Alexandrou, Anna; Saarinen, Timo; Kujala, Jan; Salmelin, Riitta

A multimodal spectral approach to characterize rhythm in natural speech

Published in:
Journal of the Acoustical Society of America

DOI:
10.1121/1.4939496

Published: 01/01/2016

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Alexandrou, A., Saarinen, T., Kujala, J., & Salmelin, R. (2016). A multimodal spectral approach to characterize rhythm in natural speech. *Journal of the Acoustical Society of America*, *139*(1), 215-226. https://doi.org/10.1121/1.4939496

A multimodal spectral approach to characterize rhythm in natural speech

Anna Maria Alexandrou, Timo Saarinen, Jan Kujala, and Riitta Salmelin

Citation: The Journal of the Acoustical Society of America **139**, 215 (2016); doi: 10.1121/1.4939496
View online: https://doi.org/10.1121/1.4939496
View Table of Contents: http://asa.scitation.org/toc/jas/139/1
Published by the Acoustical Society of America

A multimodal spectral approach to characterize rhythm in natural speech

Anna Maria Alexandrou,[a] Timo Saarinen, Jan Kujala, and Riitta Salmelin
Department of Neuroscience and Biomedical Engineering, Aalto University, FI-00076 AALTO, Finland

(Received 10 September 2015; revised 25 November 2015; accepted 22 December 2015; published online 12 January 2016)

Human utterances demonstrate temporal patterning, also referred to as rhythm. While simple oromotor behaviors (e.g., chewing) feature a salient periodical structure, conversational speech displays a time-varying quasi-rhythmic pattern. Quantification of periodicity in speech is challenging. Unimodal spectral approaches have highlighted rhythmic aspects of speech. However, speech is a complex multimodal phenomenon that arises from the interplay of articulatory, respiratory, and vocal systems. The present study addressed the question of whether a multimodal spectral approach, in the form of coherence analysis between electromyographic (EMG) and acoustic signals, would allow one to characterize rhythm in natural speech more efficiently than a unimodal analysis. The main experimental task consisted of speech production at three speaking rates; a simple oromotor task served as control. The EMG–acoustic coherence emerged as a sensitive means of tracking speech rhythm, whereas spectral analysis of either EMG or acoustic amplitude envelope alone was less informative. Coherence metrics seem to distinguish and highlight rhythmic structure in natural speech.
 [http://dx.doi.org/10.1121/1.4939496]

[SSN] Pages: 215–226

I. INTRODUCTION

Natural speech is characterized by regularities in the occurrence of its constituent elements. These temporal regularities may also be referred to as speech rhythm. A salient rhythmic structure can be observed in basic oromotor communicative gestures such as lip–smacking in primates (Ghazanfar, 2013), as well as in rudimentary forms of speech, such as babbling (Dolata *et al.*, 2008) and syllable repetition (Ruspantini *et al.*, 2012). Despite its inherently more complex structure, natural speech also displays rhythmic components which are, however, harder to detect and quantify. Rhythm is viewed as a key organizational principle of speech and considered crucial for communication (Cummins and Port, 1998; Kohler, 2009; Tilsen, 2009). Speech rhythm enables language acquisition (e.g., Petitto *et al.*, 2001; Nazzi and Ramus, 2003), development of reading skills (Flaugnacco *et al.*, 2014; Woodruff *et al.*, 2014), dynamic coupling of speech production and speech perception (Martin, 1972; Smith, 1992), and predictions about salient future events that facilitate subsequent processing (Cutler and Butterfield, 1992). The present study aims to quantify the temporal regularities in spontaneous, natural speech by examining the periodic structure of speech-related physical signals.

