118:887-906.

Chi TS, Gao YJ, Guyton MC, Ru PW, Shamma S (1999) Spectro-temporal modulation transfer functions and speech intelligibility. Journal of the Acoustical Society of America 106:2719-2732.

Ciocca V, Darwin CJ (1993) Effects of Onset Asynchrony on Pitch Perception - Adaptation or Grouping. Journal of the Acoustical Society of America 93:2870-2878.

Darwin CJ (1981) Perceptual Grouping of Speech Components Differing in Fundamental-Frequency and Onset-Time. Quarterly Journal of Experimental Psychology Section a-Human Experimental Psychology 33:185-207.

Darwin CJ (2008) Listening to speech in the presence of other sounds. Philosophical Transactions of the Royal Society B-Biological Sciences 363:1011-1021.

Darwin CJ, Ciocca V (1992) Grouping in Pitch Perception - Effects of Onset Asynchrony and Ear of Presentation of a Mistuned Component. Journal of the Acoustical Society of America 91:3381-3390.

Darwin CJ, Hukin RW (1998) Perceptual segregation of a harmonic from a vowel by interaural time difference in conjunction

with mistuning and onset asynchrony. Journal of the Acoustical Society of America 103:1080-1084.

Dau T, Kollmeier B, Kohlrausch A (1997) Modeling auditory processing of amplitude modulation .2. Spectral and temporal integration. Journal of the Acoustical Society of America 102:2906-2919.

Depireux DA, Simon JZ, Klein DJ, Shamma SA (2001) Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. Journal of Neurophysiology 85:1220-1234.

Dicke U, Ewert SD, Dau T (2007) A neural circuit transforming temporal periodicity information into a rate-based representation in the mammalian auditory system. Journal of the Acoustical Society of America 121:310-326.

Dong DW, Atick JJ (1995) Statistics of Natural Time-Varying Images. Network-Computation in Neural Systems 6:345-358.

Elhilali M, Shamma SA (2008) A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation. Journal of the Acoustical Society of America 124:3751-3771.

Eramudugolla R, McAnally KI, Martin RL, Irvine DRF, Mattingley JB (2008) The role of spatial location in auditory search. Hearing Research 238:139-146.

Ewert SD, Dau T (2000) Characterizing frequency selectivity for envelope fluctuations. Journal of the Acoustical Society of America 108:1181-1196.

Felsheim C, Ostwald J (1996) Responses to exponential frequency modulations in the rat inferior colliculus. Hearing Research 98:137-151.

Fu QJ, Zeng FG, Shannon RV, Soli SD (1998) Importance of tonal envelope cues in Chinese speech recognition. Journal of the Acoustical Society of America 104:505-510.

Haykin S, Chen Z (2005) The cocktail party problem. Neural Computation 17:1875-1902.

He NJ, Mills JH, Ahlstrom JB, Dubno JR (2008) Age-related differences in the temporal modulation transfer function with pure-tone carriers. Journal of the Acoustical Society of America 124:3841-3849.

Hukin RW, Darwin CJ (1995) Comparison of the Effect of Onset Asynchrony on Auditory Grouping in Pitch Matching and Vowel Identification. Perception & Psychophysics 57:191-196.

Jepsen ML, Ewert SD, Dau T (2008) A computational model of human auditory signal processing and perception. Journal of the Acoustical Society of America 124:422-438.

Joris PX, Schreiner CE, Rees A (2004) Neural processing of amplitude-modulated sounds. Physiological Reviews 84:541-577.

Kelly DH (1961) Visual Responses to Time-Dependent Stimuli .1. Amplitude Sensitivity Measurements. J Opt Soc Am 51:422-&.

Klein DJ, Depireux DA, Simon JZ, Shamma SA (2000) Robust spectrotemporal reverse correlation for the auditory system: Optimizing stimulus design. Journal of Computational Neuroscience 9:85-111.

Kohlrausch A, Fassel R, Dau T (2000) The influence of carrier level and frequency on modulation and beat-detection thresholds for sinusoidal carriers. Journal of the Acoustical Society of America 108:723-734.

Kowalski N, Depireux DA, Shamma SA (1996a) Analysis of dynamic spectra in ferret primary auditory cortex .2. Prediction of unit responses to arbitrary dynamic spectra. Journal of Neurophysiology 76:3524-3534.

Kowalski N, Depireux DA, Shamma SA (1996b) Analysis of dynamic spectra in ferret primary auditory cortex .1. Characteristics of single-unit responses to moving ripple spectra. Journal of Neurophysiology 76:3503-3523.

