Neuroscience and Biobehavioral Reviews 81 (2017) 181–187

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

Review article

Temporal modulations in speech and music

Nai Ding[a,e,b,c,*], Aniruddh D. Patel[d,g], Lin Chen[e,a], Henry Butler[d], Cheng Luo[a], David Poeppel[e,f]

[a] *College of Biomedical Engineering and Instrument Sciences, Zhejiang University, China*
[b] *Interdisciplinary Center for Social Sciences, Zhejiang University, China*
[c] *Neuro and Behavior EconLab, Zhejiang University of Finance and Economics, China*
[d] *Department of Psychology, Tufts University, Medford, MA, United States*
[e] *Department of Psychology, New York University, New York, NY, United States*
[f] *Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany*
[g] *Azrieli Program in Brain, Mind, & Consciousness, Canadian Institute for Advanced Research (CIFAR), Toronto, Canada*

ARTICLE INFO

Article history:
Received 14 September 2016
Received in revised form 9 February 2017
Accepted 10 February 2017
Available online 14 February 2017

Keywords:
Speech
Music
Rhythm
Temporal modulations
Modulation spectrum

ABSTRACT

Speech and music have structured rhythms. Here we discuss a major acoustic correlate of spoken and musical rhythms, the slow (0.25–32 Hz) temporal modulations in sound intensity and compare the modulation properties of speech and music. We analyze these modulations using over 25 h of speech and over 39 h of recordings of Western music. We show that the speech modulation spectrum is highly consistent across 9 languages (including languages with typologically different rhythmic characteristics). A different, but similarly consistent modulation spectrum is observed for music, including classical music played by single instruments of different types, symphonic, jazz, and rock. The temporal modulations of speech and music show broad but well-separated peaks around 5 and 2 Hz, respectively. These acoustically dominant time scales may be intrinsic features of speech and music, a possibility which should be investigated using more culturally diverse samples in each domain. Distinct modulation timescales for speech and music could facilitate their perceptual analysis and its neural processing.

Contents

1. Introduction

Rhythmic structure is a fundamental feature of both speech and music. Both domains involve sequences of events (such as syllables, notes, or drum sounds) which have systematic patterns of timing, accent, and grouping (Patel, 2008). A primary acoustic correlate

* Corresponding author at: College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Hangzhou, 310027, China.
E-mail address: ding_nai@zju.edu.cn (N. Ding).

http://dx.doi.org/10.1016/j.neubiorev.2017.02.011
0149-7634/

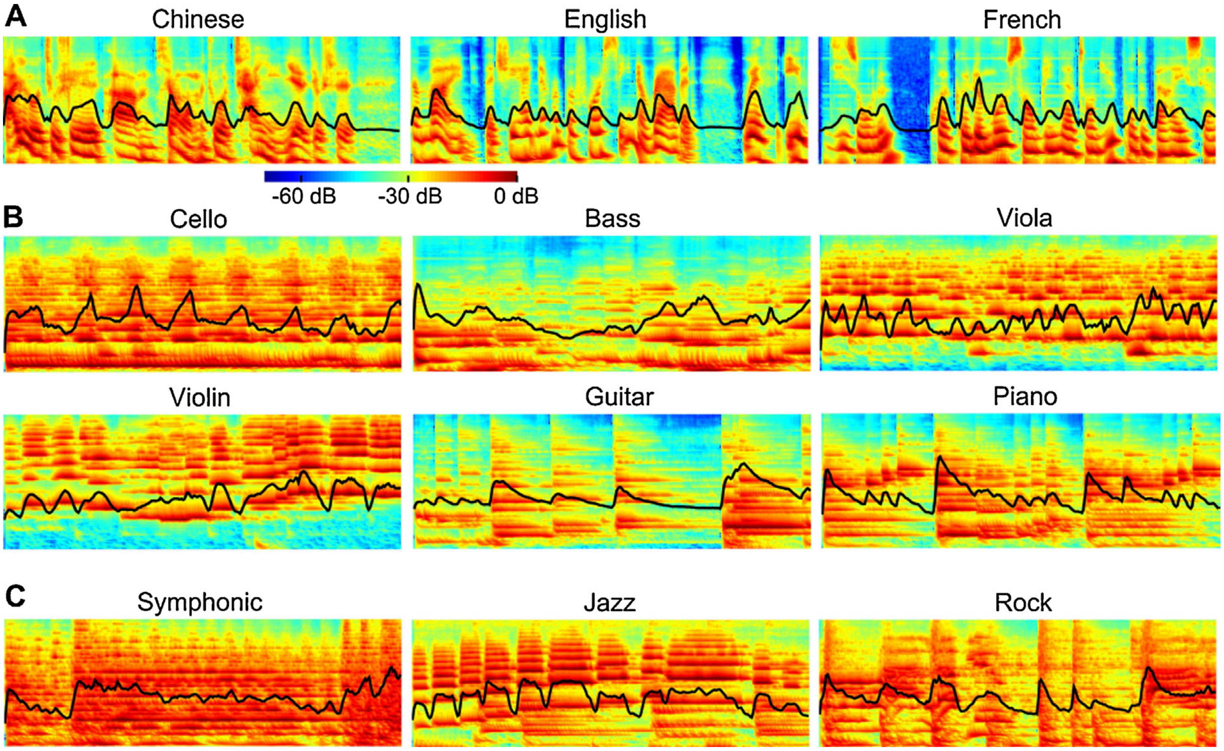

Fig. 1. Spectrograms of randomly chosen 3-s excerpts of speech (A), single-instrument music (B), and ensemble music (C). The spectrograms are simulated using a cochlear model (Yang et al., 1992). The x-axis denotes time (3 s) and the y-axis denotes frequency (180 Hz to 7.24 kHz, on a logarithmic scale). The amplitude of the spectrogram is represented in a logarithmic scale and the maximal amplitude is normalized to 0 dB. The spectrogram amplitudes summed over frequencies, which reflects how sound intensity fluctuates over time, is superimposed as black curves.