The definition of speech rhythm adopted in the present study is, hence, somewhat different from a predominant outlook on speech rhythm which emphasizes linguistic and phonetic aspects of rhythm such as meter and prosody. A conventional linguistic approach to speech rhythm divides languages into different rhythmic categories ("time-stressed" or "syllable-stressed") according to timing patterns of stressed syllables (Abercrombie, 1967). This premise of simple isochrony in speech has since been questioned (e.g., Cummins and Port, 1998; Ramus *et al.*, 1999; Kohler, 2009). Speech rhythm has subsequently been assessed through descriptive measures examining the temporal relationships between basic phonological units, using, for instance, speaking rate variations (Dellwo and Wagner, 2003; Dellwo, 2008). Alternatively, approaches based on coupled oscillators (e.g., Barbosa, 2007; O'Dell and Nieminen, 2009; Meireles and Gambarini, 2012) view speech rhythm as being composed of two interdependent recurring elements, periodicity and structure (Fraisse, 1974), or a syllabic and syllable-stress oscillator, respectively (Barbosa, 2007). This multiplicity and evolution in methods and approaches for quantifying rhythm can be seen as a testimony to the fact that speech, in general (Greenberg, 2006), and speech rhythm, in particular (Kohler, 2009), is a multi-layered phenomenon which can be studied from various viewpoints. The present paper focuses on one quantitative aspect of speech rhythm that considers periodic fluctuations in speech signals associated with the production of words and syllables. Such a mechanistic definition of speech rhythm may also be understood in terms of the syllabic oscillator part of the coupled oscillator model proposed by Barbosa (2007), and is a description of rhythm that uses measurable speech-related signals. Periodic components in a speech stream may be characterized by spectral decomposition of speech signals (Tilsen, 2008; Tilsen and Johnson, 2008), where periodicity is identified as power maxima in acoustic amplitude envelope spectra (e.g., Chandrasekaran *et al.*, 2009).

Frequency-domain signal processing tools are being increasingly employed to investigate the acoustic (Das *et al.*,

[a]Electronic mail: anna.alexandrou@aalto.fi

2008; Tilsen and Johnson, 2008; Tilsen and Arvaniti, 2013) and muscular (electromyographic, EMG) (e.g., Ruspantini *et al.*, 2012) aspects of speech signals and speech rhythm. The acoustic envelope carries temporal features which reflect rhythm in speech (Rosen, 1992). These are observed on slow timescales and chiefly consist of low-frequency amplitude fluctuations of the acoustic envelope (Rosen, 1992). In speech rhythm research, the acoustic signal has often been investigated in the time domain (e.g., Ramus *et al.*, 1999); in particular, energy fluctuations in the acoustic amplitude envelope are suggestive of syllabic rhythm (Marcus, 1981; Cummins and Port, 1998). When aiming to describe rhythm in speech, the power spectrum of the amplitude envelope of the acoustic signal seems more informative than time-domain methods since the spectral estimate does not rely on any pre-determined hypothesis about the rhythmic structure of an utterance (Tilsen and Johnson, 2008). In addition to acoustic signals, neuromuscular (i.e., EMG) signals are highly relevant markers of speech rhythm: they indirectly measure the synchronous firing of motor neurons and, hence, are indicative of motor control and activation patterns of a given muscle or muscle group. For instance, peri-oral EMG (e.g., Wohlert and Hammen, 2000) is a reliable marker of muscular activity associated with the movement of the articulators (e.g., lip and tongue). Frequency-domain analysis of surface EMG signals has demonstrated a rhythmic pattern of activation in articulatory muscles during speech-related tasks (Smith *et al.*, 1993; Ruspantini *et al.*, 2012; Shepherd *et al.*, 2012). While EMG alone captures various important aspects of speech production, articulatory muscle activity is invariably accompanied by respiratory and phonatory events from the vocal tract and vocal chords. It could be, thus, suggested that acoustic and EMG signals are interrelated in both time and frequency domains, although each signal may also differentially highlight (sub)segments of speech, such as consonants or vowels (e.g., Gracco, 1988).

Frequency-domain analyses of acoustic and EMG signals thus have, each separately, proven their usefulness in describing speech rhythm. However, it seems important to consider these two signals jointly as they represent complementary parts of the process of natural, coherent speech production. Speech is a complex signal originating from the coordination of numerous effectors with varying intrinsic timescales. Multiple processing levels involving the neuromuscular, articulatory, and respiratory systems come to play in order to produce the resulting acoustic output (Alfonso and Baer, 1982). Furthermore, the rhythmic characteristics of the output are dynamic and vary with time, thus making it difficult to accurately define and quantify speech rhythm (O'Dell *et al.*, 2007; Tilsen and Arvaniti, 2013). Because of this, it would seem unlikely that collecting data from a single modality would be sufficient to fully describe the temporal rhythmic features of the acoustic output. In accordance with previous views (Cummins, 2009, 2012), it is thus proposed that reaching a global description of speech rhythm would greatly benefit from adoption of a multimodal and integrative perspective. Coherence analysis between acoustic and EMG signals is a multimodal method which provides a quantitative measure of the correlation of these signals in the frequency domain. Coherence analysis as a measure of synchrony between two signals (for instance, EMG-EMG or EMG-cortical coherence) has valuable applications in both basic neurophysiological research and clinical applications (for a review, see Grosse *et al.*, 2002).

