A cross-linguistic study of speech modulation spectra

Léo Varnet, Maria Clemencia Ortiz-Barajas, Ramón Guevara Erra, Judit Gervain, Christian Lorenzi

A cross-linguistic study of speech modulation spectra

Léo Varnet,[1,a)] Maria Clemencia Ortiz-Barajas,[2] Ramón Guevara Erra,[2] Judit Gervain,[2] and Christian Lorenzi[1]

[1]*Laboratoire des Systèmes Perceptifs, Département d'Études Cognitives, École normale supérieure, PSL Research University, CNRS, 29 rue d'Ulm, 75005 Paris, France*

[2]*Laboratoire Psychologie de la Perception, Centre National de la Recherche Scientifique, UMR 8242, Université Paris-Descartes, 45 rue des Saints Pères, 75006 Paris, France*

(Received 23 December 2016; revised 9 August 2017; accepted 22 September 2017; published online 11 October 2017)

Languages show systematic variation in their sound patterns and grammars. Accordingly, they have been classified into typological categories such as stress-timed vs syllable-timed, or Head-Complement (HC) vs Complement-Head (CH). To date, it has remained incompletely understood how these linguistic properties are reflected in the acoustic characteristics of speech in different languages. In the present study, the amplitude-modulation (AM) and frequency-modulation (FM) spectra of 1797 utterances in ten languages were analyzed. Overall, the spectra were found to be similar in shape across languages. However, significant effects of linguistic factors were observed on the AM spectra. These differences were magnified with a perceptually plausible representation based on the modulation index (a measure of the signal-to-noise ratio at the output of a logarithmic modulation filterbank): the maximum value distinguished between HC and CH languages, with the exception of Turkish, while the exact frequency of this maximum differed between stress-timed and syllable-timed languages. An additional study conducted on a semi-spontaneous speech corpus showed that these differences persist for a larger number of speakers but disappear for less constrained semi-spontaneous speech. These findings reveal that broad linguistic categories are reflected in the temporal modulation features of different languages, although this may depend on speaking style. © *2017 Acoustical Society of America*. https://doi.org/10.1121/1.5006179

[JFL] Pages: 1976–1989

I. INTRODUCTION

The languages of the world show systematic variation in many of their syntactic, prosodic and phonological properties. While this variation is increasingly well described and understood from a linguistic point of view,[1] how it is reflected in the speech signal remains, to a large extent, unexplored. Yet, typically developing infants rely on this acoustic signal to learn the abstract lexical and grammatical regularities of their native language. It has been proposed that there may be significant correlations between certain grammatical properties and some acoustic features of the speech signal, and that young learners might use precisely these to break into language.[2–5] For instance, prominence in phonological phrases is carried by different acoustic features in languages with varying word orders: languages with a preposition-noun order use duration, i.e., lengthening, on the prominent element, the noun (e.g., Italian: *a Ro:ma* [in Rome]), while languages with noun-postposition orders typically use higher pitch or intensity (e.g., Japanese: *ˆTokyo ni* [Tokyo to]).

In the current study, we tested to what extent languages that show systematic typological variation in their linguistic rhythm and their basic word order may differ in their basic acoustic properties. Over the last century, a wealth of studies has promoted the view of speech as a modulated carrier signal.[6–8] According to this "modulation theory" of speech, speech sounds are seen as a sum of carrier signals produced by the vocal folds, the amplitude and frequency of which change slowly as a consequence of the dynamic changes of the vocal tract during phonation.[6] (To avoid confusion with the notion of "audio frequency," we will use the term "rate" to refer to the frequency of these fluctuations hereafter.) The present study focuses on low-rate ($\lesssim$50 Hz) amplitude and frequency modulations (AM and FM). The AM component relates to the fluctuations of the temporal envelope of the speech signal such as those produced by alternating high-energetic and silent segments, whereas the FM component is related to the temporal fine structure of the speech signal, primarily due to the fluctuations of the f_0 and the harmonic structure. These two components play an important role in accounting for speech intelligibility. The AM component is believed to be crucial for intelligibility, especially for speech presented in silence.[9–12] Furthermore, low-rate FM cues have been shown to play a role in intelligibility when speech is presented together with competing voices.[56,57]

The modulation information contained in a given speech signal can be characterized by the modulation spectrum,[13] i.e., the amount of modulation as a function of modulation rate. The speech modulation spectra (AM spectrum and FM spectrum) can therefore be understood as a general description of the temporal structure of the speech signal at different time scales. This description of temporal information in speech is reminiscent of the original hierarchies proposed by Rosen,[14] and more recently by Leong *et al.*[15] and Giraud and

a)Electronic mail: leo.varnet@ens.fr

Poeppel,[16] characterizing several temporal features based on dominant fluctuation rates, each feature having distinct roles in linguistic contrasts. Speech cues are typically divided into, at least, three time scales: stress rate (1–2 Hz), syllable rate (2–8 Hz), and phoneme rate (8–40 Hz), although the exact boundaries of these time scales may vary between studies.

The AM spectrum may be computed in at least two ways, denoted "AMa spectrum" and "AMi spectrum" hereafter. AMa spectrum is the long-term Fourier amplitude spectrum of the temporal envelopes across several frequency bands.[17] The AMa spectrum computed for a corpus of 300 monosyllabic English words shows a peak at 2 Hz followed by a steady decline in amplitude.[17] Other natural sounds, such as animal vocalizations or environmental sounds, show a similar decrease in amplitude in high modulation frequencies. However, the change of slope for low modulation frequencies seems to be specific to speech sounds.[18–21]

Another, more perceptually-based way of representing the AM information contained in a speech signal is the AMi spectrum.[8] In this case, the envelope in each frequency band is analyzed through a $\frac{1}{3}$-octave filterbank in the modulation domain. The obtained spectrum is then normalized by the mean value of the envelope in the frequency band. Figure 1 presents a schematic diagram of this approach. Compared to the AMa spectrum, the $\frac{1}{3}$-octave-band representation used in the AMi spectrum emphasizes components in the high modulation rate region, where the bandwidths of modulation filters are larger. This is motivated by the psychophysical demonstration of broad frequency selectivity in the AM domain in the human auditory system.[22–25] Furthermore, the output is expressed as a modulation index (depending on the ratio between the intensities of the target and the noise) instead of an absolute value.[12] As a consequence, this representation is closely linked to observed behavioral performance in the perceptual tasks. Decline in speech intelligibility caused by noise, fast-acting amplitude compression, or reverberation has been found to be systematically associated with strong changes in the AMi spectrum of the degraded speech stimuli.[8,12,26,27] In other words, the extent to which speech intelligibility is affected by a given transmission system depends on how well the AMi spectrum of speech is preserved. AMi spectra have been widely used in the past and several languages have been described, including English,[28–33] Dutch,[8,12,26,33–35] Japanese,[31] German,[33,36] and more recently Chinese, French, Swedish, Dutch, Danish, and Norwegian.[33] These previous investigations suggest that the AMi spectra of speech in these different languages are globally similar in shape: they exhibit a marked peak around 4 Hz and contain almost all of their energy in the range from 0 to 30 Hz. However, there have been very few attempts to *compare* AMa or AMi spectra across languages.

Yet, there are reasons to believe that the AM content of speech may differ cross-linguistically, as languages differ systematically in some of the features that may influence the AM spectrum, such as speech rhythm or phrasal prosody. Languages have intuitively been described to fall into one of three rhythmic classes on the basis of what perceptually

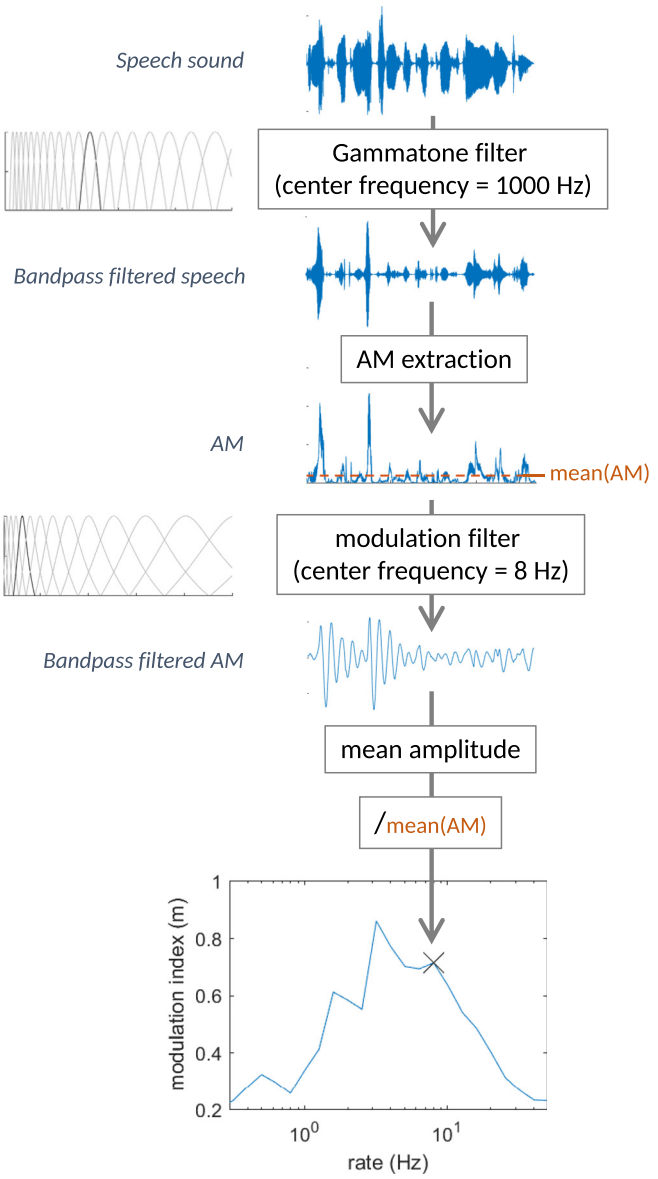

FIG. 1. (Color online) Schematic diagram of the calculation of the modulation index from a speech sound for a given frequency channel (center frequency = 1000 Hz) and a given modulation channel (center frequency = 8 Hz).