of perceived rhythm is the slow temporal modulation structure of sound, i.e. how sound intensity fluctuates over time (Fig. 1). For speech, temporal modulations below 16 Hz are related to the syllabic rhythm (Goswami and Leong, 2013; Greenberg et al., 2003) and underpin speech intelligibility (Drullman et al., 1994; Elliott and Theunissen, 2009; Shannon et al., 1995). For music, slow temporal modulations are related to the onsets and offsets of notes (or runs of notes in quick succession), which support perceptual phenomena such as beat, meter, and grouping (Gordon, 1987; Large and Palmer, 2002; Levitin et al., 2012; London, 2012; McKinney et al., 2007; Patel, 2008; Scheirer, 1998; Todd, 1994). Recently, a number of studies have investigated the neural activation patterns associated with these temporal modulations in the human brain and assessed their relevance to speech and music perception (Barton et al., 2012; Di Liberto et al., 2015; Ding and Simon, 2014; Doelling and Poeppel, 2015; Giraud and Poeppel, 2012; Hämäläinen et al., 2012; Henry and Obleser, 2012; Kerlin et al., 2010; Norman-Haignere et al., 2015; Nozaradan et al., 2011; Overath et al., 2015; Santoro et al., 2014; Schroeder et al., 2008; Steinschneider et al., 2013).

Although the importance of temporal modulations in speech and music has long been recognized, they have not been characterized in a consistent manner, leading to divergent reports (Zeng et al., 2004; Ghitza et al., 2013) and making it difficult to compare the modulation spectra across the speech and music materials analyzed in different studies. For example, some researchers emphasize that the temporal modulations in speech and music both have a characteristic 1/f spectrum (Voss and Clarke, 1975; Attias and Schreiner, 1997; De Coensel et al., 2003; Singh and Theunissen, 2003) while others argue that for speech, temporal modulations resonate around 4–5 Hz (Houtgast and Steeneken, 1985; Chi et al., 1999; Greenberg, 1999; Greenberg et al., 2003; Goswami and Leong, 2013). Here, we review different methods to calculate the modulation spectrum and compare the speech and music modulation properties using a consistent method.

2. Modulation spectrum analysis

The modulation spectrum is the spectrum of the temporal envelope of sound and reflects how fast sound intensity fluctuates over time. The modulation spectrum can be calculated using the procedure described in Fig. 2. The first step is to extract the sound envelope, which can be implemented in different ways. First, it can be extracted using either the broadband sound signal or the sound signal filtered into narrow frequency bands. Extracting sound envelopes in narrow frequency bands is a more commonly used approach and is easier to interpret neurophysiologically as will be discussed at the end of this section. Second, the sound envelope can be extracted by either applying the Hilbert transform or by rectifying and low-pass filtering the signal. The sound envelope extracted by these two methods are similar in general. Third, the amplitude of the extracted sound envelope can be rescaled using a static nonlinear function *f*(.), which could be a square function (Houtgast and Steeneken, 1985; Payton and Braida, 1999), a logarithmic function (Elliott and Theunissen, 2009), or more complex functions simulating response properties of the auditory nerve (Chi et al., 1999; Gill et al., 2006). The output of the nonlinear function is the temporal envelope.

The temporal envelope is converted into the frequency domain by the Discrete Fourier Transform (DFT). Spectral analysis of the temporal envelope reflects how fast sound intensity fluctuates over time. High modulation frequency corresponds to fast modulations and vice versa. Calculating absolute values of the DFT coefficients is equivalent to filtering the temporal envelope using a linearly spaced filter bank and calculating the root mean square value (RMS) of the output of each filter. A linearly spaced filter bank means that

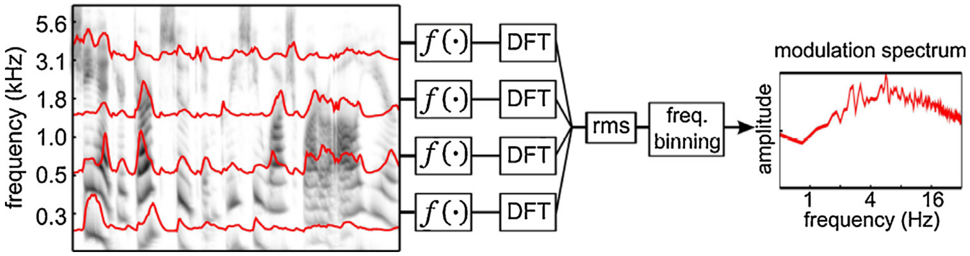

Fig. 2. Schematic illustration of how the modulation spectrum is calculated. The sound signal is first decomposed into narrow frequency bands using a cochlear model and the temporal envelope is extracted in each band (Yang et al., 1992). These operations create the spectrogram (gray-scale figure). Envelopes for 4 frequency bands are illustrated by red curves superimposed on the spectrogram. The nonlinear function *f*(.) could be a square function, a logarithmic function, or other functions. The root-mean-square (rms) of the Discrete Fourier Transform (DFT) of all narrowband power envelopes is the modulation spectrum.

the center frequency of the filters in the filter bank separates evenly in a linear frequency scale, i.e., in Hertz. Alternatively, if the filter bank is logarithmically spaced, i.e., the distance between center frequencies is constant in octaves, the frequency domain representation is the absolute value of the DFT coefficients weighted by the modulation frequency. This alternative method is the traditional way of calculating the modulation spectrum (Houtgast and Steeneken, 1985; Payton and Braida, 1999).

