Chapter 12
Modulation Representations for Speech and Music

Mounya Elhilali

Abstract The concept of modulation has been ubiquitously linked to the notion of timbre. Modulation describes the variations of an acoustic signal (both spectrally and temporally) that shape how the acoustic energy fluctuates as the signal evolves over time. These fluctuations are largely shaped by the physics of a sound source or acoustic event and, as such, are inextricably reflective of the sound identity or its timbre. How one extracts these variations or modulations remains an open research question. The manifestation of signal variations not only spans the time and frequency axes but also bridges various resolutions in the joint spectrotemporal space. The additional variations driven by linguistic and musical constructs (e.g., semantics, harmony) further compound the complexity of the spectrotemporal space. This chapter examines common techniques that are used to explore the modulation space in such signals, which include signal processing, psychophysics, and neurophysiology. The perceptual and neural interpretations of modulation representations are discussed in the context of biological encoding of sounds in the central auditory system and the psychophysical manifestations of these cues. This chapter enumerates various representations of modulations, including the signal envelope, the modulation spectrum, and spectrotemporal receptive fields. The review also examines the effectiveness of these representations for understanding how sound modulations convey information to the listener about the timbre of a sound and, ultimately, how sound modulations shape the complex perceptual experience evoked by everyday sounds such as speech and music.

Keywords Auditory cortex · Modulation spectrum · Musical signal
Spectrotemporal modulations · Spectrotemporal receptive fields · Speech

M. Elhilali (✉)
Laboratory for Computational Audio Perception, Center for Speech and Language Processing, Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD, USA
e-mail: mounya@jhu.edu

K. Siedenburg et al. (eds.), *Timbre: Acoustics, Perception, and Cognition*,
Springer Handbook of Auditory Research 69,
https://doi.org/10.1007/978-3-030-14832-4_12

12.1 Introduction

If one asks a telecommunication engineer what is “modulation”, the answer is rather simple: It is the process of multiplexing two signals: a signal that can carry information and can be physically transmitted over a communication channel (the *carrier signal*, typically a quickly varying wave) with a signal that contains the information or the message to be transmitted or broadcasted (the *modulation* or *data signal*, typically a slowly varying envelope) (Freeman 2004). This characterization provides a formal account of modulation but fails to capture the nuances of multiplexing two signals that get rather complicated depending on the domain under study. This definition presumes a priori knowledge of the identity, attributes, and behavior of such signals, which is only possible in specific applications (e.g., on/off keying—OOF—used to transmit binary 0/1 codes over a sinusoidal carrier that can be decoded directly from the signal amplitude).

On the flip side, defining modulation as a multiplexing operation is rather ineffective when it comes to the inverse problem: demodulating a signal in order to identify its modulator and carrier components. If one does not have specific constraints on these signal components, it is not trivial to untangle them because many (possibly infinite) solutions are conceivable. How one judges which solution is a reasonable one is again domain and signal specific. As such, the modulation/demodulation problem is ill-posed (Turner and Sahani 2011) but is still fundamental to understanding the information-bearing components of signals.

In the case of complex audio signals (speech, music, natural, or communication sounds), getting a clear idea of the identity of the message and carrier components remains one of the holy grails of research on the physical and perceptual underpinnings of sound. Interest in modulations of an audio signal aims to pinpoint the information-bearing components of these signals, especially given the redundant nature of the waveforms that can emanate from both mechanical (e.g., instrument, vocal tract) or electrical (e.g., computer generated) sound sources.

The problem is particularly compounded because complex audio signals, such as speech and music, contain information and modulations at multiple time scales and across various spectral constructs. In the case of speech, there is an extensive body of work dating back to the early twentieth century that explored the span and dynamics of the speech envelope. The argument that the slow envelope is the chief carrier of phonetic information in speech is quite old. In the 1930’s, Dudley advocated that the dynamics of signal envelopes are important for describing linguistic information in speech (Dudley 1939, 1940). In his view, the vocal tract is a sluggish system that slowly changes shape, with low syllabic frequencies up to 10 Hz, giving rise to varying modulating envelopes that contribute most to the intelligibility of speech.

Building on this work, numerous studies have shown that speech intelligibility is well maintained after temporal envelopes are lowpass filtered or degraded, with a critical range between 5–15 Hz that spans the range of phonemic and syllabic rates in natural speech (Greenberg 2004). Still, the modulation spectrum profile of speech is a complex one and reveals that the speech envelope contains energy of the order of a few to tens or hundreds of Hertz. This profile highlights key energy fluctuations in speech signals, ranging from hundreds of milliseconds (of the order of multiple syllables or words) to tens of milliseconds (typically spanning subsyllabic and phonemic segments) (Rosen 1992; Divenyi et al. 2006). The complexity of speech signals includes the multiplexed information across various time scales but also variations across frequency bands and in the phase relationships across bands (Pickett 1999).

In the case of music signals, a similar picture emerges spanning multiple time scales, frequency bands, and spectral profiles. The information-bearing components of a musical signal, be it an isolated note or a full orchestral piece, appear to multiplex across a complex construct of spectrotemporal dimensions. Much like speech, music signals have melodic, harmonic, and rhythmic structures that intertwine into intricate patterns (both in time and frequency) to convey the complex acoustic experience of music perception. Recent advances in computing power, signal processing techniques, and increased availability of digitized audio material have led to fairly sophisticated analysis tools to study various aspects of regularity in music, such as rhythm, melody, harmony, or timbre (Müller 2015; Meredith 2016).

Despite the intricate nature of spectrotemporal regularities in both speech and music, they share fundamental attributes reflected in their decomposition into alphabetic tokens (phonemes, syllables, word, notes, chords), assembly of sequences of events (accents, grouping, words, phrases), and rhythmic structure (time, stress), all interleaved with specific spectral patterns that reflect the sound sources (instrument, oral cavity), production style, and contextual attributes. The correlates of these regularities can be gleaned from examining the modulation patterns in the signal at multiple time scales and granularities. This chapter reviews common techniques used to represent modulations in speech and music signals and their implications for understanding the information-bearing components in these signals. Section 12.2 reviews signal processing tools commonly used to represent modulations: fundamental time-frequency representations (Sect. 12.2.1), spectrotemporal modulation profiles (Sect 12.2.2), and temporal modulation spectra (Sect 12.2.3). Sect. 12.2.4 delves into representations that are unique to speech and music signals and considers constraints imposed by the physics of the vocal tract and controlled sound production through most musical instruments. Section 12.3 offers insights into the neurophysiological interpretation of modulations, particularly encoding of the spectrotemporal signal envelope along the auditory pathway. Section 12.4 reviews key findings in psychophysical and physiological research into the role of modulation in speech and music perception. Section 12.5 provides a summary of the main ideas in the text along with perspectives on future research directions.

12.2 Representation of Modulations

12.2.1 The Time-Frequency Representation

A complete description of the information content of speech and music signals is not possible. However, one can derive a number of low-level empirical descriptors that reveal a lot about the structure of these signals. Common ways to explore the nature of these signals involve analysis of the acoustic waveform as well as its frequency content. A time-frequency profile, typically obtained via short-time Fourier transform, wavelet, or filterbank analysis (Caetano, Saitis, and Siedenburg, Chap. 11), best displays the variations of energy as the signal evolves over time. Fig. 12.1A depicts the time-frequency representation of a speech utterance produced by a male speaker saying /we think differently/. Immediately emerging from this spectrographic view of speech is the fact that the temporal envelope varies slowly over the course of tens to hundreds of milliseconds. In fact, one can easily discern the volleys of activity across frequency channels, occurring at a rate of 5–7 peaks per second, commensurate with phonemic and syllabic contours of the speech utterance. The right subpanel in Fig. 12.1A highlights a cross-section of this spectrogram around 450 Hz, which represents the half-wave rectified output of the auditory filter centered about that spectral region. The time waveform clearly shows an overall fluctuating pattern around 6 Hz, which closely follows segmental and syllabic landmarks of the speech signal (Poeppel et al. 2008). A similar structure emerges spectrally with frequency profiles that are largely coarse. The energy distribution across frequency channels appears to mostly delineate harmonic and formant peaks (bottom-left subpanel in Fig. 12.1A).

In parallel, Fig. 12.1B illustrates an example of the time-frequency spectrogram of the finale of Tchaikovsky's violin concerto in D major, Op. 35. The time-frequency spectrogram highlights the exuberant energy in this piece with a very dynamic temporal profile reflective of the vigorous nature of this finale. The clear steady tones typical of bowed string instruments are also clearly visible throughout the passage, with the spectral profile showing the clear harmonic nuances of the solo violin performance. Still, the rich energy of this final movement of the concert is not readily discernable from the spectrogram view only. The cross-section of this spectrogram along time emphasizes the nested dynamics over the course of a 6 s period. The soft onset signature of the violin is not very evident due to the multiscale rhythmic modulations in this extravagantly energetic piece with discernable Russian rhythmic undertones (Sadie 2001). The temporal envelope clearly shows a fast-paced profile modulated by a much slower rhythmic profile varying at rate of 1–3 peaks/s. The spectral cross-section shown in the bottom-left panel in Fig. 12.1B takes a closer look at the frequency profile of the signal around 2.3 s. The characteristic profile of a violin note clearly emerges with the overall envelope highlighting the resonance of the violin body with three obvious peaks (Katz 2006). Within the broad peaks, one can glimpse details of the spectral structure imposed by the mechanical constraints of the violin along with the unambiguous harmonic structure of the note.