appears to be the isochronous unit in their speech signal. In what linguists traditionally categorized as syllable-timed languages, such as French, Italian, or Spanish, the organizing time unit was believed to be the syllable. In stress-timed languages, like English, Dutch, or Arabic, the unit of isochrony was assumed to be the interstress interval. And in mora-timed languages, like Japanese and Tamil, the unit was believed to be the mora. These differences are easy to perceive. Even newborn infants are able to discriminate two unfamiliar languages, as long as they belong to two different rhythmic classes, but not if they belong to the same class.[4,37–39]

While rhythm-based language discrimination is well established, it is not clear what acoustic features of the signal correspond to the percept of rhythm or what theoretical

notions best describe it. The isochrony principle was not supported by empirical evidence,[40] and subsequent findings suggested that rhythmicity is better understood as a continuum, not as discrete categories. An operational definition was suggested by Ramus *et al.*,[41] who introduced two measures, $\%V$, i.e., the relative length of vocalic space in the speech signal, and ΔC, i.e., the variability in the length of consonant clusters, to characterize languages. When using these dimensions, it was found that languages traditionally described as mora-timed have high $\%V$ and low ΔC, syllable-timed languages have high to medium values for both, while stress-timed languages have relatively low $\%V$ and high ΔC values. The rhythm class hypothesis has since then been revisited and reframed in different ways in the literature, and has even been criticized. In addition to the measures provided by Ramus *et al.*,[41] several other metrics have been proposed,[42–45] and important points like distinguishing rhythm from timing or speaking rate[46–48] have been raised. These theoretical and methodological points notwithstanding, the phenomenological difference in "rhythm" across languages remains. It is thus not implausible to assume that they might be reflected in the AMi spectra of different languages, as they are related to the rate and structure of syllables, i.e., the 2–8 Hz peak in the AMi spectrum.[27] Specifically, we hypothesized that rhythmic differences between languages will result in changes in the most prominent AM and FM modulation rates, possibly with somewhat faster modulations for stress-timed as opposed to syllable-timed languages. This difference is predicted on the basis of the greater variability in syllable length and thus the presence of a greater number of shorter syllables in stress-timed languages. This hypothesis has already been explicitly stated by Goswami and Leong[32] and Leong *et al.*[15]

Another linguistic property that might impact the AM spectrum is the basic word order. The basic word order of a language, i.e., the relative order of syntactic Heads and their Complements (e.g., the Verb and its Object, adpositions and their nouns), has systematic prosodic correlates at the level of the phonological phrase. In Head-Complement (HC) languages, such as English, French, Italian or Spanish, that place the Verb in front of its Complement (e.g., *eat apples*), the prominence in phonological phrases is final, i.e., it falls on the Complement, and it is acoustically realized as increased duration. By contrast, in Complement-Head (CH) languages, such as Basque, Japanese or Turkish, the prosodic prominence is phrase-initial, since the Complement is initial (e.g., Hungarian: *almát eszik* [apple.acc eat]), and it is marked by increased intensity and/or pitch.[49,50] As it is known that the regular alternation of prominent and weak elements is reflected in the AM component, word order may well be reflected in the AM spectra of languages. Specifically, we hypothesized that the type of basic word order will affect both the FM and the AM modulation spectrum, as the prosodic patterns correlated with CH vs HC languages prosody are carried by pitch/intensity and duration, respectively. We thus predicted that HC languages might show stronger AM modulation than CH languages, while CH languages may exhibit stronger FM modulation. Furthermore, rhythm and basic word order are typologically correlated.[3,51] Languages with high $\%V$ values tend to have CH word order, e.g., Basque and Japanese, while languages with lower $\%V$ are typically HC. Linguistic features might thus have acoustic signatures on multiple accounts.

To the best of our knowledge, only two studies attempted to compare the AMi spectra of several languages.[31,33] The first study, conducted by Arai and Greenberg,[31] compared the AMi spectra of English, a stress-timed HC language, and Japanese, a mora-timed CH language, using semi-spontaneous speech materials (OGI multilanguage corpus, Switchboard corpus). They found that the AMi spectra were overall comparable, maybe due to a similar amount of variability in the syllable duration in both languages. However, subtle differences between English and Japanese AMi are observable in Figs. 2 and 3 of Arai and Greenberg.[31] Since the study did not report any statistical comparisons between the spectra, the significance of these differences remains to be understood. The second study conducted by Ding *et al.*[33] on a wider range of languages, compared the AMi spectra of stress-timed HC languages (English, Swedish, German, Dutch, Danish, Norwegian), and syllable-timed HC (French) or syllable-timed CH (Chinese) languages. This study used four different speech corpora of semi-spontaneous (Buckeye corpus, Switchboard corpus) or connected (audiobooks, TIMIT corpus) speech. They found that, consistent with Arai and Greenberg,[31] the AMi spectral were overall comparable. However the AMi spectra were normalized by their maximum values, precluding the observation of amplitude differences. No differences in AMi peak rates across the 9 languages and 4 corpora were reported.

Temporal modulations in speech are not limited to the AM component. Indeed, speech sounds are also modulated in frequency,[6,52–54] the frequency-modulation (FM) components reflecting at least partly the slow fluctuations in the fundamental frequency (f_0) of the speaker (the acoustic correlate of voice pitch) and its harmonics. Moreover, AM and FM components of band-limited signals are not independent and are, therefore, expected to covary to some extent in speech signals.[17,55] Sheft *et al.*[17] computed FM spectra for continuous speech using American-English material, and found FM spectra similar (in terms of peak rate) to the AM spectra computed for the same material. They interpreted this result as being due to covariations of the AM and FM components in band-limited signals. This is particularly true for speech, where slow fluctuations in both f_0 and temporal envelope are believed to convey prosodic information assisting speech segmentation and syntactic parsing.[58–62] As discussed above, prominence in the phonological phrase in CH languages is mainly carried by voice pitch or intensity. We thus expect the above-discussed typological difference in word order (HC/CH) to potentially impact the FM and even more specifically the f_0 spectrum, with CH languages having stronger modulations than HC languages.

Given the above cross-linguistic predictions, the present study compared the AM and FM spectra across a large corpus of speech recordings from ten languages, including stress-timed, syllable-timed and mora-timed languages, as well as HC and CH languages. Unlike in previous studies, a single corpus of well-controlled (semi-read) connected

TABLE I. Summary of the characteristics of the SRS corpus. Data for $\%V$, ΔC and slope Q10/f_c were provided by the authors of Mehler, Sebastian-Galls, and Nespor (Ref. 3), Ramus, Nespor, and Mehler (Ref. 41), Guevara Erra and Gervain (Ref. 64), Molnar, Carreiras, and Gervain (Ref. 65). S.D. represents standard deviation.

Language	Linguistic characteristics	Number of speakers	Number of stimuli (N)	Stimuli duration (mean ± S.D.)	mean $\%V$	mean $\Delta C \cdot 100$	slope Q10/f_c
Basque	CH, syllable-timed	4	162	2.8 s ± 0.5 s	48.0	4.41	—
Dutch	HC, stress-timed	4	228	3.0 s ± 0.4 s	42.3	5.33	1.04
English	HC, stress-timed	4	153	3.0 s ± 0.4 s	40.1	5.35	1.02
French	HC, syllable-timed	4	216	2.9 s ± 0.4 s	43.6	4.39	—
Japanese	CH, mora-timed	4	212	2.8 s ± 0.4 s	53.1	3.56	0.73
Marathi	CH, syllable-timed	2	80	3.3 s ± 0.4 s	51.5	4.30	0.61
Polish	HC, syllable-timed	4	216	3.3 s ± 0.6 s	41.0	5.14	0.82
Spanish	HC, syllable-timed	4	212	3.4 s ± 0.5 s	43.8	4.74	—
Turkish	CH, syllable-timed	4	160	3.0 s ± 0.3 s	48.4	5.15	0.78
Zulu	HC, syllable-timed, click	4	158	3.6 s ± 0.7 s	—	—	1.11

speech recordings was used here. We computed both AM spectra (AMa and AMi) previously described. We also calculated two types of FM spectra for each language. The first was computed using a technique similar to that proposed by Sheft *et al*. described above. The second one, called hereafter "f_0 modulation spectrum" (f_0M spectrum), was restricted to the modulation components of the f_0 contour of the speech material extracted using the YIN algorithm developed by de Cheveigné and Kawahara.[63]

The effects of speakers and speaking style might be potential confounds in the analysis of cross-liguistic differences. For example, Arvaniti[48] has shown that the type of speech materials used (read sentences, read stories, spontaneous speech) has a strong influence on standard rhythm metrics. For this reason, an additional corpus of semi-spontaneous speech produced by a large number of speakers has been used in a second analysis to test for the generalizability of our results across speaker numbers and speech styles.

II. MATERIALS AND METHODS

All analyses were conducted in MATLAB R2016b (The Mathworks, Natick, MA).

A. Semi-Read Speech (SRS) corpus

The main corpus used in this study, referred to hereafter as the "semi-read speech (SRS) corpus," comprises speech samples from ten languages, most of them having already been used in previous studies: Dutch, English, French, Japanese, Polish and Spanish,[41] Marathi and Turkish,[3,64] Basque,[65] and Zulu.

The stimuli were produced according to the same principles, described in detail in Ramus, Nespor, and Mehler:[41] sentences systematically varied between 15 to 21 syllables, were uttered separately by four female native speakers of each language (only two speakers for Marathi) and sampled at 16 kHz. Speakers first read a given sentence silently, then uttered it out loud for recording, introducing a certain degree of spontaneity in the production.

All stimuli were normalized in root-mean-squared (rms) power before further analysis. It must be noted that all recordings were made under similarly good recording conditions (sound-proof or sound-attenuated booth, high quality recording equipment etc.), thus it is highly implausible that the signal-to-noise ratio (SNR) played a role in the obtained results.