The modulation spectrum has a relatively straightforward neurophysiological interpretation. It can be viewed as an approximation of the spectrum of neural responses in auditory nerves or higher-level sub-cortical auditory nuclei. In the cochlea, sound signals are decomposed into narrow frequency bands and then, roughly speaking, each narrowband signal is half-wave rectified and low-pass filtered; these operations essentially extract the temporal envelope of the narrowband signal (Yang et al., 1992). Along the ascending auditory pathway, neural responses gradually lose sensitivity to fast temporal modulations (Joris et al., 2004) and become sparse in time (Delgutte et al., 1998; Schneider and Woolley, 2013). In general, however, in sub-cortical nuclei and even in primary auditory cortex, neurons relatively precisely follow the sound envelope below 30 Hz (Joris et al., 2004; Shamma, 2001). Therefore, the modulation spectrum (<30 Hz) can be roughly viewed as the spectrum of the neural responses summed over the whole neural population in a sub-cortical nucleus.

Additionally, if auditory cortical neurons are viewed as filters of the sound envelope, an influential and well supported hypothesis (Dau et al., 1997b; Shamma, 2001; Theunissen et al., 2001), the modulation spectrum can be roughly viewed as the activation level of neurons tuned to different modulation frequencies. There exists insufficient empirical evidence to determine whether the modulation tuning of cortical neurons is better approximated by a linearly spaced modulation filters (Dau et al., 1997a; Elliott and Theunissen, 2009; Singh and Theunissen, 2003) or logarithmically spaced modulation filters (Chi et al., 2005), especially for the lower frequency range below 10 Hz. A 1/f spectrum in the linear frequency scale corresponds to a flat spectrum in the logarithmic frequency scale. Therefore, it can be argued that a logarithmic frequency scale modulation spectrum describes how the modulation spectrum deviates from a 1/f spectrum, a general trend that natural sounds follow (Singh and Theunissen, 2003; Voss and Clarke, 1975).

3. Speech and music modulation spectra

3.1. Diversity and consistency in speech and music rhythms

Both speech and music have diverse rhythmic patterns. For example, speech rhythms differ between languages and have sometimes been classified into distinct rhythm categories such as ‘stress-timed’, ‘syllable-timed’, and ‘mora-timed’ (although there is debate whether rhythmic differences between languages are continuous rather than categorical (Turk and Shattuck-Hufnagel, 2013)). Furthermore, the details of speech rhythms vary across speakers: people speak at different rates and pause with different patterns. The rhythms of music are even more diverse, with patterns of tempo, grouping, and metrical structure varying dramatically across genres and performances (London, 2012). Indeed, the structural diversity of music across cultures and genres (e.g., ranging from an African drum ensemble to a modern piece of electronic music based on slowly changing noise textures with no clear rhythm) is arguably much greater than the acoustic diversity of spoken languages, creating a significant ‘sampling problem’ for any study of music’s modulation properties. The challenges of assembling a cross-cultural corpus of diverse musical genres was beyond scope of this study, and thus we focus on a more restricted subset of music, namely several genres of Western music varying in instrument type, ensemble size (solo or group), and presence/absence of vocals. We ask whether there are any consistencies in the modulations spectra of these musical samples, as a first step in characterizing the modulation spectra of music more generally. By reviewing and reanalyzing speech and music modulation properties using a consistent method, we explore whether acoustic rhythms constitute one such feature that can separate speech and music into two internally coherent categories, based on our current samples. If so, this would motivate a broader sampling of both languages and musics to see to what extent these findings generalize to other cultures.

3.2. Speech modulation spectrum

We plot the speech modulation spectrum for different languages and different speaking styles based on large corpora (Fig. 3). The analysis follows the procedure described in Fig. 2, with the nonlinear function being a square function. Long sound recordings are cut into 6-s duration segments. The modulation spectrum is calculated separately for each segment and then averaged over all segments. The analysis script is available in the supplementary materials. The spectrum shows a peak between 4 and 5 Hz, highly consistent for the 9 languages plotted (Fig. 3A). Additionally, the modulation spectrum is shown to be consistent across isolated sentences, audiobooks, and conversational speech in English (Fig. 3B). The averaged peak frequency of the speech modulation spectra is between 4.3 and 5.4 Hz for all tested speech materials (Fig. 3C and D).

3.3. Music modulation spectrum

The modulation spectrum of our Western music samples shows important differences from that of speech. The modulation spectra of classical music played by four single-voice string instruments (violin, viola, cello, and bass, which typically play one note at a time) are shown in Fig. 4A. Each shows a broad peak between 1