Fig. 12.1 (Continued) as a function of time; a *frequency cross-section* of the spectrogram around 250 ms as a function of log-frequency; and the *two-dimensional Fourier transform* (*2D FFT*) of the time-frequency spectrograms that yields the modulation power spectrum of the signal. The figure was interpolated using linear interpolation and compressed to a power of 2.5 to obtain better color contrast (for display purposes only). (**B**) The spectrotemporal details of the finale of Tchaikovsky's violin concerto in D major, Opus 35, using similar processing steps as in panel **A**

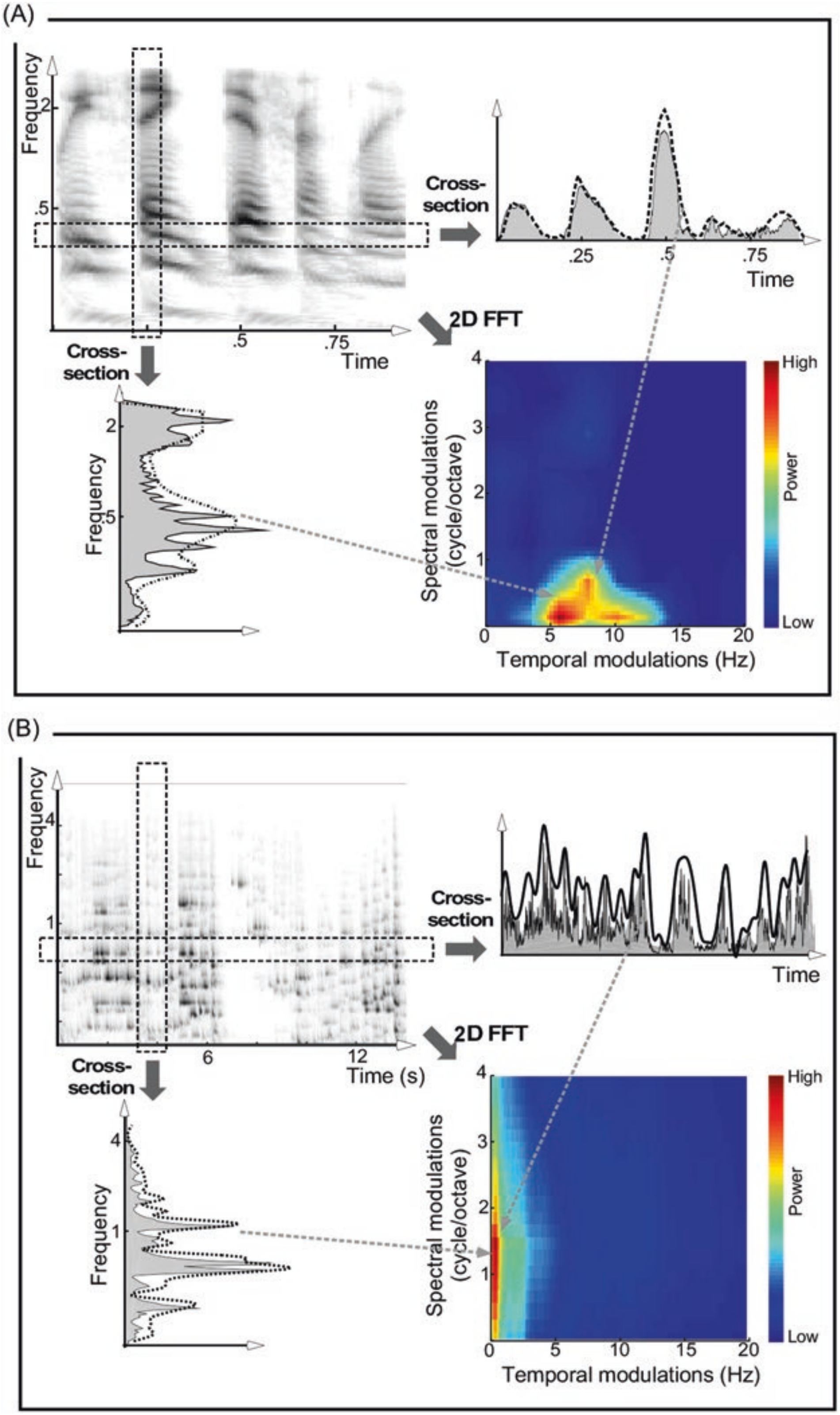

Fig. 12.1 Spectrotemporal details of speech and music. (**A**) The *time-frequency spectrogram* of a male utterance saying /we think differently/ over a time span of 1 s and frequency range of 5 octaves (note the log frequency axis); a *temporal cross-section* of the spectrogram around 450 Hz

12.2.2 The Spectrotemporal Modulation Profile

A better illustration of these spectrotemporal modulation details can be achieved in the spectral/Fourier domain, obtained by performing a two-dimensional Fourier transform of the time-frequency spectrogram (Fig. 12.1A, B, lower right panels). This operation estimates the power distribution of both spectral and temporal components over the chosen time and frequency spans and yields the *modulation spectrum* of the signal (Singh and Theunissen 2003). The modulation spectrum is an account of the distribution of time-frequency correlations of adjacent and far-away elements in the signal and, hence, is an estimate of the degree and dynamics of signal fluctuations along time and frequency axes. Immediately worth noting in this modulation spectrum is that the energy in the Fourier domain is mostly concentrated in a highly localized region of the modulation space.

For a speech signal (Fig. 12.1A), the modulation spectrum highlights what was already seen in the unidimensional profiles. For instance, the temporal envelope induces a strong activation peak between 5 and 7 Hz, while the spectral modulations reveal discernable energy at a harmonic rate (i.e., distance between harmonic peaks) or coarser (i.e., distance between formant peaks), which appear as strong activity around 1 cycle/octave and below. The modulation spectrum energy for the music signal (Fig. 12.1B) also accentuates the modulation patterns observed in the cross-sections of the spectrogram. A strong activation pattern around 1 cycle/octave clearly highlights the crisp harmonic peaks of a violin sound, while the temporal modulations show a distributed energy that is strongest below 3 Hz but spread as far as 10 Hz, highlighting the strong vibrato and clear articulation in this piece that carry the slower rhythmic structure.

Unlike conventional methods for computing the modulation spectrum (traditionally confined to a transform in the temporal dimension, discussed in Sect. 12.2.3), the two-dimensional modulation spectrum highlights *both* the spectral and temporal dynamics of the signal as well as the time alignment of these modulation patterns (i.e., cross-channel modulation phase), which is an important component for understanding spoken material and music compositions (Greenberg and Arai 2001; Hepworth-Sawyer and Hodgson 2016). The combined profile—across time and frequency—is the only mapping able to highlight subtle patterns in the original envelopes, such as frequency-modulations (FM), which are key signatures of many patterns in transitional speech sounds (e.g., diphthongs, semi-vowels), and metallic or percussive bell sounds (Chowning 1973).

Because of its span of the joint time-frequency space, the spectrotemporal modulation power spectrum (MPS) representation has been used as a dashboard to explore the precise loci of modulation energy driving the perception of speech and music. Recent work examined detailed tiling of the spectrotemporal modulation spectrum using granular techniques that inspected the perceptual contribution of various regions or building-blocks of the two-dimensional modulation profile. These methods, originally developed in vision research, aim to assign a quantifiable contribution of specific modulation energies to perceptual recognition of sound constructs using an approach referred to as "bubbles" (Gosselin and Schyns 2001). Because the spectrotemporal modulation profile is in a fact an image with temporal modulations on the *x*-axis and spectral modulations on the *y*-axis, the adoption of vision techniques can be seamlessly applied. These approaches have shown that the intelligibility of speech signals depends significantly on both spectrotemporal modulations that carry considerable modulation energy in the signal as well as those that carry linguistically relevant information (Venezia et al. 2016). A similar observation has also been reported for musical instrument recognition where low spectral and temporal modulation are the most salient regions to correlate with musical timbre, though signatures of individual instruments can be clearly discerned in the MPS space (Thoret et al. 2016). Alternative tiling techniques that use filtering (low-pass, notch filters) as well as dimensionality reduction and scaling have also been used to explore the informative regions of the MPS space (Elliott and Theunissen 2009; Elliott et al. 2013).

Overall, the MPS representation is proving to be a powerful descriptor of sound identity and timbre representation. It is also a space where joint interactions across time and frequency can be readily discerned. Still, it is not a very intuitive mapping of the acoustic waveform because it is a representation derived from the signal via at least two (typically more) transformations: from the acoustic signal to a time-frequency spectrogram and then to a time-frequency modulation spectrum (in addition to computing magnitude, power, binning operations, etc.). The models employed to perform these transformations do shape the salient details of the modulation profile and can invariably emphasize different aspects in this mapping, be it stimulus energies or perceptual energies.