The characteristics of the SRS corpus are presented in Table I.

B. Semi-Spontaneous Speech (SSS) corpus

An additional corpus of semi-spontaneous speech produced by a large number of speakers has been used in a second analysis to test for the generalizability of our results across speaker numbers and speech styles. This corpus will be referred to as "SSS corpus" hereafter.

This corpus consisted of responses of about 100 speakers to a semi-directed interview conducted by telephone (OGI Multilanguage corpus[66]). Four of the languages from the original SRS corpus were included: English, French, Japanese, and Spanish. This choice of languages was ideal as the four languages were different from one another in their rhythms and word orders. A first condition (fixed vocabulary, FV) corresponded to the speakers naming their native or common language. The second condition (topic-specific, TS) was collected in response to the prompts "Tell us something that you like about your hometown" and "Tell us about the climate in your hometown." The two conditions differed with respect to their semantic content, syntactic complexity, and stimulus duration. The FV condition corresponded to short utterances of about 1 s (e.g., "my native language is English" or "English"), whereas the TS condition corresponded to longer, more complex utterances of about 5–8 s. Even though FV and TS corresponded to the forms of spontaneous speech, the TS utterances were less homogeneous in terms of syntactic structure and lexical content. These recordings were made over the telephone, so the overall quality was very poor.

The characteristics of the SSS corpus are presented in Table II.

C. Calculating the modulation spectra

A block diagram of the algorithm used in this study is shown in Fig. 2.

TABLE II. Summary of the characteristics of the SSS corpus. S.D. represents standard deviation.

		FV condition			TS condition		
Language	Linguistic characteristics	Number of speakers	Number of stimuli (N)	stimuli duration (mean ± S.D.)	Number of speakers	Number of stimuli (N)	stimuli duration (mean ± S.D.)
English	HC, stress-timed	198	383	1.1 s ± 0.3 s	200	393	6.7 s ± 3.3 s
French	HC, syllable-timed	122	235	1.3 s ± 0.5 s	111	218	7.3 s ± 1.9 s
Japanese	CH, mora-timed	106	201	1.4 s ± 0.4 s	96	182	6.0 s ± 2.4 s
Spanish	HC, syllable-timed	125	243	1.2 s ± 0.4 s	120	237	7.1 s ± 2.1 s

The first step in the estimation of the AM and FM spectra consisted in band-pass filtering the stimuli through a bank of 30 gammatone filters, defined with the same parameters as in Hohmann.[67] An advantage of the implementation proposed by Hohmann[67] is that it provides an analytical, complex output which simplifies further processing. The filters were 1-ERB-wide and their center frequencies were equally spaced on the ERB scale (i.e., quasi-logarithmically spaced on the frequency scale) between 70 Hz and 6700 Hz with a density of 1 gammatone filter per ERB.[67] The envelope and temporal fine structure (TFS) of the resulting narrow-band signals were extracted separately from each complex gammatone response. The AM component (envelope) thus corresponds to the magnitude of the analytic signal, whereas the TFS corresponds to its unwrapped instantaneous phase.[53,68] The FM component is obtained by taking the time-derivative of the TFS. Unfortunately, in practice, this is a problem for speech as the TFS is erratic and often meaningless in low-AM segments (e.g., in silent intervals between two words), leading to an overestimation of high FM rates. Therefore, in a second step, these low-energetic segments were removed from the FM component and were replaced with NaNs for all time points where the envelope of the signal was below a threshold of 0.05 (−13.0 dB rms). A similar correction was used in the frequency amplitude modulation encoding (FAME) algorithm developed by Nie *et al.*[69]

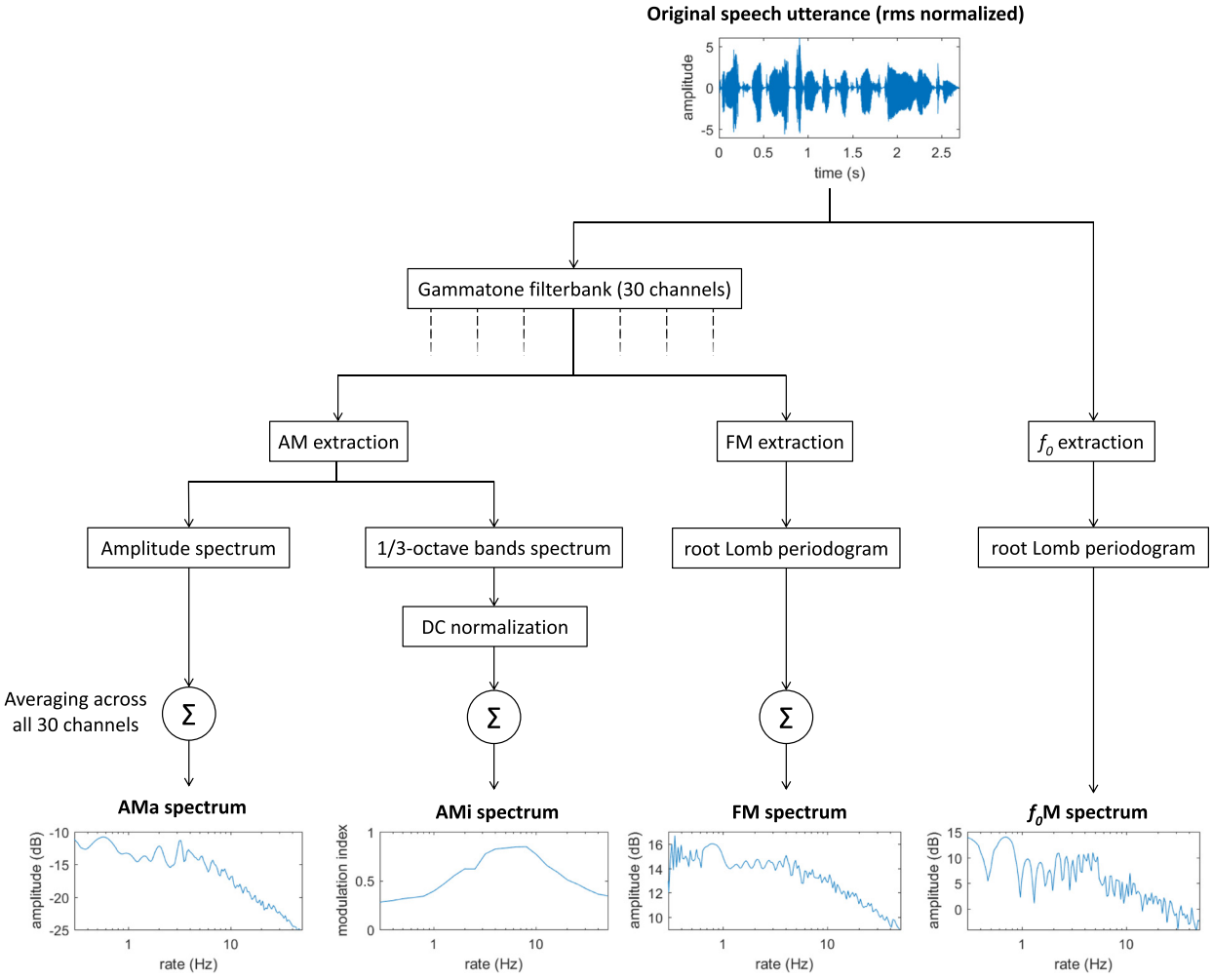

FIG. 2. (Color online) Schematic diagram of AMa, AMi, FM, and f_0M spectra calculation from a speech signal.

Three spectra were thus calculated in each analysis band. The AMa spectrum was obtained directly from the amplitude of the discrete Fourier transform of the AM component. For the FM component, however, some data points were removed (see above), making it impossible to use the Fourier spectrum. It was therefore necessary to calculate the FM spectrum by taking the root of the Lomb periodogram (also called Least-Square Spectrum), a generalization of the Fourier spectrum in the case of partially undefined functions.[70] For the AMi spectrum, a $\frac{1}{3}$-octave band spectrum was obtained by decomposing the AM component using a bank of $\frac{1}{3}$-octave wide 1st order Butterworth bandpass filters overlapping at -3 dB (center frequencies between 0.1 Hz and 410 Hz), and by taking the rms amplitude of the filtered output multiplied by a factor of $\sqrt{2}$. For each filter, a modulation index was calculated by dividing the output by the mean amplitude of the AM component for the speech sample in a given gammatone filter.[8] Finally, the 30 band-specific AMi spectra (respectively AMa and FM spectra) were averaged to generate a single AMi spectrum per utterance (respectively AMa and FM spectra). Such averaging is made possible by the great consistency of the shape of modulation spectra across frequency bands.[27,32]

A f_0 modulation spectrum (referred to as f_0M spectrum below) was also calculated for each utterance by first extracting the discontinuous f_0 contour from the speech signal using the YIN algorithm.[63] Only voiced segments (detected through an aperiodicity measure) longer than 20 ms were considered here, with upper and lower search bounds of 50 Hz and 600 Hz, respectively. Then the root Lomb periodogram of the f_0 contour was computed, as in the FM case. It must be noticed that YIN performs each single estimation on a time interval equal to the integration window size (20 ms in our case) plus the f_0 period. Therefore, for a fundamental frequency of 50 Hz or more, temporal fluctuations above 20 Hz should not be considered as reliable.

All spectra were resampled by averaging the values in logarithmically-spaced frequency segments with logarithmically increasing widths, before statistical analysis. Average modulation spectra across utterances were derived for each speaker and each language, for the purpose of visualization.

The analysis on the SSS corpus aimed at confirming the results obtained on the SRS corpus. For this reason, only the AMi spectra were calculated.

D. Statistical analysis

Statistical analyses were conducted on several relevant characteristics extracted from the spectra of individual utterances taken from the SRS corpus as dependent variables (see below). The comparison was done by means of a mixed model including a "language rhythm" factor (stress-timed vs syllable- and mora-timed), a "basic word order" factor (HC languages vs CH languages) and a random effect of speaker. The significance of the effects was then tested with a likelihood ratio test by comparing the likelihood of the model to that of the nested model with no linguistic factors (i.e., only intercept and speaker effect).