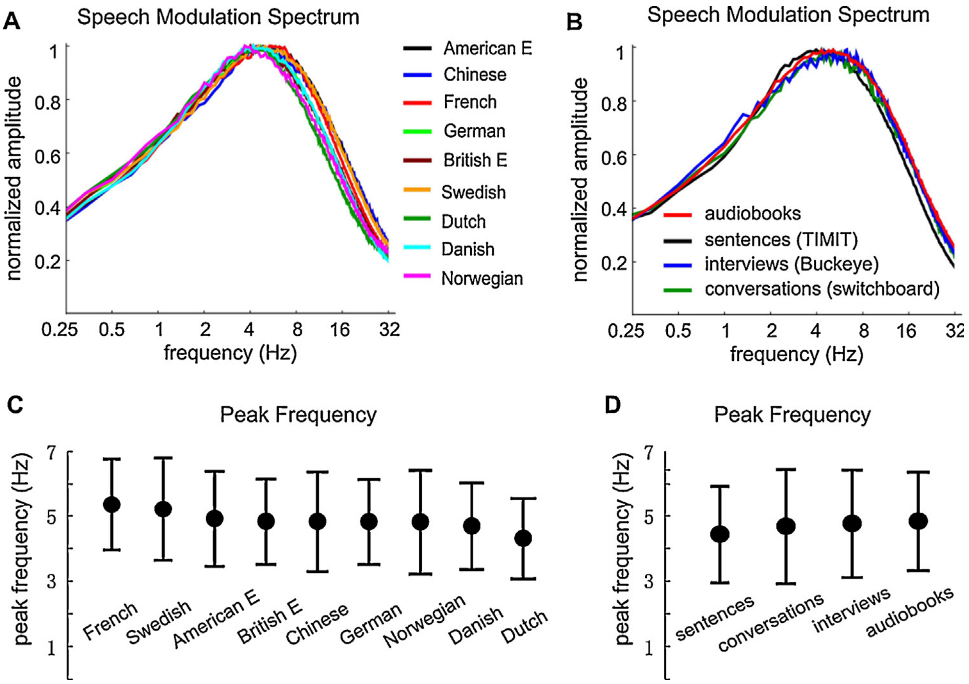

Fig. 3. The modulation spectrum of speech. A) The modulation spectrum for naturalistic, discourse-level speech across 9 languages. The modulation spectrum is consistent across languages and shows a peak near 4 Hz. Each spectrum is normalized by its peak amplitude. B) The modulation spectrum for four different corpora of American English. The modulation spectrum is consistent for speech produced in different contexts, including spontaneous and read speech. The peak frequencies for the modulation spectra are shown in C and D. The error bar represents one standard deviation on each side across all the 6-s duration speech recording segments tested for each type of material.

and 2 Hz, substantially lower than the 4–5 Hz peak frequency for speech. The modulation spectra for two multi-voice instruments (piano and guitar), which typically play more than one note at a time, are shown in Fig. 4B and do not differ substantially from the average modulation spectrum for the single-voice instruments. Across the six different solo instruments studied here, the modulation spectrum is largely independent of instrument at values below 8 Hz. The modulation spectra for viola and guitar show secondary peaks between 16 and 32 Hz, which may be related to vibratory properties of these instruments.

The modulation spectra of three types of Western ensemble music are shown in Fig. 4C. The average modulation spectrum of single-voice instruments is also shown for comparison. The modulation spectrum is generally consistent for all analyzed musical styles below 4 Hz, although symphonic music tends to contain more modulation energy at very low frequencies, below about 1.5 Hz. Above 4 Hz, the modulation spectra of rock and symphonic music contain more high-frequency energy than the modulation for single-voice instruments and jazz music (the latter two have very similar modulation spectra). For the musical recordings analyzed here, only rock music contains vocals, and therefore it is compared directly with speech in Fig. 4D. Rock music has a broad modulation peak around 2–3 Hz, with considerably more modulation energy than speech in lower frequencies and somewhat less modulation energy than speech between 2 and 16 Hz.

The modulation spectrum is directly compared for different languages and musical instruments/genres in Fig. 4D. The spectra show a striking difference: the modulation peak for music is evidently below 2 Hz while the peak for speech is evidently above 4 Hz. Fig. 4E shows that the modulation peak for speech (averaged over all materials) is consistently above the peak for music for all musical styles analyzed here. The modulation peak for rock music remains in the same range when the analysis is restricted to periods without vocals (for all the 6-s rock music segments being analyzed, 14% has no vocals, which is consistently rated by 3 people with amateur musical training). Fig. 4F shows the lowest and highest frequencies at which the modulation spectrum amplitude exceeds 75% of the peak amplitude. The difference between the highest and lowest frequencies can be viewed as a measure of the width of the modulation spectral peak. The highest frequency at which rock music modulation spectrum falls below 75% of the peak amplitude is higher than other musical styles, whether it contains vocals or not, but remains lower than speech. Preliminary analysis actually suggests that humans can classify speech and music purely based on temporal modulation information (see Supplementary Materials).

4. Modulation properties and time scales in speech and music processing

4.1. Time scales in speech processing

The modulation spectrum of speech, a purely acoustic metric, shows the greatest power between 2 and 10 Hz, peaking at around 5 Hz, a remarkably consistent pattern across many languages. Recent studies also show that the mean syllabic rate of speech, a linguistic metric, is 5–8 Hz across many languages (Pellegrino et al., 2011). These two perspectives indicate likely universal rhythmic properties of human speech. The 2–10 Hz rhythm is prevalent in other aspects of the speech communication chain, observed for example in motor cortex (Ruspantini et al., 2012) and articulator movements (Chandrasekaran et al., 2009; Ghazanfar et al., 2012) during speech production and in widely distributed cortical areas including auditory cortex during speech perception (Ding et al., 2016; Zion Golumbic et al., 2013). Therefore, the ∼5 Hz rhythm is likely an intrinsic attribute of speech, possibly imposed by the underlying neurodynamic properties of the speech production and perception systems and, building on that, the biomechanical properties of the human articulators (Chandrasekaran et al., 2009).