The representation shown in Fig. 12.1 employs a straightforward two-dimensional Fourier transform to the time-frequency spectrogram. Other approaches have been proposed, including the use of two-dimensional wavelet transforms (Anden and Mallat 2014), bio-mimetic affine transforms mimicking receptive fields in mammalian auditory cortex (Chi et al. 2005), or even physiologically recorded receptive fields from single neurons in primary auditory cortex (Patil et al. 2012). Naturally, incorporating nonlinearities as reported in auditory processing can further color the readout of such modulation profiles, though limited work has been done that can shed light on the biological and perceptual relevance of nonlinearly warping the modulation space (as discussed in Sec. 12.5).

One of the other limitations of the modulation spectrum stems from the fundamental limit in precision by which modulations can be measured simultaneously in time and frequency. Much like the uncertainty principle is applied in a time-frequency spectrogram, the same is true in the modulation domain, which is effectively a transformation of the original space. The uncertainty principle, or Heisenberg principle, articulates the trade-off that one can achieve when attempting to represent time and frequency with infinite precision (Cohen 1995; Grochenig 2001). The smaller the window in time used to perform the analysis, the larger the bandwidth of spectral resolution afforded by this analysis because these two quantities have effectively a fixed product. Similarly, the temporal and spectral modulations derived within these constraints are also restricted relative to each other and, as such, provide a limited view of the spectrotemporal modulation profile of a signal (Singh and Theunissen 2003). How the brain deals with these limitations remains unknown, though they may explain the multi-resolution mapping of modulation in auditory cortical networks, as discussed in Sec. 12.3.

12.2.3 The Temporal Modulation Spectrum

As discussed in Sect. 12.2.2, the notion of modulation aims at identifying the patterns of inflection or change imposed on a signal. While the formal definition of such change does not necessarily identify the dimension on which it needs to operate, there is a large body of work that has focused on the temporal envelope. The temporal envelope is the main carrier of rhythmic fluctuations in the signal, and therefore its timescale and timespan are crucial information-bearing components of the signal. It is important to note that modulations along frequency also play a crucial role (as mentioned in Sect. 12.2.2; this issue will be expanded further in Sect. 12.4). Still, the temporal profile has garnered particular interest because of its simple mathematical derivation yet powerful importance in speech and music perception (Patel 2008).

The temporal modulation spectrum is obtained through a series of transformations that pass a signal *x[n]* through a bank of *M* bandpass filters in order to derive the envelope of each filter output. While this process is traditionally done on band-limited signals at the output of each filter, the premise of the computation does not preclude using broadband signals nor does it confine the bandwidth of the filterbank to a specific range. Naturally, the fluctuations of the filter outputs will be dictated by the choice of filterbank parameters, bandwidths, and frequency span.

Techniques used in the literature vary from using simple Fourier-like spectral decompositions (e.g., Fig. 12.1) to more perceptually grounded spectral mappings based on critical bands or a Bark scale (Moore 2003). The output of this filterbank analysis is an array of M filter outputs:

$$x_m\left[n\right]; m = 1, \ldots, M$$

The fluctuations of these output signals are then further isolated using an envelope extraction technique (either using the Hilbert transform or other transformations such as half-wave rectification and low-pass filtering), which results in a smooth envelope of each filter output ($E[x_m]$) whose variations are bounded both by the original bandwidth of the filterbank as well as the constraints of the envelope-tracking technique (Lyons 2011). Typically, this process is followed by a nonlinear mapping that compresses the linear envelope output using a nonlinear scaling function, such as square, logarithm, or a biologically motivated nonlinearity-mimicking nonuniform gain compression in the activation of the auditory nerve (Yang et al. 1992; Zhang et al. 2001). The compression is also used to counter the strong exponential nature of envelope amplitudes in natural sounds (Attias and Schreiner 1997).

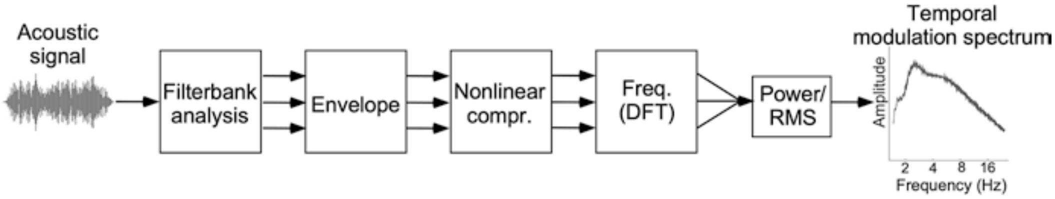

Fig. 12.2 Schematic of processing stages to derive the temporal modulation spectrum from an acoustic signal. The acoustic signal undergoes an initial analysis to map it onto a time-frequency representation before transformations of this spectrogram extract a temporal modulation spectrum from the envelope across different frequency channels. *DFT*, discrete Fourier transform; *RMS*, root-mean-squared

The readout of the fluctuations in the envelope signal is then obtained in the Fourier domain by mapping the time-domain signals onto a frequency-axis profile that is then summed across channels and transformed into power, root-mean-squared energy, or compressed magnitudes (Fig. 12.2).

Historically, this approach has been developed in the room acoustics literature via the concept of a modulation transfer function (MTF) (Houtgast and Steeneken 1985) and thus has relied on modulation filters employed to analyze the energy in the envelope signal at specific modulation points chosen along a logarithmic scale. An equivalent readout can be obtained using linearly spaced filters or by directly employing a Fourier transform on the compressed envelope signals. In either case, the resulting profile can then be combined across frequency bands and properly binned and scaled to yield an amplitude modulation spectrum that reflects envelope energies along different modulation frequencies. A major underlying assumption in this transformation is that such modulation frequencies of interest are below the pitch range, focusing primarily on the true envelope patterns or slow fluctuations in the signal. A number of constraints in the design of the processing steps must be considered in order to avoid artifacts or distortions that could mislead the readout of the spectrum profile (Qin Li and Les Atlas 2005).

12.2.4 Domain-Centric Representations

Some approaches have considered more structured analyses of the signal. In the case of speech sounds, the source-filter model of speech production has led to widely used techniques such as *Linear Predictive Coding* (LPC) (Schroeder and Atal 1985; see Caetano, Saitis, and Siedenburg, Chap. 11). The approach builds on the minimal but powerful simplification of speech production as a coupling of a vibrating source that generates the carrier signal with a filter that colors this carrier, hence giving speech its spectral shapes. As such, being able to decompose the signal into these two fundamental components disentangles the voicing characteristics primarily present in the source from the timbral cues primarily shaped by the filter, though there is a strong interaction between the two. From a linear systems point of view, separating the source (glottal signal) from the system (parameters of the vocal tract) means that the current speech sample can be closely approximated as a linear combination of past samples (hence the name linear predictive coding) (Rabiner and Schafer 2010). While an oversimplification of the complex dynamics of speech production, LPC modeling offers an indirect, yet effective, account of the spectral modulations shaping phonetic tokens of speech signals, though the temporal dynamics are often ignored by assuming the system (vocal tract) is quasi-stationary over short periods of time that span the analysis window.

A similar decomposition of source and filter cues underlies the widely popular cepstral decomposition, which provides a transformation of the filter characteristics in the *cepstral domain*. The cepstrum (a rotated version of the word spectrum) is an application of homomorphic signal processing techniques that apply a nonlinear mapping to a new domain wherein two components of a signal can be disentangled or deconvolved (Rabiner and Schafer 2010). Applied to speech signals, the *power cepstrum of a signal* is defined as the squared magnitude of the inverse Fourier transform of the logarithm of the magnitude of the Fourier transform of a signal (Caetano, Saitis, and Siedenburg, Chap. 11). Effectively, the cepstrum domain separates the slowly varying envelope (or modulation) signal from the rapidly varying excitation carrier signal, allowing the analysis of each component separately. The result is cepstral coefficients (and the related mel-frequency cepstral coefficients or MFCC) that offer an effective account of phoneme-dependent signal characteristics (Childers et al. 1977). Much like LPC, cepstral analysis remains limited to static representation of short segments of the speech signal (typically of the order of a phoneme) and focuses solely on the spectral characteristics of the modulating envelope.

Other approaches have been used to extend these representations to the time domain by computing derivative and acceleration over time, often referred to as delta and delta-delta coefficients of the signal, in an attempt to capture some of the temporal dynamics in the speech signal driven by prosodic and syllabic rhythms. While derivatives are rather simplistic extensions to capture the intricate temporal structure of the vocal tract during speech production, techniques such as LPC and MFCC remain powerful tools that provide a basic bread-and-butter analysis of speech signals with a formidable impact on many applications of speech analysis (Chen and Jokinen 2010; Hintz 2016). Their popularity speaks to the tremendous redundancies in speech signals as well as the powerful impact of a simple source-filter model in capturing some of the nuances of how speech signals are shaped and how they carry information.