When necessary, *post hoc* analyses were carried out to disentangle the respective effect of each linguistic parameter by testing a mixed model including only one single factor and the random effect.

We used the following dependent variables. The "constant bandwidth" spectra (i.e., AMa, FM and f_0M spectra) were characterized by their high-frequency and low-frequency slopes and their mean amplitude in the 2–8 Hz range (corresponding to a period duration of 0.125–0.5 s, i.e., approximately one longer or two–three shorter syllables). Slightly narrowing or widening the range does not affect the results considerably. The slopes were estimated by linear fitting on the 0.3–1.5 Hz and 10–50 Hz ranges, respectively, and expressed in dB/oct. For the f_0M and FM spectra, only the high frequency slope was measured because of the absence of a clear linear part in the low frequencies. AMi spectra were characterized by the value and position of their maximum.

To compare the modulation results with existing phonological and acoustic measures derived from these same recordings, we calculated correlations, wherever appropriate, between our dependent variables and the rhythm metrics $\%V$ and ΔC[41] as well as with a measure of optimal, non-redundant coding $Q10/f_c$.[64] This latter measure was obtained by calculating maximally independent representations for speech in different languages using independent component analysis (ICA) and characterizing the obtained filter populations by their sharpness (Q10) as a function of their central frequency (f_c). For more details, see Guevara Erra and Gervain.[64] Note that these phonological and statistical measures were not available for all ten languages tested here, so the correlations were run on smaller sets of languages.

Because of the restricted number of languages under study in the SSS corpus (English, French, Spanish and Japanese), *t*-test comparisons were carried out on pairs of languages according to the results obtained on the SRS corpus, instead of a mixed model.

III. RESULTS

The present study sought to determine whether speech modulation spectra reflect systematically varying linguistic characteristics. To facilitate the comparison, Fig. 3 presents all modulation spectra gathered with a color code highlighting the linguistic characteristics of the languages in the SRS corpus, whereas Fig. 4 shows the normalized AM, FM and f_0M separately for each language.

A. Results for the AMa spectra

The average high-frequency slope of the AMa spectra across languages was -3.9 dB/oct. There was no significant effect of any of the linguistic factors on this dependent variable in the mixed effect model [$\chi^2(2) = 0.77, p = 0.68$]. The mean low-frequency slope across all languages was -0.71 dB/oct. The effect of linguistic factors on this dependent variable was marginally significant [$\chi^2(2) = 5.85, p = 0.054$], mainly because of the effect of syllable-timed languages being slightly steeper at low rates ($p = 0.029$).

The mixed effect model ran on the averaged (log) AMa in the 2–8 Hz range revealed a significant difference between

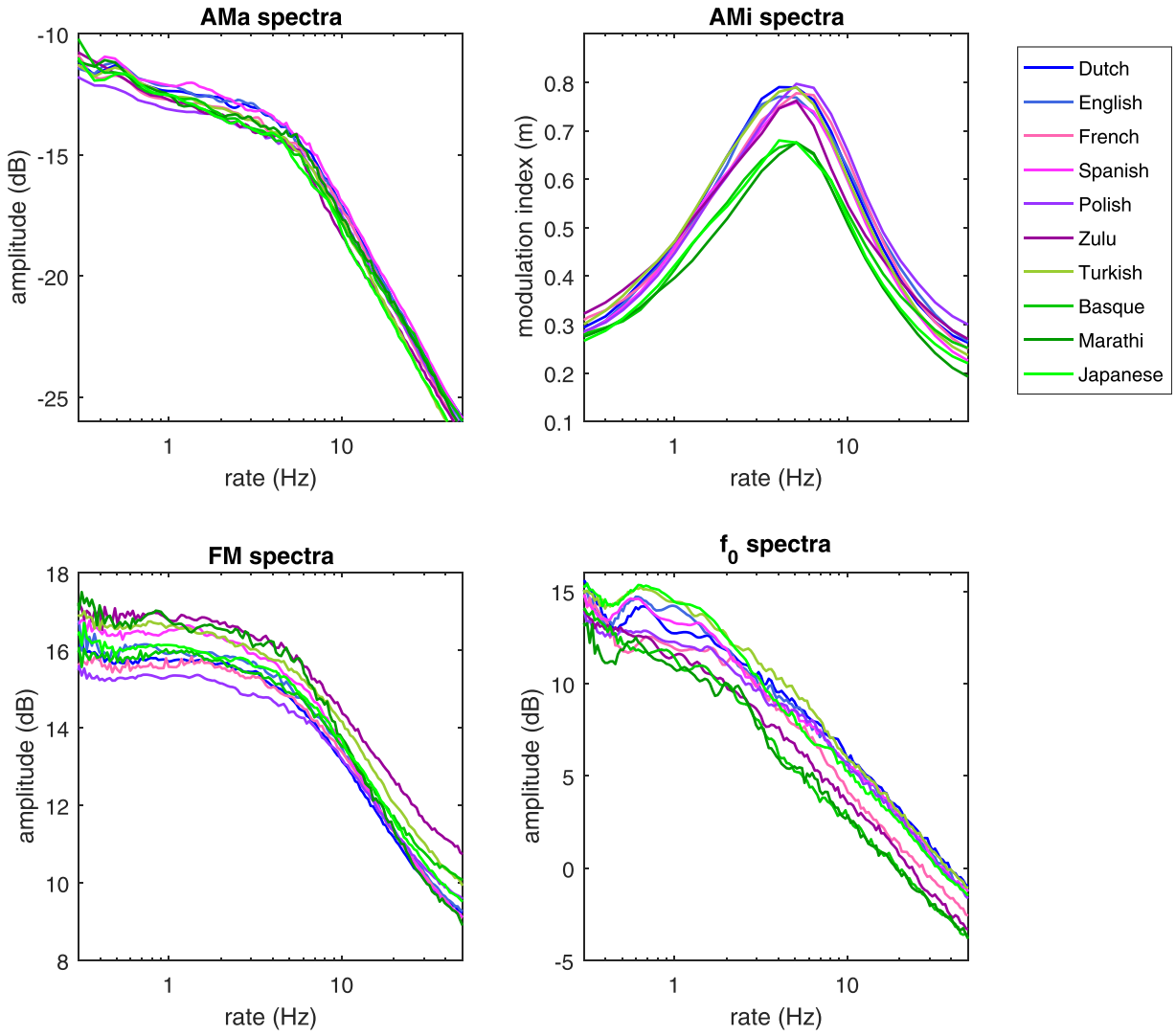

FIG. 3. (Color online) Averaged modulation spectra for all languages of the SRS corpus. Blue lines: HC, stress-timed languages; indigo lines: HC, syllable-timed languages; green lines: CH, syllable-timed or mora-timed languages.

linguistic groups [$\chi^2(2) = 9.52, p = 8.6 \times 10^{-3}$]. This effect was mainly due to the "language rhythm" factor, with stress-timed languages showing higher amplitude in this region than syllable-timed languages ($p = 5.3 \times 10^{-3}$). However, the amount of AMa in the 2–8 Hz range was not significantly correlated with $\%V$, ΔC or Q10/f$_c$ ($p > 0.1$).

B. Results for the AMi spectra

The mixed effect model assessing the maximum values of the AMi spectra showed a significant difference among linguistic groups [$\chi^2(2) = 13.90, p = 9.6 \times 10^{-4}$], as for AMa, the difference being in this case magnified by the $\frac{1}{3}$-octave splitting and the normalization by the mean of the band-limited envelope. This effect was due to the "basic word order" parameter ($p = 2.1 \times 10^{-4}$). There was a significant correlation between average maximum AMi value and $\%V$ [$r(7) = -0.78, p = 0.012$], ΔC [$r(7) = 0.77, p = 0.014$], but not slope Q10/f$_c$ [$r(5) = 0.66, p = 0.11$].

The mixed effect model assessing the location of the peak was also significantly affected by language group [$\chi^2(2) = 15.018, p = 5.5 \times 10^{-4}$], due to an effect of the "language rhythm" factor ($p = 5.1 \times 10^{-3}$). A *post hoc* analysis restricted to HC languages only, showed that maxima for syllable-timed languages were slightly shifted towards higher rates (5.24 Hz on average), compared to those of stress-timed languages (4.40 Hz on average) [$\chi^2(1) = 11.52, p = 6.9 \times 10^{-4}$]. However, this measure was not significantly correlated with $\%V$, ΔC or Q10/f$_c$ (all $p > 0.4$).

These two measures thus defined three groups of languages: (i) French, Spanish, Polish, Zulu, i.e., HC, syllable-timed languages (ii) Dutch, English, and Turkish, i.e., HC, stressed-timed languages with the exception of Turkish, and (iii) Marathi, Japanese, and Basque.

C. Results for the FM and f$_0$M spectra

Although the FM and f$_0$M spectra had clearly different slopes, we found the f$_0$M to match almost perfectly the FM spectra of the lowest gammatones (70–300 Hz).