For speech, it has been proposed that the slow temporal modulations serve as acoustic landmarks to trigger an initial coarse analysis of speech features, which is followed by more fine-grained phonetic analysis (Stevens, 2002). Consistent with this hypothesis,

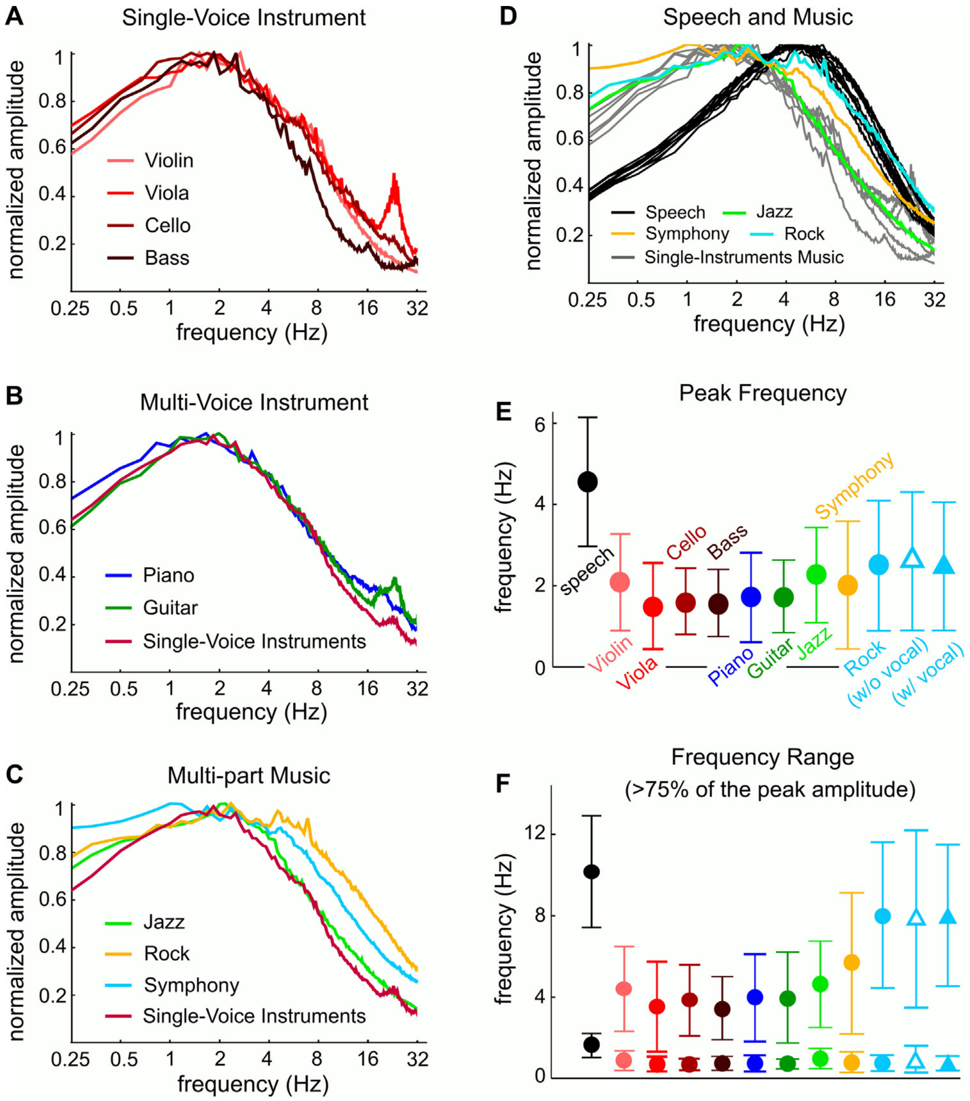

Fig. 4. Modulation spectra of Western music. A) The modulation spectrum of classical music played by single-voice string instruments consistently shows a peak below 2 Hz. B) The modulation spectrum of multi-voice instruments, e.g., piano and guitar, is consistent with that of the single-voice string instruments. C) The modulation spectrum of multi-part music, e.g. symphonic music, jazz, and rock, shows a broader peak than single-instrument music, especially for symphonic music and rock. D) The modulation spectrum of speech (reproduced from Fig. 3A), single-instrument music (reproduced from Fig. 4AB), and multi-part music (reproduced from Fig. 4C). Speech and music modulation spectra show distinct peak frequencies, and music contains more power at modulation frequencies below 4 Hz. E) The peak frequency of the modulation spectrum is consistently lower for music than speech. The error bar represents 1 standard deviation over all the 6-s duration musical recording segments analyzed for each type of material. For rock music, 6-s segments containing vocals and not containing vocals are separately analyzed. (14% of the segments has no vocals, which is consistently rated by 3 people with amateur musical training) F) The highest and lowest frequencies at which the modulation spectrum exceeds 75% of the peak amplitude.

it has been shown that neural activity in auditory cortex is synchronized to the slow temporal modulations of speech (Ding and Simon, 2014; Luo and Poeppel, 2007; Zion Golumbic et al., 2013). Since the slow temporal modulations correspond to the time scale of syllables, syllable-sized acoustic chunks have been proposed as the basic unit for initial speech analysis (Ghitza, 2013; Greenberg, 1999; Poeppel et al., 2008).

4.2. Time scales in music processing

The modulation spectrum of Western music also shows consistency across genres and musical instruments, despite differences in instrument and ensemble type in our sample. In all cases, the strongest activation is between 0.5 and 3 Hz, suggesting that the different musical instruments studied here, namely violin, viola, cello, bass, piano, and guitar, do not impose strong constraints on the slow temporal modulations of music. For ensemble music, jazz has a modulation spectrum very similar to that of single-instrument music, while symphonic and rock music show broader spectral speaks. For the music reviewed here, only rock music contains vocals, which may be one reason that its modulation spectrum is slightly closer to the modulation spectrum of speech.

Why would the modulation spectrum of music show a substantially lower peak than that for speech? One possibility is the nature of the ‘articulators’ involved. All of our musical samples involved instruments played with the hands/arms, and it is possible that the natural frequencies of these parts of the motor system are slower than the natural frequencies of speech articulators. Here ‘natural frequencies’ refer to frequency ranges where movement is most efficient, which depends on the lengths, weights, and synergies of articulators. For example, the arm’s preferred frequency is about 1.5 Hz (van der Wel et al., 2009), which is in the range of the modulation peak we found for music.