While this source-filter view is rather unique to the speech production system, it is also applicable and quite popular for music analysis (Collins 2009; also see Caetano, Saitis, and Siedenburg, Chap. 11). Many musical instruments can be viewed as pairings of a source (a vibrating object such as a string) coupled with a filter (the body of the instrument that shapes the sound produced). Unlike a unitary model of source-filter analysis in speech, a common production system cannot be applied across instruments since the production may depend on vibrating strings, membranes, or air columns. As such, the distinction between the source and the filter is not as distinct as it is in speech and poses some challenges when applied to music signals, especially for non-Western music or polyphonic music (Muller et al. 2011).

While many approaches for music analysis borrow from a long tradition of speech processing, a number of elegant techniques have been developed specifically for music analysis particularly applied to domains of pitch, harmony, beat, tempo, and rhythm. The modulatory fluctuations in music, of both the spectral profile as well as the temporal envelope, have inspired a number of clever decompositions of music in order to hone in on the modulatory fluctuations in the signal. Some of these techniques extend the concept of a temporal modulation spectrum across multiple time scales. For instance, a family of modulation spectra spanning fast tempi (called *meter vectors*) offer a hierarchy of modulation spectra that summarizes the temporal patterning of events in a music signal nested across multiple time constants (Schuller 2013).

Overall, the analysis of modulations in speech and music signals is often informed by particular aspects of signal perception or production under study or with the ultimate goal of identification, recognition, or tracking. As such, the field enjoys a wide variety of tools developed from different perspectives that represent various facets of modulation. Ultimately, the modulation spectrum (in its many forms) has rather direct neurophysiological interpretations, as discussed in Sec. 12.3, though the elucidation of the exact substrate of specific forms of modulation encoding remains an open area of research.

12.3 Neurophysiological Interpretation of Modulations

The mapping of the informative acoustic attributes of an incoming signal takes different forms and varying levels of granularity as the signal is analyzed along the auditory pathway (Eggermont 2001). As early as cochlear processing, a sound signal entering the ear is decomposed along many bandpass frequency regions that span the basilar membrane, resulting in a time-frequency representation much like a short-term Fourier transform. The intricate details of sensory hair cell transduction shape the response across cochlear channels through a number of processing stages often modeled using half-wave rectification, low-pass filtering, and nonlinear compressions (Yang et al. 1992; Ibrahim and Bruce 2010). This process, analogous to deriving the envelope of an analytic signal using the Hilbert transform, effectively tracks the temporal variations of the signal along different frequency bands, which not only highlights the overall temporal patterns of the signal but specifically underscores the profiles of onsets and sustained activity as well as rhythmic changes (e.g., temporal cross-sections in Fig. 12.1).

The details in this temporal profile are encoded with gradually lower resolutions along the auditory pathway where the neural code appears to be increasingly selective to the slower dynamics that modulate the signal profile (Miller et al. 2002;

Escabi and Read 2003). This selectivity is reflected in the tuning parameters of neurons from the midbrain all the way to primary auditory cortex. Neural tuning characteristics are typically summarized using spectrotemporal receptive fields (STRF) (Elhilali et al. 2013). The STRF is a powerful tool in studying the selectivity of neurons to particular patterns in the stimulus. It typically treats a neuron as a system with a known input (the sound stimulus) and a measured output (the neural response). As is common in systems theory, the characteristics of a system (i.e., the system function) can be derived from its input and output or a class of inputs and corresponding outputs. This system function allows one to think of a neuron as a filter with a STRF that reflects the characteristics of the stimulus that best induces a strong response.

This STRF representation has been invaluable in shedding light on tuning characteristics of neurons along the auditory pathway. Of particular interest to the current discussion is the selectivity of neurons at the level of auditory cortex. While there is a great deal of variability across species and cortical layers, most auditory cortical neurons are sensitive to slow temporal and spectral modulation patterns (Depireux et al. 2001; Liu et al. 2003) commensurate with scales and dynamics of interest in modulation profiles, as discussed in Sect. 12.2. Unlike tuning in peripheral auditory nuclei, which captures mostly tonotopic energy across frequency, cortical neurons exhibit tuning sensitivity across at least three dimensions: (1) best frequencies (BF) that span the entire auditory range; (2) bandwidths that span a wide range from very broad (~2 octaves) to narrowly tuned (< 25% of an octave) (Schreiner and Sutter 1992; Versnel et al. 1995); and (3) temporal modulation dynamics that range from very slow to fast (1–30 Hz) (Miller et al. 2002).

Interpreting this representation from the vantage point of signal modulations, neural responses of a whole population of cortical neurons are mostly driven by temporal dynamics in the signal that are commensurate with the sound envelope (< 30 Hz). As a population, ensemble tuning of cortical neurons can therefore be tied to the temporal modulation spectrum of natural and complex sounds (Depireux et al. 2001; Miller et al. 2002). Complementing this axis are the spectral dynamics of the neural response across a cortical ensemble of neurons, which also spans spectral energies typical in signals with a characteristic resonance structure (extended over many octaves), that are able to extract harmonic and subharmonic structures in the spectrum (Schreiner and Calhoun 1995; Kowalski et al. 1996). The spectral selectivity of cortical neurons appears to match rather well the distinctive profile of spectral shapes in natural sounds, supporting the theory of a faithful alignment between acoustic modulation energy and neural encoding of such spectral modulations, which ultimately guides processing and perception of complex sounds (Leaver and Rauschecker 2010). Taking both dimensions into account, the considerable match between the modulation spectrum (derived directly from a signal corpus) and the tuning characteristics of an ensemble of cortical STRFs has been argued in the literature as possible evidence for the underlying role of the mammalian auditory cortex in encoding information-bearing components of complex sounds (Singh and Theunissen 2003).

While this view—at the ensemble level—reveals a formidable match between the acoustic properties of the signal and cortical neural tuning, the details of how these contours are derived are important to bear in mind because they impose a number of constraints on the modulation profiles under study and their interpretations. As mentioned earlier, the STRF is commonly interpreted through a systems theory view that deduces a system function based on the mapping between the input stimulus and the recorded neural response. Given interest in a system function that spans both time and frequency, a spectrotemporal representation of the stimulus is often preferred. However, the exact signal processing transformation used to map the spectrotemporal space dictates, to a great degree, the view and details emerging about the STRF. For instance, the tiling of the time-frequency space, the detailed resolution or sampling of such space, and the scaling of the amplitude energy profile of the stimulus can greatly affect the readout emerging from the neural mapping of this transformation and its match to the brain responses induced by complex acoustic stimuli.

Of particular interest is whether the use of a wavelet-based representation (based on logarithmic filter spacing) versus a spectrogram approach (akin to a Fourier transform) is more informative about the modulation spectrum and its neural underpinnings. On the one hand, wavelet-based analyses are generally preferred in explaining a number of perceptual findings, including modulation-tuning thresholds (Chi et al. 1999), given the closer biological realism in mimicking the frequency resolution provided by the auditory periphery. On the other hand, the time-frequency resolution tradeoff allows more modulation dynamics at the higher frequency bands of a wavelet representation and could magnify the effect of faster temporal dynamics. As such, linearly spaced filters have been preferred for deriving modulation spectra, especially when considering the temporal dynamics (Jepsen et al. 2008; Elliott and Theunissen 2009).

Though it is difficult to objectively quantify and compare the adequacy of different time-frequency mappings, a common technique used in the literature is to assess the goodness-of-fit for different mappings. A report by Gill et al. (2006) performed a systematic study of sound representations in an effort to elucidate the importance of certain factors in the derivation of neuronal STRFs. The study examined a number of parameters, particularly the use of linear versus logarithmic spacing of modulation filters, in deriving the time-frequency representation of the signal. Gill et al. (2006) found little evidence for a clear advantage in using linear versus logarithmic filter tiling for the derivation of time-frequency spectrograms of the stimulus and, subsequently, for the goodness-of-fit models of auditory neurons in the songbird midbrain and forebrain.

In contrast to the different ways of spectral tiling, which show little to no effect, Gill et al. (2006) reported stronger effects of adaptive gain control and amplitude compression of the stimulus in assessing auditory tuning. Those two aspects reflect the need for nonlinear transformations (both static and dynamic) in characterizing the neural underpinnings of auditory tuning to sound modulations. Nonlinear mappings of the time-frequency profile of the stimulus not only reflect the complex nature of neural processing along the auditory pathway, they also highlight the multiplexed layers of information-bearing components of natural sounds (Santoro et al. 2014). Reducing the concept of modulations to an envelope riding on top of a carrier is too simple to explain its role in timbre perception, especially for complex sounds.