Both the FM and f$_0$M spectra were characterized by their high-frequency slopes (on average −1.8 dB/oct and −2.9 dB/oct, respectively) and their mean log-amplitude value in the 2–8 Hz range. There was no significant

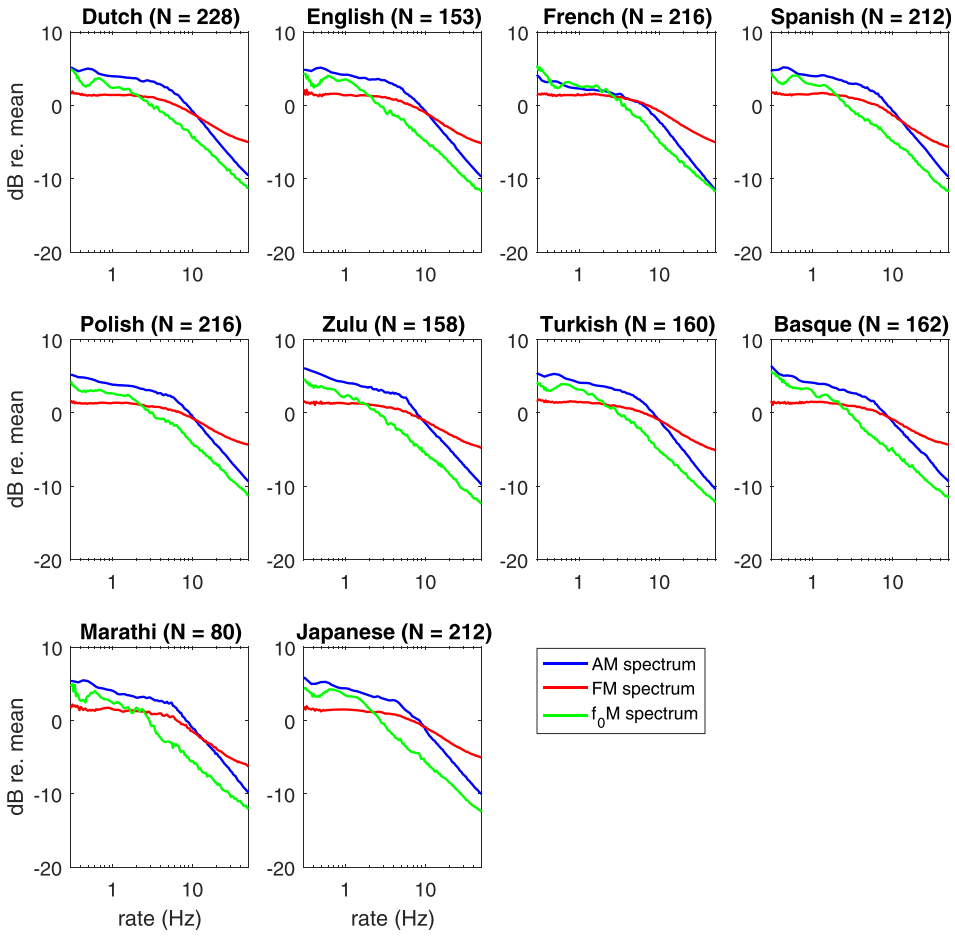

FIG. 4. (Color online) "Constant-bandwidth" modulation spectra for all languages of the SRS corpus, averaged across speakers. Blue lines: AMa spectra; red lines: FM spectra; green lines: f_0M spectra. All spectra are plotted in dB relative to their mean amplitude in the 0.3–50 Hz range.

difference between linguistic groups, either for the slope of the f_0M spectra [$\chi^2(2) = 5.5, p = 0.07$] or for that of the FM spectra [$\chi^2(2) = 0.038, p = 0.98$].

The mixed effect model assessing the mean amplitude of the FM and f_0M spectra in the 2–8 Hz range showed no significant difference [$\chi^2(2) = 1.96, p = 0.38$ and $\chi^2(2) = 3.3$, $p = 0.19$, respectively].

D. AMi results for the SSS corpus

In an attempt to replicate the above AMi cross-linguistic differences, we carried out a complementary analysis on a second corpus (SSS corpus) that was split into two subsets of speech materials (FV and TS conditions). Because of the restricted number of languages under study (English, French, Spanish, and Japanese), t-test comparisons were carried out on pairs of languages according to the above results. The properties under testing were (1) the maximum value between HC and CH languages (i.e., English vs Japanese, French vs Japanese, and Spanish vs Japanese) and (2) the peak position between HC stress-timed and HC syllable-timed languages (i.e., English vs French and English vs Spanish).

All tested pairwise comparisons were significant in the FV condition (all $p < 5.5 \times 10^{-5}$), but did not reach the significance level in the TS condition (all $p > 0.2$).

IV. DISCUSSION

The aim of the present study was to determine whether different groups of languages can be distinguished on a purely acoustic basis in the modulation domain. On the basis of previous work on the role of temporal modulations in speech intelligibility, it was hypothesized that features derived from the speech amplitude- or frequency-modulation spectra could reflect linguistic parameters such as speech rhythm and basic word order. We therefore measured the AMa, AMi, FM and f_0M spectra for a large corpus of sentences comprising ten languages (SRS corpus) as well as for a smaller corpus of four of the ten languages with a larger number of speakers and a different speech style (SSS corpus).

A. General similarities across languages

Overall, the "constant bandwidth" modulation spectra (AMa, FM and f_0M spectra) turned out to be very similar across languages (Fig. 4). They were all low-pass in shape, reflecting the fact that speech signals mostly comprise relatively slow temporal modulations. This is a common property for all natural sounds.[19,20,71] However, unlike music or environmental sounds, speech modulation spectra have been shown to plateau or even decrease at low rates,[17,18,21] which is consistent with Fig. 3. Furthermore, the cut-off rate appears to be very similar between the AMa and FM spectra as already discussed by Sheft *et al.*[17]

The disparity between the FM and f_0M spectra for each language may seem surprising, given that these two characteristics relate to the TFS of the signal. However, when restricting our analysis to the gammatone filters tuned

between 70 and 300 Hz, corresponding roughly the frequency range of the f_0 contour, FM spectra were found to be highly similar to the f_0M spectrum. This observation confirms that for speech, FM information is partly based on the relatively slow (<5 Hz) f_0 modulations.

Sheft *et al.*[17] calculated the AM and FM spectra of 300 monosyllabic English words using an approach similar to the present one. The comparison of their spectra with the AMa and FM spectra obtained here for English sentences (second panel of Fig. 4) makes it clear that stimulus type (words vs sentences) has an impact on the shape of the modulation spectrum, with the average duration of short isolated words generating low-rate modulations (at around 2 Hz) in the analysis performed by Sheft *et al.*[17] In the present study, stimuli were longer and probably more variable in duration (see Table I), resulting in a low-pass shape for the average AMa and FM spectra. This interpretation is supported by the fact that, using 15-min long speech samples, Attias and Schreiner[18] obtained AM power spectra mostly flat in the low-frequency region.

High-frequency slopes estimated on the present data seem to be fairly consistent with those obtained by Sheft *et al.*,[17] with a sharper decrease for AM (−3.9 dB/oct on average) compared to FM (−1.8 dB/oct on average). Following the work of Voss and Clarke,[19,20] Attias and Schreiner[18] have suggested that AM *power* spectra for natural sounds, including speech, have a $1/f^{\alpha}$ shape at large f, with $1 \leq \alpha \leq 2.5$. The high-frequency slopes estimated on the AMa spectra in the present study are consistent with this result as they correspond to a value of $\alpha \approx 2.56$ for the AM power spectra averaged across all languages.

The AMi spectrum (second panel of Fig. 3) offers a more perceptually plausible representation of the AM information contained in the corpus, by integrating the AM amplitude in $\frac{1}{3}$-octave bands and normalizing it by the mean of the band-limited envelope (see Methods section). As noted in the Introduction, such splitting of the AM spectrum emphasizes the medium- and high-rate regions relative to the low rates, where the modulation filters have very narrow bandwidths. Therefore, all AMi spectra reach a maximum around 5 Hz. This band-pass shape has already been observed in many studies calculating the modulation index in specific frequency bands[26,34] or across the whole frequency range.[36]

B. Cross-linguistic differences in the modulation spectra

The statistical analyses showed that linguistic factors had a significant effect on both AMa and AMi spectra. This twofold result reflects the fact that the two spectra correspond to alternative representations of the same AM information in the speech signal. In other words, any change in the temporal envelope of speech should be reflected in both modulation spectra. However, as explained above, the AMi spectrum is a more perceptually plausible representation of AM information than the AMa spectrum, as it takes into account the tuning of the auditory system in the AM domain (logarithmic bandwidths). Hence, the fact that the observed effect of linguistic factors was enhanced in the AMi spectra is consistent with the notion that the human auditory system is optimized for the processing of temporal modulation cues important for language comprehension (as demonstrated for other species, for a large repertoire of animal vocalizations[21]).

1. Word order is reflected in AMi maximum

The study of Japanese and English semi-spontaneous materials by Arai and Greenberg[31] marked a first attempt to compare the modulation spectra of two languages. As in the present study, they found the AMi spectra of the two languages to be strikingly similar, with a clear maximum around 5 Hz. They interpreted this finding as resulting from the high variability of syllabic segments in both languages. However, these qualitative observations leave open the possibility that more subtle differences separate the two languages. Here, we carried out statistical tests to assess possible quantitative differences between the AMi spectra. Consistent with our initial expectations, this comparison demonstrated that the maximum modulation index reached at the peak was significantly lower in Japanese and other CH languages (green lines in Fig. 3) than in English and other HC languages (blue and indigo lines). Note that the difference in maximum modulation index between English and Japanese in our study is consistent with the results of Arai and Greenberg[31] (cf. their Figs. 2 and 3, although they only represent the result for the 1–2 kHz frequency band).

The identification of reliable acoustical correlates of linguistic parameters is an important question in psycholinguistics as they are assumed to "bootstrap" the acquisition of syntax in young children.[5,38,41,72] Previous studies have shown that the syntactic organization of a given language is reflected in the physical realization of prosody.[49,50,65] Namely, in CH languages, prosodic prominence in phonological phrases is carried by higher intensity and/or higher pitch, whereas in HC languages, it is carried by longer duration. The present study suggests that HC and CH languages can be distinguished on the basis of AM features. It may be the case that in HC languages, long, stressed syllables interleaved with short, non-stressed syllables result in a stronger AM at the syllabic rate than in CH languages.