Another possibility is that the plateau of the musical modulation spectrum, i.e. 0.5–3 Hz, corresponds to the typical frequency range of musical beats. For example, dance music pieces tend to

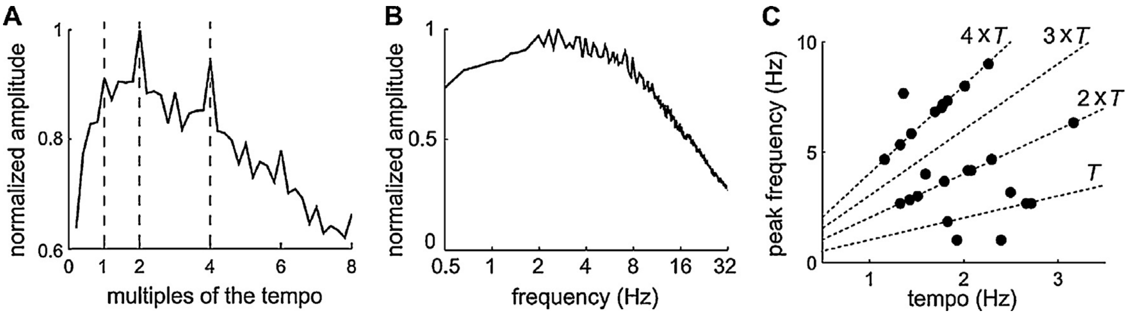

Fig. 5. Relationship between the modulation spectra and musical tempi for 25 musical excerpts with a steady beat. A) The modulation spectrum averaged over 25 excerpts. The modulation frequency axis is normalized based on the tempo of each excerpt. The average modulation spectrum shows peaks at the tempo rate and its multiples. The strongest peaks are seen at 2 times the tempo and 4 times the tempo. B) The modulation spectrum when the frequency axis is represented in Hz. C) The relationship between musical tempo and the frequency at which the modulation spectrum shows maximal power, for all 25 excerpts. If the music tempo is denoted as *T* and the modulation spectrum peak frequency is denoted as *F*, dotted lines are plotted for the relationships: *F* = *T*, *F* = 2*T*, *F* = 3*T*, and *F* = 4*T*. For most excerpts, the modulation spectrum peak frequency appears at 2*T* or 4*T*.

have tempi between 94 and 176 beats per minute (BPM), which corresponds to a rate of 1.6 Hz to 2.9 Hz for beats (Noorden and Moelants, 1999).

To further demonstrate how the modulation spectrum might be related to the rate of musical beats for individual musical pieces, we show the relationship between modulation peak and music tempo in Fig. 5, based on 25 selected excerpts of music with a clear beat (See Music Catalog in the Supplementary Materials). These excerpts were chosen from musical genres in which we could identify extended passages with a strong and steady beat: rock, funk, blues, and electronic music. In these 25 excerpts the modulation spectrum generally shows a broad peak centered around 1 and 2 times the tempo, with additional narrow peaks at 1, 2, and 4 times the tempo. This suggests that the modulation spectrum peak in music reflects acoustic fluctuations near the beat rate, with additional resonance reflecting a metrical subdivision of the beat, either one metrical level below the beat (2 × the beat) or two metrical levels below the beat (4 × the beat). Regular subdivision of the beat is a prominent feature of Western music, and research on musical meter suggests that having one or two levels of subdivision below the primary beat level is typical for Western music (London, 2012). The temporal modulations may provide an acoustic correlate of this musical phenomenon.

Beats are fundamental features organizing the temporal structure of much of Western music, including the genres studied here. Since the acoustically salient modulations correspond to the frequency range for common musical beats and subdivisions, it suggests that these rhythms possibly construct initial time scales for musical analysis (Farbood et al., 2013; Kraus and Slater, 2015), in parallel with the landmark hypothesis in speech (Stevens, 2002). Indeed, during music listening, cortical activity has been shown to be synchronized to the perceived musical beats (Nozaradan et al., 2011; Nozaradan et al., 2012).

5. Summary

We compare the slow temporal modulation properties of speech and music based on larger corpora than have been previously analyzed for this purpose. The speech recordings (over 25 h total) contain 9 languages and speech produced in different manners, e.g. during reading or telephone conversation. The music recordings (over 39 h of Western music total) contain single-instrument music from six musical instruments, as well as symphonic, jazz, and rock music. Based on these recordings, a high degree of consistency in the modulation spectra is found within the categories of speech and music – but not across them. These phenomena suggest that the statistical regularities of slow temporal modulations may be intrinsic signatures of speech and music more generally, a possibility that deserves investigation with broader cross-culturally array of linguistic and musical samples. In particular, it will be important to test to what extent the current conclusions generalize to languages with diverse prosodic characteristics (Jun, 2005) and to music from non-Western cultures (Brown and Jordania, 2013) and singing without instrumental accompaniment. Additionally, it is worth bearing in mind that the analysis here concerns only 'normal' speech. The modulation spectrum can certainly be affected by speech production pathologies (Falk et al., 2012) or when a talker speaks exceptionally clearly or slowly (Krause and Braida, 2004). It will also be interesting to use the methods presented here to study speech and music that are primarily rhythmic in nature, e.g. metrically regular poems and North Indian tabla music (Patel and Iversen, 2003; Turner and Pöppel, 1983), as well as music that contains no clear rhythmic beats e.g. Gregorian chant or Chinese Ch'in music (Van Gulik, 1969). Finally, we note that the distinctions between speech and music modulation properties reported here may be a productive attribute when addressing the distinctions between speech and music perceptual analysis.