12.4 How Informative are Modulations?

12.4.1 Modulations in Speech

What does the speech modulation spectrum reveal about understanding spoken language? Work dating a few decades back showed that comprehension of speech material is highly impaired in acoustic environments where distortions attenuate energies between 2–8 Hz (Steeneken and Houtgast 1979; Houtgast and Steeneken 1985). Those observations were further corroborated by later work in different languages that showed a dramatic decline in intelligibility if the integrity of the temporal modulation profile of speech was altered (with operations such as low-pass or bandpass filtering) (Drullman et al. 1994; Arai et al. 1999). Similar distortions disrupting the integrity of the spectral modulation profile by phase jitter or bandpass filtering are also equally detrimental to intelligibility, even if they do not alter the temporal envelope profile of speech (Arai and Greenberg 1998; Elhilali et al. 2003). In contrast, numerous studies have argued that any manipulations of speech that do not disrupt the integrity of its spectrotemporal modulations are harmless to its intelligibility (Shannon et al. 1995; Zeng et al. 2005). All in all, there is growing evidence that the spectrotemporal features captured by the speech MPS (see Sect. 12.2.2) offer a representation that closely maintains the phonetic identity of the sound as perceived by human listeners (Elliott and Theunissen 2009). The fidelity of the speech MPS correlates closely with intelligibility levels of speech in the presence of ambient noise and other distortions (Elhilali and Shamma 2008). The more a noise distorts the speech MPS, the more the decline of speech intelligibility. Conversely, noises that fall outside the core acoustic energy of the speech MPS have little effect on its intelligibility levels (Carlin et al. 2012).

The role of the spectrotemporal modulations of speech as information-bearing components has been leveraged extensively to sample speech signals for many applications, particularly automatic speech recognition (ASR) in the presence of background noise. Modulation-based analysis has enjoyed a lot of success as front-ends for ASR systems. Most studies have focused on the temporal evolution of the signal envelope to quantify modulation spectra (Kingsbury et al. 1998; Moritz et al. 2011), or estimations of the envelope *pattern* using temporal envelopes (Hermansky and Sharma 1999; Morgan et al. 2004), or using frequency-domain linear prediction (FDLP) (Athineos and Ellis 2003; Ganapathy et al. 2010). Also, a few attempts have been made to extend the analysis of modulations to both spectral and temporal domains; these studies have focused mainly on using two-dimensional Gabor filters (or other variants) as localized features for analysis of speech (Kleinschmidt 2003; Meyer et al. 2011). Across all of these different representations, the common thread is that once the speech signal is mapped onto a space that directly highlights its modulation content, the fidelity of that representation is sufficient to maintain the speech content and facilitate its robust recognition (Nemala et al. 2013). As such, this robustness provides empirical corroboration that such envelope modulations are indeed important information-bearing components of speech.

A faithful representation of speech signals has direct relevance for hearing prosthetics, particularly cochlear implants (CI), for which the fidelity of the signal has direct perceptual implications for the user (for more on timbre perception by CI users, see Marozeau and Lamping, Chap. 10). Speech modulations along the spectral axis are of particular interest in the case of cochlear implants because they dictate the resolution of the frequency axis and, ultimately, the channel capacity of the prosthetic device. Numerous studies have reported minimal disruption of speech comprehension in noise-free environments when only a few frequency channels are present over a range of hundreds of Hertz below 4 kHz (Shannon et al. 1995). Importantly, as few as four channels (i.e., a spectral resolution as low as 1.6 cycles/octave) are sufficient to maintain intelligibility. Such resolution is generally too low for acceptable levels of speech recognition in noise and also results in impoverished music perception (as discussed in Sec. 12.4.2). By the same token, it has been argued that fully resolving formant spectral peaks (up to 2 cycles/octave) results in great improvement in intelligibility, especially when speech is corrupted with noise (Friesen et al. 2001; Elliott and Theunissen 2009). The tradeoff between the spectral resolution sufficient for speech perception in quiet settings and the spectral resolution necessary for speech recognition in the presence of noise remains a matter of debate (Friesen et al. 2001; Croghan et al. 2017). This is especially important given the variability across listeners in their ability to utilize the spectrotemporal cues available to them.

The debate over modulations and spectrotemporal resolutions necessary for speech perception highlight the fact that there is more to speech than just its envelope (Moore 2014). While the view of modulations as an envelope fluctuation riding a fast carrier is true to a great extent, that view conceals the complex role played by the underlying fast structure of the signal in complementing the representation, and ultimately the perception, of speech signals. The temporal fine-structure and spectral details play key roles in speech perception in noise (Qin and Oxenham 2003; Shamma and Lorenzi 2013), sound localization (Smith et al. 2002), lexical-tone perception (Xu and Pfingst 2003), repetition or residue pitch perception (deBoer 1976), and fundamental frequency discrimination (Houtsma and Smurzynski 1990). Psychophysical evidence suggests that one of the advantages that normal subjects have over hearing-impaired listeners is improved local target-to-masker ratios, especially in the presence of spectrally and temporally fluctuating backgrounds (Peters et al. 1998; Qin and Oxenham 2003). The notion of listening in the spectral and temporal "dips" of the masker sounds is less realizable for hearing impaired listeners because of poor spectral selectivity and reduced temporal resolution (Glasberg and Moore 1992).

Fine details of speech (especially along the spectrum) are also crucial for dealing with stationary and narrowband noises and pitch-centric speech processing

(McAuley et al. 2005; Wang and Quatieri 2012). Hence, one has to be careful in interpreting the perceptual salience of the slow envelope for speech perception as an exhaustive account of the speech signal. Reducing the speech signal to a dichotomy consisting of two independent components—envelope and fine-structure—is a flawed premise. The envelope and fine-structure components are not only impossible to tease apart, but they also convey complementary information about the speech signal, especially in everyday listening environments (Shamma and Lorenzi 2013).

12.4.2 Modulations in Music

Much like speech, music signals carry a multiplexed and highly layered structure of dynamics both spectrally and temporally. Music perception evokes a complex experience that spans multiple elements that include pitch, melody, timbre, and rhythm among others. The representations of signal modulations in their different forms directly encode many facets of these musical attributes (see Caetano, Saitis, and Siedenburg, Chap. 11). Among musical elements, modulations have a very tight affiliation with the perception of timbre both in terms of sound identity but also as musical quality.

Acoustically, a musical note is shaped by the physical constraints of the instruments as well as the motor control of the player. These constraints whittle the acoustic signal with modulatory envelopes that carry some of the timbral properties of music. The acoustic signature of these constraints naturally shapes both spectral and temporal profiles of the acoustic signal, and they ultimately inform the perceptual experience as these cues are decoded by the auditory system. Numerous perceptual studies have shed light on these acoustic correlates (McAdams, Chap. 2; Agus, Suied, and Pressnitzer, Chap. 3) with spectrum as the most obvious candidate. The spectral shape of a musical note is naturally shaped by the vibration mode and resonances of the instrument and that modulates not only the spectral energy profile but also frequency peaks, spectral sharpness and brightness, amplitudes of harmonic partials, spectral centroid, and spectral irregularities. The temporal envelope of the signal is also heavily modulated, and correlates of timbre can also be gleaned from the energy buildup, onset information, attack over time, and the spectral flux over time. All these attributes, spanning both spectral and temporal modulations, not only determine the identity of a musical instrument but also the perceived timbral quality of musical-instrument sounds.

In a study directly relating spectrotemporal modulations to the perception of timbre, Patil et al. (2012) explored the fidelity of neural activation patterns in mammalian auditory cortex in accurately replicating both classification of musical instruments as well as perceptual judgements of timbre similarities. The study examined the ability of a cortical mapping to reflect instrument-specific characteristics. Patil et al. (2012) specifically assessed whether a processing pipeline that mimicked the transformation along the auditory pathway up to primary auditory cortex was able to capture the instrument identity from a wide variety of isolated notes from eleven instruments playing 30–90 different pitches with 3–10 playing styles, 3 dynamic levels, and 3 manufacturers for each instrument (an average of 1980 tones per instrument). The model was able to distinguish the identity of different instruments with an accuracy of 98.7%, corroborating the hypothesis that timbre percepts can be effectively explained by the joint spectrotemporal analysis performed at the level of mammalian auditory cortex.

Patil et al. (2012) also examined a more stringent constraint to explore how well this cortical mapping reflected distances between instruments that correlated with the perceptual judgements of timbre similarity by human listeners. In other words, it is not sufficient to judge whether a timbre representation is able to distinguish a violin from a cello, but can it also discern that a violin is perceived as more similar to a cello than it is to a bassoon. The representation based on spectrotemporal receptive fields was indeed able to project notes from individual instruments onto a space that maintains their relative distances according to similarity judgements of human listeners. The faithful representations of spectrotemporal modulations in the cortical space were correlated with human similarity judgements with an accuracy of $r = 0.944$.

While the relation between spectrotemporal modulation tuning at the level of primary auditory cortex and timbre perception is quite strong, it is important to note a number of observations. The fact that the timbre space spans a complex interplay of spectral and temporal dimensions is not surprising and has been established through a large body of work spanning many decades (see Siedenburg, Saitis, and McAdams, Chap. 1). What timbre analysis via a biomimetic cortical model sheds light on is the fact that the decoding of acoustic modulations along both time and frequency over a rich representational space appears to be necessary and sufficient to almost fully capture the complete set of acoustic features pertinent to instrument identity and timbre similarity. It also pushes forth the debate about the cardinality of a timbre space, one that extends beyond few descriptors to require a high number of dimensions. This direct relationship between modulations and timbre perception reinforces the theories tying modulation with information-bearing components of the musical signal.