Langemann U, Zokoll MA, Klump GM (2005) Analysis of spectral shape in the barn owl auditory system. Journal of Comparative Physiology a-Neuroethology Sensory Neural and Behavioral Physiology 191:889-901.

Linden JF, Liu RC, Sahani M, Schreiner CE, Merzenich MM (2003) Spectrotemporal structure of receptive fields in areas AI and AAF of mouse auditory cortex. Journal of Neurophysiology 90:2660-2675.

Malayath N, Hermansky H (2003) Data-driven spectral basis functions for automatic speech recognition. Speech Commun 40:449-466.

Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. J Neurosci Meth 164:177-190.

Maris E, Schoffelen JM, Fries P (2007) Nonparametric statistical testing of coherence differences. J Neurosci Meth 163:161-175.

Nassiri R, Escabi MA (2008) Illusory spectrotemporal ripples created with binaurally correlated noise. Journal of the Acoustical Society of America 123:El92-El98.

O'Connor KN, Barruel P, Sutter ML (2000) Global processing of spectrally complex sounds in macaques (Macaca mullata) and humans. Journal of Comparative Physiology a-Neuroethology Sensory Neural and Behavioral Physiology 186:903-912.

Osmanski MS, Marvit P, Depireux DA, Dooling RJ (2009) Discrimination of auditory gratings in birds. Hear Res.

Pirenne MH, Denton EJ (1952) Accuracy and sensitivity of the human eye. Nature 170:1039-1042.

Roberts B (2005) Spectral pattern, grouping, and the pitches of complex tones and their components. Acta Acustica United with Acustica 91:945-957.

Roberts B, Bregman AS (1991) Effects of the Pattern of Spectral Spacing on the Perceptual Fusion of Harmonics. Journal of the Acoustical Society of America 90:3050-3060.

Roberts B, Bailey PJ (1993) Spectral Pattern and the Perceptual Fusion of Harmonics .1. The Role of Temporal Factors. Journal of the Acoustical Society of America 94:3153-3164.

Roberts B, Bailey PJ (1996) Regularity of spectral pattern and its effects on the perceptual fusion of harmonics. Perception & Psychophysics 58:289-299.

Shamma SA (1996) Auditory cortical representation of complex acoustic spectra as inferred from the ripple analysis method. Network-Computation in Neural Systems 7:439-476.

Versnel H, Zwiers MP, van Opstal AJ (2009) Spectrotemporal response properties of inferior colliculus neurons in alert monkey. Journal of Neuroscience In Press.

Wang X, Lu T, Bendor D, Bartlett E (2008) Neural coding of temporal information in auditory thalamus and cortex. Neuroscience 154:294-303.

Wichmann FA, Hill NJ (2001a) The psychometric function: I. Fitting, sampling, and goodness of fit. Perception & Psychophysics 63:1293-1313.

Wichmann FA, Hill NJ (2001b) The psychometric function: II. Bootstrap-based confidence intervals and sampling. Perception & Psychophysics 63:1314-1329.

Supplements

Symmetry analyses

All systematic differences between upward and downward moving ripples is given (Fig. S1).

Spectral-temporal separability

Fig. S2 shows the distribution of *a* and *r* values as retrieved from SVD simulations. In these simulations the ST-MTFs were scrambled before SVD analysis was done. These scrambled ST-MTFs were more inseparable than our data, indicating, such that less than 1% of the *a*-values were as small as the value that was obtained experimentally.

Fig. S3 shows the pure spectral and temporal MTFs for each subject separately. Red lines denote the simulated pure MTFs, the green line is measured data on spectral-temporal ripples, and the blue line depicts measured data at spectral or temporal modulated data (ω=0 or Ω=0). Error bars depict 95% confidence intervals.

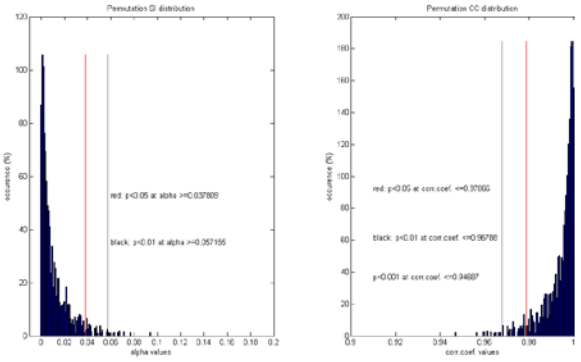

Figure S2. Distribution of separability indices (left) and correlation coefficients (right) that resulted from 50.000 simulations, and the resulting critical values at p<0.05 (red) and p<0.01 (black).

Note that a separability index close to zero and a correlation coefficient close to 1 indicate separability of the underlying MTF.