Acknowledgements

We thank Erik Broess of Tufts University for help in selecting the music materials and Jess Rowland for help in creating the music catalog. Work supported by National Natural Science Foundation of China 31500873 (ND), Zhejiang Provincial Natural Science Foundation of China LR16C090002 (ND), Fundamental Research Funds for the Central Universities (ND), and NIH 2R01DC05660 (DP).

Appendix A. Supplementary data

Supplementary data associated with this article and the analysis script can be found, in the online version, at http://dx.doi.org/10.1016/j.neubiorev.2017.02.011.

References

Barton, B., Venezia, J.H., Saberi, K., Hickok, G., Brewer, A.A., 2012. Orthogonal acoustic dimensions define auditory field maps in human cortex. Proc. Natl. Acad. Sci. 109, 20738–20743.

Brown, S., Jordania, J., 2013. Universals in the world's musics. Psychol. Music 41, 229–248.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., Ghazanfar, A.A., 2009. The natural statistics of audiovisual speech. PLoS Comput. Biol. 5, e1000436.

Chi, T., Gao, Y., Guyton, M.C., 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, S.A., 2005. Multiresolution spectrotemporal analysis of complex sounds. J. Acoust. Soc. Am. 118, 887–906.

Dau, T., Kollmeier, B., Kohlrausch, A., 1997a. Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. J. Acoust. Soc. Am. 102, 2892–2905.

Dau, T., Kollmeier, B., Kohlrausch, A., 1997b. Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. J. Acoust. Soc. Am. 102, 2906–2919.

De Coensel, B., Botteldooren, D., Muer, T.D., 2003. 1/f noise in rural and urban soundscapes. Acta Acust.United Acust. 89, 287–295.

Delgutte, B., Hammond, B., Cariani, P., 1998. Neural coding of the temporal envelope of speech: relation to modulation transfer functions. In: Palmer, A., Reese, A., Summerfield, A., Meddis, R. (Eds.), Psychophysical and Physiological Advances in Hearing. Whurr Publishing, London, pp. 595–603.

Di Liberto, G.M., O'Sullivan, J.A., Lalor, E.C., 2015. Low-Frequency cortical entrainment to speech reflects phoneme-Level processing. Curr. Biol. 25, 2457–2465.

Ding, N., Simon, J.Z., 2014. Cortical entrainment to continuous speech: functional roles and interpretations. Front. Hum. Neurosci. 8, http://dx.doi.org/10.3389/fnhum.2014.00311.

Ding, N., Melloni, L., Zhang, H., Tian, X., Poeppel, D., 2016. Cortical tracking of hierarchical linguistic structures in connected speech. Nat. Neurosci. 19, 158–164.

Doelling, K.B., Poeppel, D., 2015. Cortical entrainment to music and its modulation by expertise. Proc. Natl. Acad. Sci. 112, E6233–E6242.

Drullman, R., Festen, J.M., Plomp, R., 1994. Effect of reducing slow temporal modulations on speech reception. J. Acoust. Soc. Am. 95, 2670–2680.

Elliott, T., Theunissen, F., 2009. The modulation transfer function for speech intelligibility. PLoS Comput. Biol. 5.

Falk, T.H., Chan, W.-Y., Shein, F., 2012. Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. Speech Commun. 54, 622–631.

Farbood, M.M., Marcus, G., Poeppel, D., 2013. Temporal dynamics and the identification of musical key. J. Exp. Psychol. Hum. Percept. Perform. 39, 911–918.

Ghazanfar, A.A., Takahashi, D.Y., Mathur, N., Fitch, W.T., 2012. Cineradiography of monkey lip-smacking reveals putative precursors of speech dynamics. Curr. Biol. 22, 2012.

Ghitza, O., 2013. The theta-syllable: a unit of speech information defined by cortical function. Front. Psychol. 4.

Ghitza, O., Giraud, A.-L., Poeppel, D., 2013. Neuronal oscillations and speech perception: critical-band temporal envelopes are the essence. Front. Hum. Neurosci., 6.

Gill, P., Zhang, J., Woolley, S.M., Fremouw, T., Theunissen, F.E., 2006. Sound representation methods for spectro-temporal receptive field estimation. J. Comput. Neurosci. 21, 5–20.

Giraud, A.-L., Poeppel, D., 2012. Cortical oscillations and speech processing: emerging computational principles and operations. Nat. Neurosci. 15, 511–517.

Gordon, J.W., 1987. The perceptual attack time of musical tones. J. Acoust. Soc. Am. 82, 88–105.

Goswami, U., Leong, V., 2013. Speech rhythm and temporal structure: converging perspectives? Lab. Phonol. 4, 67–92.

Greenberg, S., Carvey, H., Hitchcock, L., Chang, S., 2003. Temporal properties of spontaneous speech—a syllable-centric perspective. J. Phon. 31, 465–485.

Greenberg, S., 1999. Speaking in shorthand—a syllable-centric perspective for understanding pronunciation variation. Speech Commun. 29, 159–176.

Hämäläinen, J.A., Rupp, A., Soltész, F., Szücs, D., Goswami, U., 2012. Reduced phase locking to slow amplitude modulation in adults with dyslexia: an MEG study. Neuroimage 59, 2952–2961.

Henry, M.J., Obleser, J., 2012. Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. Proc. Natl. Acad. Sci. 109, 20095–20100.

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

Joris, P.X., Schreiner, C.E., Rees, A., 2004. Neural processing of amplitude-modulated sounds. Physiol. Rev. 84, 541–577.

Jun, S.-A., 2005. Prosodic typology. In: Jun, S.-A. (Ed.), Prosodic Typology: The Phonology of Intonation and Phrasing. Oxford University Press, Oxford, UK, pp. 430–458.

Kerlin, J.R., Shahin, A.J., Miller, L.M., 2010. Attentional gain control of ongoing cortical speech representations in a cocktail party. J. Neurosci. 30, 620–628.