One of caveats to this theory (that the study by Patil and colleagues brought to light) is that the modulation space cannot be a separable one, spanning marginally along time and frequency (Patil et al. 2012). Rather, the *joint* representation along both directions is crucial, emphasizing spectrotemporal dynamics in the timbre profile (see McAdams, Chap. 2). For instance, frequency modulations (FM), such as vibrato, impose rich dynamics in music signals, and they can only be discerned reliably by examining the joint spectrotemporal space. The role of spectrotemporal modulations that underly music perception has been directly reported using psychophysical studies that correlate music perception abilities and modulation detection thresholds for time alone, frequency alone, and joint time-frequency (Choi et al. 2018). The correlations are stronger with spectrotemporal modulation-detection thresholds, further corroborating the idea that the configuration of the timbre space directly invokes a modulation space based on *joint* spectrotemporal dynamics (Elliott et al. 2013).

Another important observation from the Patil et al. (2012) study is that timbre representation in a biomimetic spectrotemporal modulation space is only effective at replicating human judgements when augmented by a nonlinear mapping boundary. A number of studies, in fact, have established this nonlinear behavior, especially at the level of auditory cortex, as it pertains to encoding of complex sound patterns (Sadagopan and Wang 2009). The exact nature, neural underpinnings, and the specificity of this nonlinearity to different sound classes remain unclear. As such, the quest for a direct mapping between spectrotemporal modulations and a timbre space remains unfulfilled.

12.4.3 Common and Unique Modulation Profiles in Speech and Music

As one examines the relationship between modulation profiles and the perception of speech and music, a natural question that arises pertains to commonalties and differences between profiles of these two sound classes. While temporal dynamics of speech are widely diverse and multiscale (e.g., variations across speakers, languages, prosodic profiles), variations in musical temporal patterns are even more diverse across genres, performances, and arrangements (Patel 2008). An analysis of modulation temporal profiles contrasting speech with Western musical samples shows drastic differences between these two sound classes (Fig. 12.3). This analysis, reproduced from (Ding et al. 2017), depicts temporal modulation profiles obtained by computing a discrete Fourier transform (DFT) of narrowband power envelope signals representing the root-mean-squared of the outputs of cochlear channels that correspond to four frequency bands. This processing contrasts the dynamics of a speech corpus, consisting of nine languages (American English, British English, Chinese, Danish, Dutch, French, German, Norwegian, and Swedish), against datasets of Western music samples that include classical music by single-voice string instruments and multi-voice instruments, symphonic ensembles, jazz, and rock (for details, see Ding et al. 2017). Immediately notable is the shift in the peak temporal modulation between speech and music. While speech has the now established peak around 4–8 Hz (typically attributed to physical dynamics of speech production articulators), the music dataset analyzed in this study shows visibly lower peaks with a plateau between 0.5–3 Hz. A number of physical and perceptual constraints can offer some explanations for the disparity. The kinematics of hand movements in music production (for the Western samples analyzed) impose a natural constraint on the temporal rates of movement with a preferred frequency of arm movements at around 1.5 Hz (Van Der Wel et al. 2009). There is also a relationship between emergent temporal modulations of music signals and underlying beats of the musical phrasing that also tend to highlight a rate of 1.5–3 Hz (van Noorden and Moelants 1999).

In addition to temporal modulation profiles, the distinction between speech and musical sounds is also very prominent with respect to their spectral profiles. Speech perception remains effective even over rather coarse sampling of the spectral axis. A case in point is the effectiveness of cochlear implants at conveying intelligible

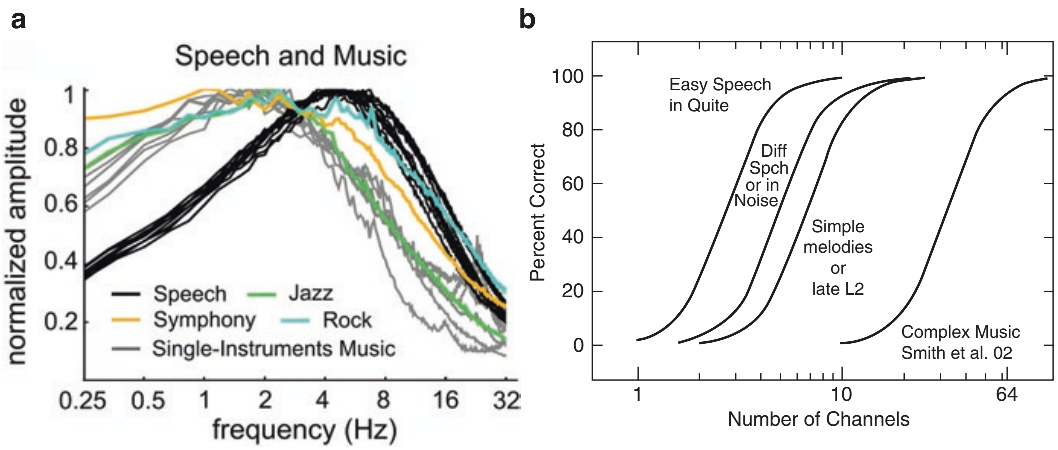

Fig. 12.3 Modulation profiles in speech and music. (**A**) The modulation spectrum of speech (*black*), single-instrument (*gray*), and multi-part music (*colors*). (**B**) Meta-analysis incorporating results across many studies to examine speech and music recognition (*y-axis*) as a function of the number of spectral channels (*x-axis*) in a noise band vocoder (**A** reprinted from Ding et al. 2017; **B** reprinted from Shannon 2005; both used with permission from Elsevier)

speech with very few channels, at least in favorable listening conditions (Wilson 2004). That is far from being the case for music perception (McDermott 2004), for which poor spectral resolution directly impacts melody recognition as well as timbre perception, two crucial aspects of the complex experience that constitutes music perception. Fig. 12.3 reproduces an illustration by Shannon (2005) that highlights the effects of spectral resolution on the perception of speech and music signals in addition to the effect of difficulty of listening. Panel B provides a meta-analysis across a number of studies that examine speech and music recognition rates as a function of the number of spectral channels in a noise-band vocoder. Speech detection in quiet listening conditions is contrasted with the same task under more challenging situations (including more difficult sentences, background noise, recognition in a second language, etc.). The trends show a clear need for improved spectral resolution under challenging conditions. This requirement for finer spectral resolution is further underscored when a task of melody recognition in the presence of competing melodies is used. This latter study results in the interesting contrast between speech versus melody identification: as low as 3 channels to achieve 75% correct identification of speech sentences in quiet listening conditions to as high as 40 channels to achieve 75% correct identification of melodies (Smith et al. 2002).

An interesting question regarding the distinction between spectral and temporal modulations of speech and music signals is how the perceptual system integrates across these modulation cues. For speech signals, joint spectrotemporal modulations capture temporal fluctuations of certain spectral peaks (e.g., formant transitions or speech glides). But work on automatic speech recognition suggests that joint spectrotemporal modulations are not necessary to improve recognition of words in the presence of distortions (Schädler and Kollmeier 2015).

These results argue that capturing signal transitions along both time and frequency may be less crucial for recognizing speech in noise. Instead, a reduced representation of spectral and temporal modulations (separately) is argued to yield comparable recognition as the joint-modulation representation. Unfortunately, there have been limited extensions of this exploration to definitely rule out a role of joint spectrotemporal modulations in speech recognition.

In contrast, the role of joint spectrotemporal modulations in musical timbre has been clearly demonstrated. There is strong evidence that a separable space, spanning time and frequency separately, is insufficient to capture the nuances of timbre required for distinguishing the timbre of different musical instruments. Instead, a modulation representation of both time and frequency axes is important to explicitly encode key musical constructs such as frequency modulations common in string vibrato (Patil et al. 2012; Elliott et al. 2013).

The divergence in acoustic attributes of both sound classes offers a potential rationale for different neural circuits that underlie the processing of speech and music in the brain (Zatorre et al. 2002; Norman-Haignere et al. 2015). The left hemisphere plays a more prominent role in complex linguistic functions; whereas, the right hemisphere appears to notably favor tasks involving tonal patterns or spectral processing, two aspects that are most related to the perception of music (Liégeois-Chauvel et al. 1998). This specialization beyond auditory cortex builds on an underlying common circuitry of mid-level and primary cortical representations that appear to focus primarily on extracting spectrotemporal modulations in incoming complex sound patterns. These very modulations appear to be a crucial backbone needed to carry information about complex sounds such as speech and music.

12.5 Summary

Theoretically, modulation is nothing but a mapping of an acoustic signal that highlights its fluctuations or indicates how its energy changes over time and frequency. These modulations are shaped by the source from which the signal emanates; hence, they can inform about the physics of that source and ultimately the signal's timbre. In practice, however, quantifying modulations is a nontrivial endeavor that takes many formulations and interpretations. Modulations of complex signals, such as speech and music, are a multifaceted construct that varies along multiple time scales and granularities, and they are shaped as much by the physics of the source as by the neural representations of acoustic energy in the brain. This chapter reviews some of the common representations of modulations and reflects on their perceptual and neural interpretation.