To clarify further the origin of the difference in the AMi spectra, a detailed analysis was conducted on the waveforms of the sentences with minimum and maximum AMi spectra peaks. AMi spectra and waveforms of the two sentences yielding the maximum ("high AMi") and minimum ("low-AMi") peaks for Dutch and Japanese taken as representative examples of the HC and CH language classes, are shown in Fig. 5. This figure reveals that high-AMi sentences are composed of short segments interleaved with brief silent intervals, whereas low-AMi sentences have more slowly fluctuating envelopes, with longer, more sustained segments. These observations suggest that the amplitude of the peak in the AMi spectrum is in fact related to the rate of the most prominent envelope fluctuation in the speech signal. As explained above, the $\frac{1}{3}$-octave band splitting used for calculating the AMi spectrum favors the representation of high

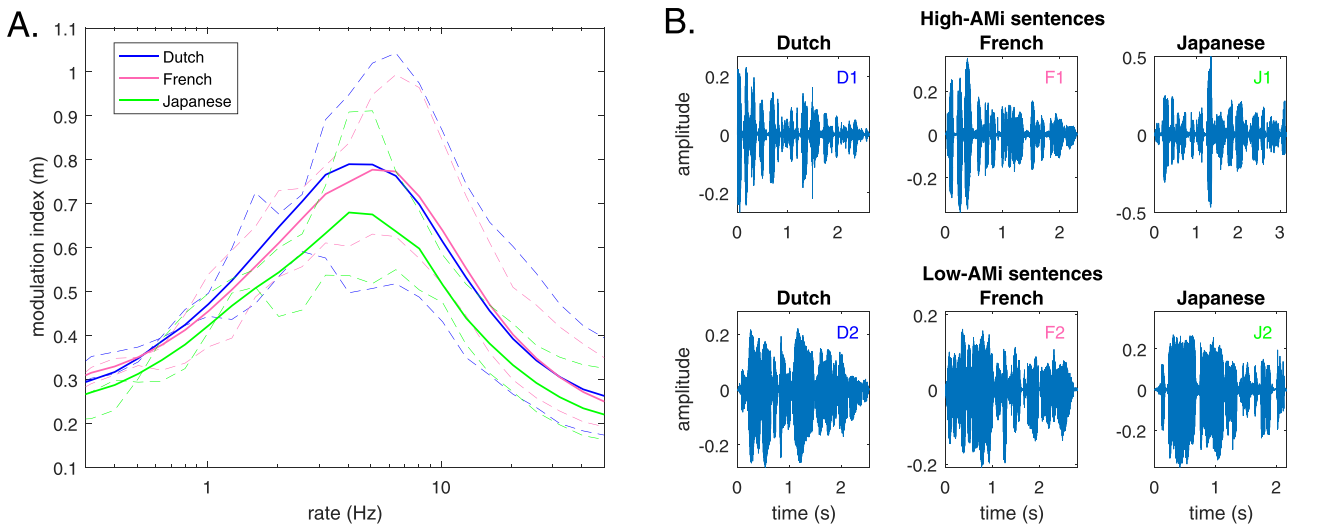

FIG. 5. (Color online) (A) Mean AMi spectra for Dutch, French and Japanese (thick lines) and AMi spectra of the two sentences yielding the maximum and minimum peaks in each language (thin lines). (B) Waveforms of the high-AMi and low-AMi sentences (D1: *"erwtensoep met worst is nog steeds zijn favoriete gerecht"*; D2: *"de juryleden waren unaniem in hun beoordelingen"*; F1: *"Les parents se sont approchés de l'enfant sans faire de bruit"*; F2: *"Les récents évenements ont bouleversé l'opinion internationale"*; J1: *"shi-tokyoku ga rekishi-kuiki no sai-kaihatsu ni tchakushu shita"*; J2: *"opera-za no saigo no konsato wa seiko datta"*).

modulation rates (where the bandwidth of the modulation filters is larger).

2. Language rhythm is reflected in AMi peak rate

In line with our initial expectations, we observed significant differences in peak rate between stress-timed and syllable-timed languages. However, contrary to our predictions, stress-timed AMi spectra showed a lower peak than syllable-timed languages. Ding *et al.*[33] reported no difference in AMi peak rate when comparing semi-spontaneous speech materials from nine languages. However, consistent with the current findings, the peak rate for French, a syllable-timed language, was higher than peak rates for other stress-timed languages (Swedish, English, German, Norwegian, Danish, and Dutch).

To clarify the origin of the difference in AMi peak rate between stress-timed and syllable-timed languages, individual AMi spectra were scrutinized for each language. Figure 6 shows the AMi spectra of the two sentences yielding the maximum and minimum peak rates for a stress-timed (Dutch) and a syllable-timed (French) language of the SRS corpus. Both languages show sentences with a secondary peak around 1–2 Hz in their AMi spectra. This suggests that the downward shift in peak rate for stress-timed languages originates from a greater occurrence of secondary peaks in this low-frequency region of the AMi spectrum, rather than from a mere transposition of the modulation spectrum towards lower rates. This low-rate secondary peak is likely a correlate of the stress pattern in the modulation domain.[11,32] It is unclear why this secondary peak would be less frequent in syllable-timed languages. It may be the case that the greater variability in syllable length for stress-timed languages alters the balance between the primary peak at syllable rate and the secondary peak associated with stress patterns, resulting in an increased prominence of modulation energy in the 1–2 Hz region. This approach stresses the importance of considering the detailed structure of AM spectra for individual utterances in addition to the analyses to mean AM spectra across utterances.

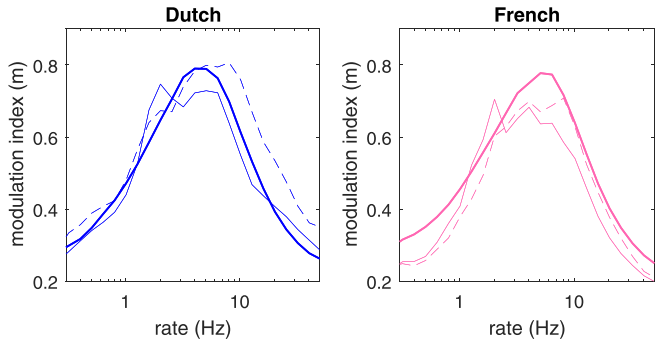

FIG. 6. (Color online) Mean AMi spectrum (thick line) and AMi spectra of the two sentences yielding the maximum (dotted line) and minimum (thin line) peak rates. Left panel: Dutch; Right panel: French.

3. Relationship with rhythmic metrics

A previous study by Ramus *et al.*[41] has provided phonological correlates of rhythmic classes by calculating the respective proportion and variability of vocalic and consonantal intervals. Using the $\%V$ and ΔC metrics, Ramus *et al.*[41] found that languages clustered into groups similar to the traditional linguistic categories. The present study confirms that these linguistic properties are reflected in the temporal modulations of the speech in different languages, by using purely acoustic metrics (i.e., not depending on a preliminary segmentation of the signal by the experimenter), as also suggested by the strong correlations between our AMi measures and $\%V/\Delta C$. Figure 7 plots the languages investigated in the present study on a "AMi peak value" vs "AMi peak position" space, similar to the $\%V$ / ΔC space.

Interestingly in this respect, it can be seen from Fig. 7 that Turkish, a CH language, reaches a high value of AMi at

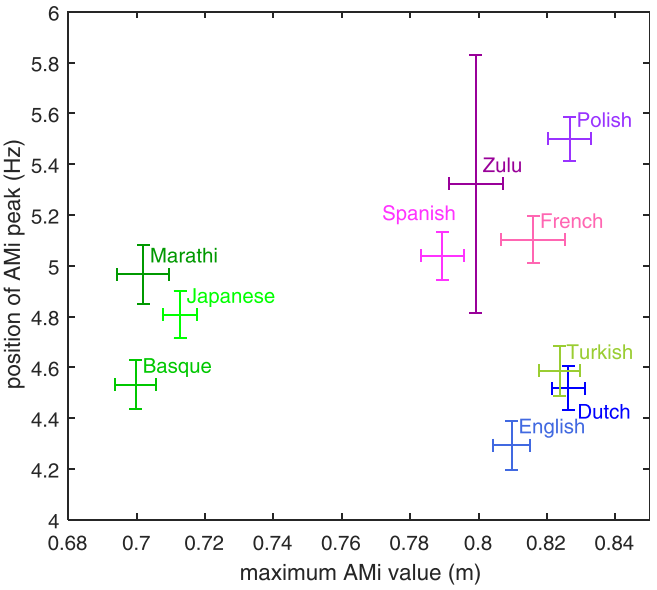

FIG. 7. (Color online) Value and location of the AMi spectrum peak of all the languages in the SRS corpus. Error bars represent ± 1 standard error of the mean.

peak rate whereas all the other CH languages under study show low AMi values. This is consistent with previous language classifications based on *%V* and ΔC[3] showing that Turkish has surprisingly high ΔC, similar to the current HC languages. Moreover, visual inspection of the waveforms of the Turkish corpus revealed that Turkish sentences are composed of short segments interleaved with brief silent intervals, similar to Dutch or English. This lends further support to the idea of a relationship between the consonant structure of a language and its AMi spectrum.

While the rhythmic metrics of *%V* and ΔC showed strong correlations with some of our modulation spectrum measures, the Q10/f_c measure did not show significant correlations with the modulation spectra, except for a weak trend towards a correlation with AMi. This implies that encodings of the speech signal that are efficient and optimal from an information theoretical perspective, capturing the greatest amount of information using the simplest representations, might capture statistical properties of the speech signal other than its temporal modulation. One hypothesis is that they capture its spectral transience as suggested in Guevara Erra and Gervain[64] and Stilp and Lewicki.[73] However, as the Q10/f_c measure was available for only seven of our ten languages, the correlations with this measure have low statistical power. Strong conclusions are thus unwarranted, and future research will need to address this issue in more detail.

Evidence for AMi differences similar to our results also comes from machine learning. Cummins *et al.*[59] have shown that a neural network trained on a database of conversational speech is able to distinguish between pairs of languages based only on the AM component of the speech utterances. Moreover, they suggest that the performance of the network is related to the rhythmic structure of the languages to be compared, with two languages from different groups being correctly discriminated (e.g., Spanish vs English) whereas two languages falling in the same group are not (e.g., Spanish vs French). Although the language clusters derived from the discrimination responses of the network are far from perfect (Japanese and English appear to be indistinguishable, for example), this provides a further indication that the AM component is a cue for identifying the rhythmic structure of a language. Note that, contrary to the present study, this result is based on the (temporal) AM component itself, not on a spectral representation of this component such as the AMa or AMi spectra. However it is not implausible that the envelope cue used by Cummins, Gers, and Schmidhuber's neural network is the same as the one revealed by the modulation spectra.