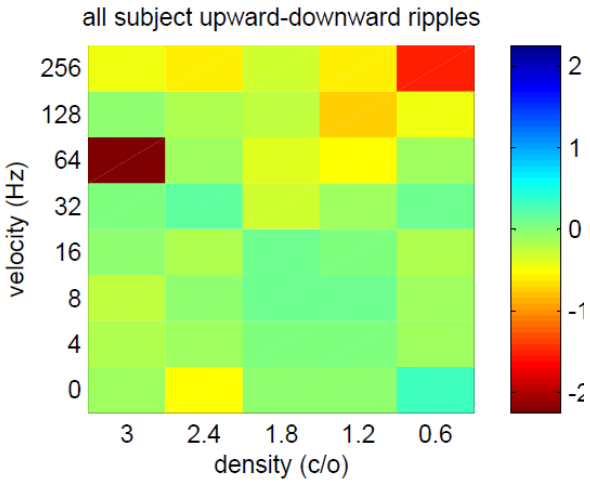

Figure S1. Difference between upward and downward ripples per ripple for all subjects. Before averaging over subjects each difference was normalized by division through the 95% confidence interval size. Note that this means that a difference larger than 0.75 corresponds to a significant difference (p<0.05). However this is only true for 2 out of 40 pixels, i.e. (64,3) and (256,0.6).

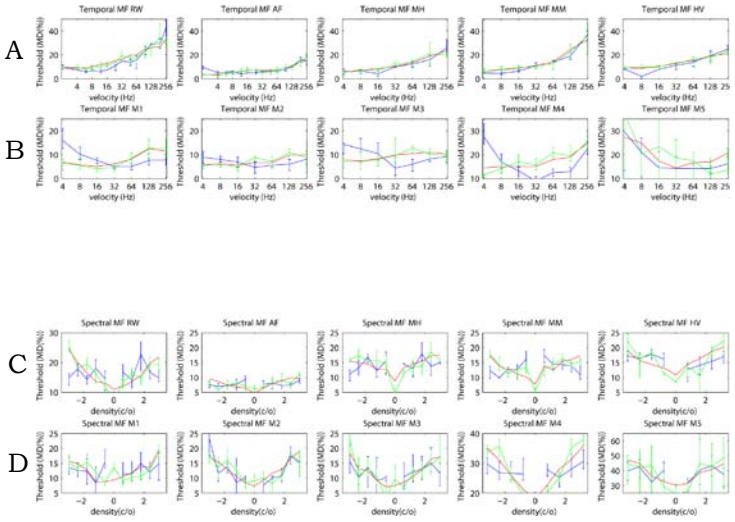

Figure S3. Pure temporal (A,B) and pure spectral (C,D) modulation functions for all subjects (humans: A,C; monkeys: B,D). The red line depicts the pure modulation function as predicted by the SVD analysis, whereas the blue line depicts the measured pure modulation functions for zero density (A,B) or zero velocity (C,D) respectively. The green line depicts the measured temporal (spectral) modulation function for a nonzero density (velocity). Error bars depict 95% confidence intervals.

FM-sweeps

In Fig. S4 the decomposition of an FM-sweep into two separable functions is shown. Note that only two eigenvalues are nonzero.

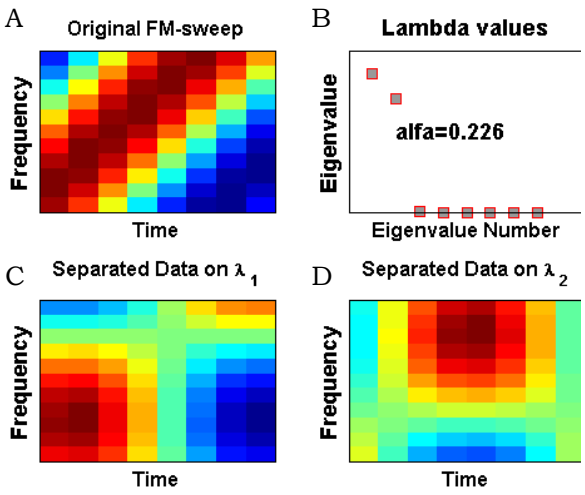

Figure S4. FM-sweeps (A) can be fully reconstructed from the addition of two fully separable functions (C and D). B. FM-sweeps are clearly inseparable sounds, as indicated by the separability index *a*. However, only the first two eigenvalues (λ) are nonzero, whereas all other eigenvalues are zero. This means that the spectrogram of an FM-sweep can be fully reconstructed by the addition of the two separable functions corresponding to these nonzero eigenvalues. These reconstructions based on only the first or second λ nonzero are displayed in C and D, respectively.