A number of questions surrounding the representation and role of modulations remain open. For example, what is the contribution of nonlinearities, which are pervasive in brain networks, in shaping the encoding of signal modulations in the auditory system? As discussed throughout this chapter, most constructs of modulations rely on transformation of the signal energy to another domain via spectral mappings such as the Fourier transform. These transformations maintain operations in the original vector space of time-frequency and, as such, are limited in their ability to manipulate or warp the mapping of spectrotemporal modulations. This is also true in the case of biomimetic constructs, such as spectrotemporal receptive fields, used to analyze neural activity in the central auditory system (Depireux and Elhilali 2013). While the receptive field view of auditory processing offers a rich set of tools to explore the encoding of sound characteristics, they are very much limited by approximative assumptions of linearity that are often compensated for in backend systems by means of nonlinear kernels that are often used in machine learning (Hemery and Aucouturier 2015). Understanding these nonlinearities is not only essential in the study and modeling of brain networks but also crucial to truly grasp the role played by sound modulations in informing perception.

The encoding of modulations is likely to be further shaped by active engagement in listening tasks and deployment of cognitive processes, notably attention. These top-down processes are known to greatly modulate neural encoding of incoming signals (Shamma and Fritz 2014), yet their role in shaping the representation of signal modulations remains largely unknown. Future research efforts addressing these questions will shed light on aspects of modulations that the brain hones in on when listening in multisource environments, for instance, their function in helping the auditory system deal with the cocktail party problem (Elhilali 2017).

Acknowledgements Dr. Elhilali's work is supported by grants from the National Science Foundation (NSF), the National Institutes of Health (NIH), and the Office of Naval Research (ONR).

Compliance with Ethics Requirements Mounya Elhilali declares she has no conflicts of interest.

References

Anden J, Mallat S (2014) Deep scattering spectrum. IEEE Trans Signal Process 62:4114–4128. https://doi.org/10.1109/TSP.2014.2326991

Arai T, Greenberg S (1998) Speech intelligibility in the presence of cross-channel spectral asynchrony. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, p 933–939

Arai T, Pavel M, Hermansky H, Avendano C (1999) Syllable intelligibility for temporally filtered LPC cepstral trajectories. J Acoust Soc Am 105:2783–2791

Athineos M, Ellis DPW (2003) Frequency-domain linear prediction for temporal features. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU):261–266

Attias H, Schreiner CE (1997) Temporal low-order statistics of natural sounds. In: Adv. Neural Inf. Proc. sys. (NIPS). MIT Press: Cambridge, MA, p 27–33

Carlin MA, Patil K, Nemala SK, Elhilali M (2012) Robust phoneme recognition based on biomimetic speech contours. In: Proceedings of the 13th annual conference of the international speech communication association (INTERSPEECH), p 1348–1351

Chen F, Jokinen K (2010) Speech technology: theory and applications, 1st edn. Springer, New York

Chi T, Gao Y, Guyton MC, Ru P, Shamma S (1999) Spectro-temporal modulation transfer functions and speech intelligibility. J Acoust Soc Am 106:2719–2732

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

Childers DG, Skinner DP, Kemerait RC (1977) The cepstrum: a guide to processing. Proc IEEE 65:1428–1443. https://doi.org/10.1109/PROC.1977.10747

Choi JE, Won JH, Kim CH, Cho Y-S, Hong SH, Moon IJ (2018) Relationship between spectro-temporal modulation detection and music perception in normal-hearing, hearing-impaired, and cochlear implant listeners. Sci Rep. 8(1). https://doi.org/10.1038/s41598-017-17350-w

Chowning JM (1973) The synthesis of complex audio spectra by means of frequency modulation. J Audio Eng Soc 21:1–10

Cohen L (1995) Time-frequency signal analysis, 1st edn. Prentice-Hall, Englewood Cliffs

Collins N (2009) Introduction to computer music, 1st edn. Wiley, Chichester/West Sussex

Croghan NBH, Duran SI, Smith ZM (2017) Re-examining the relationship between number of cochlear implant channels and maximal speech intelligibility. J Acoust Soc Am 142:EL537–EL543. https://doi.org/10.1121/1.5016044

deBoer E (1976) On the "residue" and auditory pitch perception. In: Keidel W, Neff D (eds) Auditory system (handbook of sensory physiology). Springer, Berlin, pp 479–583

Depireux DA, Elhilali M (eds) (2013) Handbook of modern techniques in auditory cortex. First. Nova Science Publishers, Inc., New York

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

Ding N, Patel AD, Chen L, Butler H, Luo C, Poeppel D (2017) Temporal modulations in speech and music. Neurosci Biobehav Rev 81:181–187

Divenyi P, Greenberg S, Meyer G (eds) (2006) Dynamics of speech production and perception. IOS Press, Amsterdam, p 388

Drullman R, Festen JM, Plomp R (1994) Effect of temporal envelope smearing on speech reception. J Acoust Soc Am 95:1053–1064

Dudley H (1939) Remaking speech. J Acoust Soc Am 11:169–177

Dudley H (1940) The carrier nature of speech. Bell Syst TechJ 19:495–513

Eggermont JJ (2001) Between sound and perception: reviewing the search for a neural code. Hear Res 157:1–42

Elhilali M (2017) Modeling the cocktail party problem. In: Middlebrooks J, Simon JZ, Popper AN, Fay RR (eds) The auditory system at the cocktail party. Springer, New York, pp 111–135

Elhilali M, Shamma S (2008) Information-bearing components of speech intelligibility under babble-noise and bandlimiting distortions. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), p 4205–4208

Elhilali M, Chi T, Shamma SA (2003) A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. Speech Commun 41:331–348. https://doi.org/10.1016/S0167-6393(02)00134-6

Elhilali M, Shamma SA, Simon JZ, Fritz JB (2013) A linear systems view to the concept of STRF. In: Depireux D, Elhilali M (eds) Handbook of modern techniques in auditory cortex. Nova Science Pub Inc, New York, pp 33–60

Elliott TM, Theunissen FE (2009) The modulation transfer function for speech intelligibility. PLoS Comput Biol 5:e1000302

Elliott TM, Hamilton LS, Theunissen FE (2013) Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. J Acoust Soc Am 133(1):389–404. https://doi.org/10.1121/1.4770244

Escabi MA, Read HL (2003) Representation of spectrotemporal sound information in the ascending auditory pathway. Biol Cybern 89:350–362

Freeman R (2004) Telecommunication system engineering, fourth edn. Wiley-Interscience, New York

Friesen LM, Shannon RV, Baskent D, Wang X (2001) Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. J Acoust Soc Am 110:1150–1163

Ganapathy S, Thomas S, Hermansky H (2010) Robust spectro-temporal features based on autoregressive models of Hilbert envelopes. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), p 4286–4289

Gill P, Zhang J, Woolley S, Fremouw T, Theunissen F (2006) Sound representation methods for spectro-temporal receptive field estimation. J Comput Neurosci 21:5. https://doi.org/10.1007/s10827-006-7059-4

Glasberg BR, Moore BC (1992) Effects of envelope fluctuations on gap detection. Hear Res 64:81–92

Gosselin F, Schyns PG (2001) Bubbles: a technique to reveal the use of information in recognition tasks. Vis Res 41(17):2261–2271. https://doi.org/10.1016/S0042-6989(01)00097-9

Greenberg S (2004) Temporal properties of spoken language. In: Proceedings of the international congress on acoustics. Kyoto, Japan, p 441–445

Greenberg S, Arai T (2001) The relation between speech intelligibility and the complex modulation spectrum. In: Proceedings of the 7th European conference on speech communication and technology (Eurospeech-2001), p 473–476

Grochenig K (2001) Foundations of time-frequency analysis. Birkhauser, Boston

Hemery E, Aucouturier J-J (2015) One hundred ways to process time, frequency, rate and scale in the central auditory system: a pattern-recognition meta-analysis. Front Comput Neurosci 9(80). https://doi.org/10.3389/fncom.2015.00080

Hepworth-Sawyer R, Hodgson J (2016) Mixing music, First edn. Routledge, New York/London

Hermansky H, Sharma S (1999) Temporal patterns (TRAPs) in ASR of noisy speech. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), p 292

Hintz M (2016) Digital speech technology:pProcessing, recognition and synthesis. Willford Press

Houtgast T, Steeneken HJM (1985) A review of MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J Acoust Soc Am 77:1069–1077

Houtsma AJM, Smurzynski J (1990) Pitch identification and discrimination for complex tones with many harmonics. J Acoust Soc Am 87:304–310

Ibrahim R, Bruce I (2010) Effects of peripheral tuning on the auditory nerve's representation of speech envelope and temporal fine structure cues. In: Lopez-Poveda EA, Palmer AR, MR (eds) The neurophysiological bases of auditory perception. Springer, New York, pp 429–438