4. No cross-linguistic differences in FM spectra

It may seem surprising that the observed differences in AM spectra are not reflected in FM spectra, given the covariation of AM and FM components in narrowband signals[53,55] and the fact that lexical stress is marked by changes in level and fundamental frequency. It may be the case that cross-linguistic differences in FM cues were blurred by signal-processing artifacts generated by the current demodulation technique (instantaneous frequency behaving badly in silent intervals). It may also be the case that FM contrasts can only be observed for more salient linguistic features. Further work is warranted to investigate this issue.

C. Effects of number of speakers and speaking style

The initial analysis was based on a corpus using only four speakers per language and semi-read sentences, two factors which potentially limit the possibility to generalize the above findings. It is unlikely at first sight that idiosyncratic factors such as speaking rate are responsible for the present pattern of results because these variations are taken into account by the random factor "speaker" in the statistical analyses. Nonetheless, we also conducted an additional analysis to directly address these issues. The analysis of the SSS corpus indicated that the cross-linguistic differences in AMi spectra could be replicated on a subset of our original languages (English, French, Spanish, and Japanese) using more than 100 speakers per language and a corpus of semi-spontaneous speech (Fig. 8, solid lines). This demonstrates that the observed differences were not solely due to idiosyncratic differences such as speech rate. However, this initial result was obtained for speech materials with limited lexical content, low syntactic complexity and short duration (approximately 1 s). An additional comparison conducted on another corpus of semi-spontaneous speech showed that with more variable, less controlled material, the cross-linguistic differences in AMi spectra disappeared (Fig. 8, dotted lines). It should also be noted that, due to the overall lower quality of the recordings, AMi spectra from the SSS corpus were less peaky compared to those obtained on the SRS corpus.

D. General discussion and limitations

The present study aimed to compare different language classes based on their modulation spectra. We have found important cross-linguistic differences coinciding with

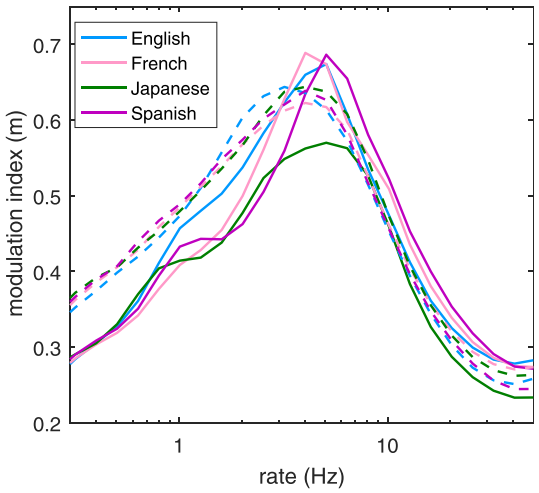

FIG. 8. (Color online) Averaged AMi spectra for all languages of the SSS corpus. Solid lines: FV condition; dotted lines: TS condition.

language classes mainly in the AMi and to a lesser extent the AMa spectra. These results were replicated on another corpus of semi-spontaneous speech using a larger number of speakers, provided that utterance complexity was kept relatively low.

A limitation of the corpora under study is the restricted number of languages in each linguistic category. Further work on a larger corpus including more languages and various stimulus types is needed to confirm the results of the present study.

The additional analysis indicated that these cross-linguistic differences in modulation spectra may disappear for truly spontaneous speech as in daily conversations. This finding certainly limits the ecological value of these cross-linguistic differences which appear to be relatively subtle and require homogeneous speech material to be observed. However, these modulation cues may be perceptible and used in certain "real-life" training situations associated with specific speaking styles, such as infant-directed speech or "clear," hyper-articulated speech.[28,29] Moreover, the current modeling approach taken here (AMi) is still relatively crude and it may well be the case that these modulation cues are enhanced by the auditory system of real listeners. In particular, the AMi approach focuses on temporal-envelope power at the output of audio-frequency and AM channels, and does not take into account short-term adaptation effects, envelope phase, and across-channel contributions that are known to play a role in AM perception.[23,74]

Even though, as demonstrated here, the shape of the AMi spectrum stems from more basic acoustic features in the stimulus such as the proportion of silent intervals, the present analysis is biologically grounded and therefore relevant for the study of speech perception. Indeed, consistent with auditory processing, the acoustic content of the stimulus is decomposed into frequency bands and modulation bands and expressed as a modulation index (i.e., relative to the mean of the band-limited envelope). While other studies have already revealed correlates of linguistic parameters in the acoustic properties of speech,[41] the AMi spectrum offers an insight into the cues that are perceptually accessible. Furthermore, the AM-based representation more comprehensively captures the multiple, frequency-dependent, periodicities in the speech signal than the simple measure of silent intervals, limited to amplitude fluctuations reaching low minimal amplitude.

Another biologically-inspired representation of modulations is the 2-D Fourier transform of the log-spectrogram.[71,75] In the future, it would be valuable to replicate the present analysis using this alternative tool.

V. SUMMARY AND CONCLUSIONS

The aim of this study was to analyze a large corpus of read speech taken from ten languages in terms of their amplitude and frequency modulation spectra to test the hypothesis that broad linguistic categories are reflected in temporal modulations features. The results of the acoustic analyzes can be summarized as follows:

(1) AM and FM spectra are highly similar across all investigated languages, when spectra are expressed in terms of absolute value.
(2) When the AM spectrum is expressed in terms of "modulation index," a more perceptually-based metrics, three linguistic groups can be differentiated based on their AM content: CH languages, HC stress-timed languages and HC syllable-timed languages.
(3) These findings persist for a larger number of speakers. Speaking style, however, has an influence on these acoustic differences that should be taken into account in future studies.

ACKNOWLEDGMENTS

The authors want to thank Marina Nespor for her helpful comments on the manuscript, as well as two anonymous reviewers for their suggestions that improved substantially the current study. We also wish to thank Marina Nespor, Jacques Mehler, Franck Ramus, and Thierry Nazzi for contributing the speech recordings. This study was funded by the ANR grant "SpeechCode" (ANR-15-CE37-0009-01) to J.G. and C.L., the Human Frontiers Science Program Young Investigator Grant (RGY-0073-2014) to J.G., ANR-11-0001-02 PSL* and ANR-10-LABX-0087.

[1]M. S. Dryer and M. Haspelmath, *WALS Online* (Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, 2013).
[2]P. Jusczyk, "Learning language: What infants know about it, and what we don't know about that," in *Language, Brain, and Cognitive Development: Essays in Honor of Jacques Mehler*, edited by E. Dupoux (MIT Press, Cambridge, MA, 2001), pp. 363–377.
[3]J. Mehler, N. Sebastian-Galls, and M. Nespor, "Biological foundations of language: Language acquisition, cues for parameter setting and the bilingual infant," in *The New Cognitive Neuroscience*, edited by M. S. Gazzaniga (MIT Press, Cambridge, MA, 2004), pp. 825–836.
[4]F. Ramus, M. D. Hauser, C. Miller, D. Morris, and J. Mehler, "Language discrimination by human newborns and by cotton-top tamarin monkeys," Science **288**, 349–351 (2000).
[5]J. L. Morgan and K. Demuth, eds., *Signal to Syntax: Bootstrapping From Speech To Grammar in Early Acquisition* (Psychology Press, Mahwah, NJ, 1996).
[6]H. Dudley, "The carrier nature of speech," Bell Syst. Tech. J. **19**, 495–515 (1940).