Kraus, N., Slater, J., 2015. Music and language: relations and disconnections. In: Celesia, G.G., Hickok, G. (Eds.), The Human Auditory System. Fundamental Organization and Clinical Disorders.

Krause, J.C., Braida, L.D., 2004. Acoustic properties of naturally produced clear speech at normal speaking rates. J. Acoust. Soc. Am. 115, 362–378.

Large, E.W., Palmer, C., 2002. Perceiving temporal regularity in music. Cogn. Sci. 26, 1–37.

Levitin, D.J., Chordia, P., Menon, V., 2012. Musical rhythm spectra from Bach to Joplin obey a 1/f power law. Proc. Natl. Acad. Sci. 109, 3716–3720.

London, J., 2012. Hearing in Time. Oxford University Press, New York.

Luo, H., Poeppel, D., 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54, 1001–1010.

McKinney, M.F., Moelants, D., Davies, M.E., Klapuri, A., 2007. Evaluation of audio beat tracking and music tempo extraction algorithms. J. New Music Res. 36, 1–16.

Noorden, L.V., Moelants, D., 1999. Resonance in the perception of musical pulse. J. New Music Res. 28, 43–66.

Norman-Haignere, S., Kanwisher, N.G., McDermott, J.H., 2015. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron 88, 1281–1296.

Nozaradan, S., Peretz, I., Missal, M., Mouraux, A., 2011. Tagging the neuronal entrainment to beat and meter. J. Neurosci. 31, 10234–10240.

Nozaradan, S., Peretz, I., Mouraux, A., 2012. Selective neuronal entrainment to the beat and meter embedded in a musical rhythm. J. Neurosci. 32, 17572–17581.

Overath, T., McDermott, J.H., Zarate, J.M., Poeppel, D., 2015. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nat. Neurosci. 18, 903–911.

Patel, A.D., Iversen, J.R., 2003. Acoustic and perceptual comparison of speech and drum sounds in the north indian tabla tradition: an empirical study of sound symbolism. In: Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, pp. 925–928.

Patel, A.D., 2008. Music, Language, and the Brain. Oxford University Press, New York, NY.

Payton, K.L., Braida, L.D., 1999. A method to determine the speech transmission index from speech waveforms. J. Acoust. Soc. Am. 106, 3637–3648.

Pellegrino, F., Coupé, C., Marsico, E., 2011. Across-language perspective on speech information rate. Language 87, 539–558.

Poeppel, D., Idsardi, W.J., Wassenhove, V.v., 2008. Speech perception at the interface of neurobiology and linguistics. Phil. Trans. R. Soc. B: Biol. Sci. 363, 1071–1086.

Ruspantini, I., Saarinen, T., Belardinelli, P., Jalava, A., Parviainen, T., Kujala, J., Salmelin, R., 2012. Corticomuscular coherence is tuned to the spontaneous rhythmicity of speech at 2–3 Hz. J. Neurosci. 32, 3786–3790.

Santoro, R., Moerel, M., Martino, F.D., 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, e1003412.

Scheirer, E.D., 1998. Tempo and beat analysis of acoustic musical signals. J. Acoust. Soc. Am. 103, 588–601.

Schneider, D.M., Woolley, S., 2013. Sparse and background-invariant coding of vocalizations in auditory scenes. Neuron 79, 141–152.

Schroeder, C.E., Lakatos, P., Kajikawa, Y., Partan, S., Puce, A., 2008. Neuronal oscillations and visual amplification of speech. Trends Cogn. Sci. 12, 106–113.

Shamma, S., 2001. On the role of space and time in auditory processing. Trends Cogn. Sci. 5, 340–348.

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

Singh, N.C., Theunissen, F.E., 2003. Modulation spectra of natural sounds and ethological theories of auditory processing. J. Acoust. Soc. Am. 114.

Steinschneider, M., Nourski, K.V., Fishman, Y.I., 2013. Representation of speech in human auditory cortex: is it special? Hear. Res., 57–73.

Stevens, K.N., 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. J. Acoust. Soc. Am. 111, 1872–1891.

Theunissen, F.E., David, S.V., Singh, N.C., Hsu, A., Vinje, W.E., Gallant, J.L., 2001. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Netw. Comput. Neural Syst. 12, 289–316.

Todd, N.P.M., 1994. The auditory primal sketch: a multiscale model of rhythmic grouping. J. New Music Res. 23, 25–70.

Turk, A., Shattuck-Hufnagel, S., 2013. What is speech rhythm? A commentary on Arvaniti and Rodriquez, Krivokapić, and Goswami and Leong. Lab. Phonol. 4, 93–118.

Turner, F., Pöppel, E., 1983. The neural lyre: poetic meter, the brain, and time. Poetry 142, 277–309.

Van Gulik, R.H., 1969. The Lore of the Chinese Lute. Sophia University.

van der Wel, R.P., Sternad, D., Rosenbaum, D.A., 2009. Moving the arm at different rates: slow movements are avoided. J. Mot. Behav. 42 (1), 29–36.

Voss, R.F., Clarke, J., 1975. 1/f 'noise' in music and speech. Nature 258, 317–318.

Yang, X., Wang, K., Shamma, S.A., 1992. Auditory representations of acoustic signals. IEEE Trans. Inf. Theory 38, 824–839.

Zeng, F.-G., Nie, K., Liu, S., Stickney, G., Rio, E.D., Kong, Y.-Y., Chen, H., 2004. On the dichotomy in auditory perception between temporal envelope and fine structure cues (L). J. Acoust. Soc. Am. 116, 1351.

Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z., Poeppel, D., Schroeder, C.E., 2013. Mechanisms underlying selective neuronal tracking of attended speech at a cocktail party. Neuron 77, 980–991.