Jepsen ML, Ewert SD, Dau T (2008) A computational model of human auditory signal processing and perception. J Acoust Soc Am 124:422–438

Katz M (2006) The violin: a research and information guide. Routledge Taylor and Francis Group, London/New York

Kingsbury B, Morgan N, Greenberg S (1998) Robust speech recognition using the modulation spectrogram. Speech Commun 25:117–132

Kleinschmidt M (2003) Localized spectro-temporal features for automatic speech recognition. In: Proceedings of Eurospeech, p 2573–2576

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

Leaver AM, Rauschecker JP (2010) Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J Neurosci 30:7604–7612

Qin Li, Les Atlas (2005) Properties for modulation spectral filtering. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP), p 521–524

Liégeois-Chauvel C, Peretz I, Babaï M, Laguitton V, Chauvel P (1998) Contribution of different cortical areas in the temporal lobes to music processing. Brain 121:1853–1867. https://doi.org/10.1093/brain/121.10.1853

Liu RC, Miller KD, Merzenich MM, Schreiner CE (2003) Acoustic variability and distinguishability among mouse ultrasound vocalizations. J Acoust Soc Am 114:3412–3422

Lyons RG (2011) Understanding digital signal processing, third edn. Prentice Hall, Upper Saddle River

McAuley J, Ming J, Stewart D, Hanna P (2005) Subband correlation and robust speech recognition. IEEE Trans Speech Audio Process 13:956–963. https://doi.org/10.1109/TSA.2005.851952

McDermott HJ (2004) Music perception with cochlear implants: a review. Trends Amplif 8:49–82

Meredith D (ed) (2016) Computational music analysis. Springer International Publishing, Cham

Meyer B, Ravuri S, Schaedler M, Morgan N (2011) Comparing different flavors of spectro-temporal features for ASR. In: Proceedings of the 12th annual conference of the international speech communication association (INTERSPEECH), p 1269–1272

Miller LM, Escabí MA, Read HL, Schreiner CE, Escabi MA, Read HL, Schreiner CE (2002) Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J Neurophysiol 87:516–527. https://doi.org/10.1152/jn.00395.2001

Moore BCJ (2003) An introduction to the psychology of hearing, 5th edn. Emerald Group Publishing Ltd, Leiden

Moore BCJ (2014) Auditory processing of temporal fine structure: Effects of age and hearing loss, 1st edn. World Scientific Publishing, Co, Hackensack/New Jersey

Morgan N, Chen BY, Zhu Q, Stolcke A (2004) Trapping conversational speech: extending TRAP/tandem approaches to conversational telephone speech recognition. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), p 40 vol.1

Moritz N, Anemuller J, Kollmeier B (2011) Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), p 5492–5495

Müller M (2015) Fundamentals of music processing. Springer International Publishing, Cham

Muller M, Ellis DPW, Klapuri A, Richard G (2011) Signal processing for music analysis. J IEEE, Sel Top Signal Process 5:1088–1110. https://doi.org/10.1109/JSTSP.2011.2112333

Nemala SK, Patil K, Elhilali M (2013) A multistream feature framework based on bandpass modulation filtering for robust speech recognition. IEEE Trans Audio Speech Lang Process 21:416–426. https://doi.org/10.1109/TASL.2012.2219526

Norman-Haignere S, Kanwisher NG, McDermott JH (2015) Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron 88:1281–1296. https://doi.org/10.1016/j.neuron.2015.11.035

Patel AD (2008) Music, language, and the brain, First edn. Oxford University Press, Oxford

Patil K, Pressnitzer D, Shamma S, Elhilali M (2012) Music in our ears: the biological bases of musical timbre perception. PLoS Comput Biol 8:e1002759. https://doi.org/10.1371/journal.pcbi.1002759

Peters RW, Moore BC, Baer T (1998) Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people. J Acoust Soc Am 103:577–587

Pickett JM (1999) The acoustics of speech communication: fundamentals, speech perception theory, and technology. Allyn & Bacon, Boston

Poeppel D, Idsardi WJ, van Wassenhove V (2008) Speech perception at the interface of neurobiology and linguistics. PhilosTransR Socl B BiolSci 363:1071–1086

Qin MK, Oxenham AJ (2003) Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. J Acoust Soc Am 114:446–454

Rabiner L, Schafer R (2010) Theory and applications of digital speech processing, First edn. Pearson, Upper Saddle River

Rosen S (1992) Temporal information in speech: acoustic, auditory and linguistic aspects. Philos Trans R Soc B-Biological Sci 336:367–373

Sadagopan S, Wang X (2009) Nonlinear spectrotemporal interactions underlying selectivity for complex sounds in auditory cortex. J Neurosci 29:11192–11202

Sadie S (ed) (2001) The new grove dictionary of music and musicians, Second edn. Macmillan, London

Santoro R, Moerel M, De Martino F, Goebel R, Ugurbil K, Yacoub E, Formisano E (2014) Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS Comput Biol 10(1). https://doi.org/10.1371/journal.pcbi.1003412

Schädler MR, Kollmeier B (2015) Separable spectro-temporal Gabor filter bank features: reducing the complexity of robust features for automatic speech recognition. J Acoust Soc Am 137:2047–2059. https://doi.org/10.1121/1.4916618

Schreiner C, Calhoun B (1995) Spectral envelope coding in cat primary auditory cortex: properties of ripple transfer functions. J Audit Neurosci 1:39–61

Schreiner CE, Sutter ML (1992) Topography of excitatory bandwidth in cat primary auditory cortex: single-neuron versus multiple-neuron recordings. J Neurophysiol 68:1487–1502

Schroeder M, Atal B (1985) Code-excited linear prediction(CELP): high-quality speech at very low bit rates. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP), p 937–940. doi: https://doi.org/10.1109/ICASSP.1985.1168147

Schuller B (2013) Applications in intelligent music analysis. Springer, Berlin/ Heidelberg

Shamma S, Fritz J (2014) Adaptive auditory computations. Curr Opin Neurobiol 25:164–168. https://doi.org/10.1016/j.conb.2014.01.011

Shamma S, Lorenzi C (2013) On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system. J Acoust Soc Am 133:2818–2833. https://doi.org/10.1121/1.4795783

Shannon RV (2005) Speech and music have different requirements for spectral resolution. Int Rev Neurobiol 70:121–134

Shannon R, Zeng F, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily temporal cues. Science 270:303–304

Singh N, Theunissen F (2003) Modulation spectra of natural sounds and ethological theories of auditory processing. J Acoust Soc Am 106:3394–3411

Smith ZM, Delgutte B, Oxenham AJ (2002) Chimaeric sounds reveal dichotomies in auditory perception. Nature 416:87–90. https://doi.org/10.1038/416087a

Steeneken HJ, Houtgast T (1979) A physical method for measuring speech-transmission quality. J Acoust Soc Am 67:318–326

Thoret E, Depalle P, McAdams S (2016) Perceptually salient spectrotemporal modulations for recognition of sustained musical instruments. J Acoust Soc Am 140(6). https://doi.org/10.1121/1.4971204

Turner RE, Sahani M (2011) Demodulation as probabilistic inference. IEEE Trans Audio Speech Lang Process 19(8):2398–2411

Van Der Wel RPRD, Sternad D, Rosenbaum DA (2009) Moving the arm at different rates: slow movements are avoided. J Mot Behav 42:29–36. https://doi.org/10.1080/00222890903267116

van Noorden L, Moelants D (1999) Resonance in the perception of musical pulse. J New Music Res 28:43–66. https://doi.org/10.1076/jnmr.28.1.43.3122

Venezia JH, Hickok G, Richards VM (2016) Auditory "bubbles": efficient classification of the spectrotemporal modulations essential for speech intelligibility. J Acoust Soc Am 140(2):1072–1088. https://doi.org/10.1121/1.4960544

Versnel H, Kowalski N, Shamma SA (1995) Ripple analysis in ferret primary auditory cortex. III. Topographic distribution of ripple response parameters. J Audit Neurosci 1:271–286

Wang TT, Quatieri TF (2012) Two-dimensional speech-signal modeling. IEEE Trans Audio Speech Lang Process 20:1843–1856. https://doi.org/10.1109/TASL.2012.2188795

Wilson BS (2004) Engineering design of cochlear implants. 20:14–52

Xu L, Pfingst BE (2003) Relative importance of temporal envelope and fine structure in lexical-tone perception. J Acoust Soc Am 114:3024–3027

Yang X, Wang K, Shamma SA (1992) Auditory representations of acoustic signals. IEEE Trans Inf Theory 38:824–839

Zatorre RJ, Belin P, Penhune VB (2002) Structure and function of auditory cortex: music and speech. Trends Cogn Sci 6:37–46

Zeng F-G, Nie K, Stickney GS, Kong Y-Y, Vongphoe M, Bhargave A, Wei C, Cao K (2005) Speech recognition with amplitude and frequency modulations. Proc Natl Acad Sci 102:2293–2298

Zhang X, Heinz MG, Bruce IC, Carney LH (2001) A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. J Acoust Soc Am 109:648–670