[7]R. Plomp, "The role of modulation in hearing," in *HEARING–Physiological Bases and Psychophysics*, edited by D. R. Klinke and D. R. Hartmann (Springer, Berlin, Germany, 1983), pp. 270–276.
[8]T. Houtgast and H. J. M. Steeneken, "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 (1985).
[9]J. C. R. Licklider and I. Pollack, "Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech," J. Acoust. Soc. Am. **20**, 42–51 (1948).
[10]R. Drullman, "Temporal envelope and fine structure cues for speech intelligibility," J. Acoust. Soc. Am. **97**, 585–592 (1995).
[11]R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," Science **270**, 303–304 (1995).
[12]K. Saberi and D. R. Perrott, "Cognitive restoration of reversed speech," Nature **398**, 760 (1999).
[13]H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**, 318–326 (1980).
[14]S. Rosen, "Temporal information in speech: Acoustic, auditory and linguistic aspects," Philos. Trans. R. Soc. Lond., B, Biol. Sci. **336**, 367–373 (1992).
[15]V. Leong, M. A. Stone, R. E. Turner, and U. Goswami, "A role for amplitude modulation phase relationships in speech rhythm perception," J. Acoust. Soc. Am. **136**, 366–381 (2014).
[16]A.-L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: Emerging computational principles and operations," Nat. Neurosci. **15**, 511–517 (2012).
[17]S. Sheft, V. Shafiro, C. Lorenzi, R. McMullen, and C. Farrell, "Effects of age and hearing loss on the relationship between discrimination of stochastic frequency modulation and speech perception," Ear. Hear. **33**, 709–720 (2012).
[18]H. Attias and C. E. Schreiner, "Temporal low-order statistics of natural sounds," in *NIPS* (MIT Press, Cambridge, MA, 1997), pp. 27–33.
[19]R. Voss and J. Clarke, "'1/f noise' in music and speech," Nature **258**, 317–318 (1975).
[20]R. F. Voss and J. Clarke,"'1/*f* noise' in music: Music from 1/*f* noise," J. Acoust. Soc. Am. **63**, 258–263 (1978).
[21]F. A. Rodríguez, C. Chen, H. L. Read, and M. A. Escabí, "Neural modulation tuning characteristics scale to efficiently encode natural sound statistics," J. Neurosci. **30**, 15969–15980 (2010).
[22]T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," J. Acoust. Soc. Am. **102**, 2892–2905 (1997).
[23]T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration," J. Acoust. Soc. Am. **102**, 2906–2919 (1997).
[24]T. Houtgast, "Frequency selectivity in amplitude-modulation detection," J. Acoust. Soc. Am. **85**, 1676–1680 (1989).
[25]S. P. Bacon and D. W. Grantham, "Modulation masking: Effects of modulation frequency, depth, and phase," J. Acoust. Soc. Am. **85**, 2575–2580 (1989).
[26]R. Plomp, "The negative effect of amplitude compression in multichannel hearing aids in the light of the modulation-transfer function," J. Acoust. Soc. Am. **83**, 2322–2327 (1988).
[27]S. Greenberg, H. Carvey, L. Hitchcock, and S. Chang, "Temporal properties of spontaneous speech–A syllable-centric perspective," J. Phonet. **31**, 465–485 (2003).
[28]J. C. Krause and L. D. Braida, "Acoustic properties of naturally produced clear speech at normal speaking rates," J. Acoust. Soc. Am. **115**, 362–378 (2004).
[29]J. C. Krause and L. D. Braida, "Evaluating the role of spectral and envelope characteristics in the intelligibility advantage of clear speech," J. Acoust. Soc. Am. **125**, 3346–3357 (2009).
[30]S. Greenberg and T. Arai, "The relation between speech intelligibility and the complex modulation spectrum," in *Proceedings of the 7th European Conference on Speech Communication and Technology*, Aalborg, Denmark (2001), pp. 473–476.
[31]T. Arai and S. Greenberg, "The temporal properties of spoken japanese are similar to those of English," in *Proceedings of Eurospeech* (1997), pp. 1011–1014.
[32]U. Goswami and V. Leong, "Speech rhythm and temporal structure: Converging perspectives?," in *Linguistic Rhythm and Literacy*, Trends in Language Acquisition Research No. 17, edited by J. Thomson and L. Jarmulowicz (John Benjamins, Amsterdam, the Netherlands, 2016), pp. 111–132.
[33]N. Ding, A. D. Patel, L. Chen, H. Butler, C. Luo, and D. Poeppel, "Temporal modulations in speech and music," Neurosci. Biobehav. Rev. (published online 2017).
[34]S. J. van Wijngaarden and T. Houtgast, "Effect of talker and speaking style on the Speech Transmission Index (L)," J. Acoust. Soc. Am. **115**, 38–41 (2004).
[35]F. Dubbelboer and T. Houtgast, "A detailed study on the effects of noise on speech intelligibility," J. Acoust. Soc. Am. **122**, 2865–2871 (2007).
[36]A. Schlueter, U. Lemke, B. Kollmeier, and I. Holube, "Intelligibility of time-compressed speech: The effect of uniform versus non-uniform time-compression algorithms," J. Acoust. Soc. Am. **135**, 1541–1555 (2014).
[37]F. Ramus, "Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues," Ann. Rev. Language Acquis. **2**, 85–115 (2002).
[38]J. Mehler, P. Jusczyk, G. Lambertz, N. Halsted, J. Bertoncini, and C. Amiel-Tison, "A precursor of language acquisition in young infants," Cognition **29**, 143–178 (1988).
[39]T. Nazzi, J. Bertoncini, and J. Mehler, "Language discrimination by newborns: Toward an understanding of the role of rhythm," J. Exp. Psychol. Hum. Percept. Perform. **24**, 756–766 (1998).
[40]R. M. Dauer, "Stress-timing and syllable-timing reanalyzed," J. Phon. **11**, 51–62 (1983).
[41]F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," Cognition **73**, 265–292 (1999).
[42]V. Dellwo, "Rhythm and speech rate: A variation coefficient for ΔC," in *Language and Language-Processing: Proceedings of the 38th Linguistic Colloquium*, Frankfurt, Germany (2006), pp. 231–241.
[43]A. Loukina, G. Kochanski, B. Rosner, E. Keane, and C. Shih, "Rhythm measures and dimensions of durational variation in speech," J. Acoust. Soc. Am. **129**, 3258–3270 (2011).
[44]E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," in *Laboratory Phonology 7* (De Gruyter, Boston, MA, 2002).
[45]L. Wiget, L. White, B. Schuppler, I. Grenon, O. Rauch, and S. L. Mattys, "How stable are acoustic metrics of contrastive speech rhythm?," J. Acoust. Soc. Am. **127**, 1559–1569 (2010).
[46]V. Dellwo and P. Wagner, "Relations between language rhythm and speech rate," in *Proceedings of the International Congress of Phonetics Science*, Barcelona, Spain (2003), pp. 471–474.
[47]A. Arvaniti, "Rhythm, timing and the timing of rhythm," Phonetica **66**, 46–63 (2009).
[48]A. Arvaniti, "The usefulness of metrics in the quantification of speech rhythm," J. Phonetics **40**, 351–373 (2012).
[49]J. Gervain and J. F. Werker, "Prosody cues word order in 7-month-old bilingual infants," Nat. Commun. **4**, 1490 (2013).
[50]M. Nespor, M. Shukla, R. V. D. Vijver, C. Avesani, H. Schraudolf, and C. Donati, "Different Phrasal Prominence Realizations in VO and OV Languages," Lingue Linguaggio **7**, 1–28 (2008).
[51]G. Fenk-Oczlon and A. Fenk, "Crosslinguistic correlations between size of syllables, number of cases, and adposition order," in *Sprache und natürlichkeit, gedenkband für Willi Mazerthaler* (*Language and Naturalness, Commemorative Book for Willi Mazerthaler*) (Narr, Tübingen, Germany, 2005).
[52]F.-G. Zeng, K. Nie, G. S. Stickney, Y.-Y. Kong, M. Vongphoe, A. Bhargave, C. Wei, and K. Cao, "Speech recognition with amplitude and frequency modulations," Proc. Natl. Acad. Sci. U.S.A. **102**, 2293–2298 (2005).
[53]S. Sheft, M. Ardoint, and C. Lorenzi, "Speech identification based on temporal fine structure cues," J. Acoust. Soc. Am. **124**, 562–575 (2008).
[54]J. Obleser, B. Herrmann, and M. J. Henry, "Neural oscillations in speech: Don't be enslaved by the envelope," Front. Hum. Neurosci. **6**, 250 (2012).
[55]A. Papoulis, "Random modulation: A review," IEEE Trans. Acoust. Speech Signal Process. **31**, 96–105 (1983).
[56]C. Binns and J. F. Culling, "The role of fundamental frequency contours in the perception of speech against interfering speech," J. Acoust. Soc. Am. **122**, 1765–1776 (2007).
[57]S. E. Miller, R. S. Schlauch, and P. J. Watson, "The effects of fundamental frequency contour manipulations on speech intelligibility in background noise," J. Acoust. Soc. Am. **128**, 435–443 (2010).
[58]J. Vaissière, "Language-independent prosodic features," in *Prosody: Models and Measurements* (Springer, New York, 1983), pp. 53–65.

[59]F. Cummins, F. Gers, and J. Schmidhuber, "Comparing prosody across many languages," *Tech. Rep.* (Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale, Lugano, Switzerland, 1999).
[60]P. Welby, "The role of early fundamental frequency rises and elbows in French word segmentation," Speech Commun. **49**, 28–48 (2007).
[61]S.-A. Jun and C. Fougeron, "A phonological model of French intonation," in *Intonation*, *Text, Speech and Language Technology No. 15*, edited by A. Botinis (Springer, Amsterdam, the Netherlands, 2000), pp. 209–242.
[62]D. R. Ladd, *Intonational Phonology* (Cambridge University Press, Cambridge, MA, 1996).
[63]A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. **111**, 1917–1930 (2002).
[64]R. Guevara Erra and J. Gervain, "The efficient coding of speech: Cross-linguistic differences," PLoS One **11**, 0148861 (2016).
[65]M. Molnar, M. Carreiras, and J. Gervain, "Language dominance shapes non-linguistic rhythmic grouping in bilinguals," Cognition **152**, 150–159 (2016).
[66]R. Cole and Y. Muthusamy, "OGI Multilanguage Corpus LDC94s17" (Linguistic Data Consortium, Philadelphia, 1994).
[67]V. Hohmann, "Frequency analysis and synthesis using a Gammatone filterbank," Acta Acust. Acust. **88**, 433–442 (2002).
[68]Z. M. Smith, Bertrand Delgutte, and A. J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," Nature **416**, 87–90 (2002).
[69]K. Nie, G. Stickney, and F.-G. Zeng, "Encoding frequency modulation to improve cochlear implant performance in noise," IEEE Trans. Biomed. Eng. **52**, 64–73 (2005).
[70]W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, 2nd ed. (Cambridge University Press, Cambridge, MA, 1992).
[71]N. C. Singh and F. E. Theunissen, "Modulation spectra of natural sounds and ethological theories of auditory processing," J. Acoust. Soc. Am. **114**, 3394–3411 (2003).
[72]M. Nespor, "About parameters, prominence and bootstrapping," in *Language, Brain, and Cognitive Development: Essays in Honor of Jacques Mehler*, edited by Emmanuel Dupoux (MIT Press, Cambridge, MA, 2001), pp. 127–142.
[73]C. E. Stilp and M. S. Lewicki, "Statistical structure of speech sound classes is congruent with cochlear nucleus response properties," J. Acoust. Soc. Am. **134**, 4229 (2013).
[74]M. L. Jepsen, S. D. Ewert, and T. Dau, "A computational model of human auditory signal processing and perception," J. Acoust. Soc. Am. **124**, 422–438 (2008).
[75]Y. E. Cohen, F. Theunissen, B. E. Russ, and P. Gill, "Acoustic features of rhesus vocalizations and their representation in the ventrolateral prefrontal cortex," J. Neurophysiol. **97**, 1470–1484 (2007).