In this study, a multimodal approach including coherence analysis of EMG and acoustic signals is employed to investigate rhythm in conversational speech. Acoustic and EMG signals are collected during natural speech production at different speaking rates. Speaking rate is a complex temporal variable determined by both articulation time and pause time (Grosjean and Deschamps, 1975). Habitual speaking rates are behaviorally expressed as phonemic (10–12 Hz), syllabic (4–5 Hz), and word (2–3 Hz) production frequencies (Levelt, 1999; Poeppel *et al.*, 2008). Speaking rate displays remarkable flexibility: one may voluntarily modulate the rate of an utterance so that it is faster or slower than the habitual rate (Grosjean and Lane, 1976). In running speech, linguistic units such as words and syllables recur in a semi-regular fashion as a function of time. This quasi-periodic recurrence of linguistic units results in a distinctive, albeit time-varying, rhythmic pattern in speech signals (Tilsen and Arvaniti, 2013). Speaking rate is viewed as a global parameter that affects the entire command sequence for an utterance. Modulations in speaking rate induce phonetic modifications which alter the temporal features of an utterance and, therefore, its rhythmic structure (Smith *et al.*, 1995; Dellwo, 2008; Meireles and Barbosa, 2008). Physically, these changes are reflected as shifts in the spectral power distribution of speech-related signals (Kelso *et al.*, 1986; Smith *et al.*, 2002). In the present experimental design, speaking rate is employed as an independent variable that serves to alter the power spectral distribution of the measured signals in a controlled manner and, in the subsequent signal analysis, helps to determine the relevance and adequacy of our multimodal approach in discerning rhythmic patterns in speech. The natural speech production tasks are complemented by a /pa/ syllable repetition task as a control (Ruspantini *et al.*, 2012). Syllable repetition represents a rudimentary form of speech (Davis and MacNeilage, 2002) that offers a simple and clear-cut rhythmic motor task to serve as a frame of reference when investigating the rhythmic features of the more complex natural speech.

The present study addresses the question of an effective means of measuring how rhythm is encoded in natural speech. Given that speech production is inherently multimodal, coherence analysis between EMG and acoustic signals could reveal the shared, functionally most relevant frequencies of operation of the human speech production apparatus. A further key point of interest is whether these operational frequencies correlate with behaviorally estimated production frequencies of linguistic units such as words and syllables. If proven efficient, a multimodal approach, such as the one presented here, could shed more light on the nature of speech rhythm and contribute to a better understanding of the underlying mechanisms of the production of rhythmic linguistic output.

II. METHODS

A. Participants

Twenty healthy Finnish-speaking volunteers (11 females; 9 males; all right-handed; mean age 24.5 yr, range 19–35 yr) gave their informed consent to participate in the study, as approved by the Aalto University Ethics Committee.

B. Experimental design

The participants were asked to produce connected speech prompted by questions (in Finnish) randomly derived from six distinct thematic categories (own life, preferences, people, culture/traditions, society/politics, general knowledge; Table I). To avoid repetition and learning effects, each thematic question was presented only once during the experiment. When replying, the participants were asked to speak casually, as if talking to a friend, at one of three rates: natural/normal, slow, or fast. With regard to the slow rate, they were asked to aim for 50% of their normal speaking rate, by preferably increasing their articulation time rather than their pause time. For the fast rate, they were instructed to speak as fluently and continuously as possible at the highest speaking rate possible, however, without severely compromising the intelligibility or the correct articulation of the produced speech.

A training phase preceded the actual experiment to help the participants to outline and modify their speaking rate range. The participants were presented with a speaking rate continuum (modified from Tsao *et al.*, 2006) that represented the range schematically and in which 100% stood for the spontaneous, natural speaking rate. The continuum consisted of several anchoring points at 25%, 50%, 75%, 125%, 150%, and 200% of the normal speaking rate. Participants were presented with a training set of thematic questions (different than those used in the actual experiment) to be answered first at normal rate (i.e., at 100%), and then at a faster (∼150% of normal) or slower (50% of normal) rate than normal speech, aided by the anchoring points. Subsequently, in the actual experiment, speaking rate variations were carried out based on the subjective perception of the participants; no external pacing device was used.

A single speech production block consisted of a spoken thematic question (duration 3–9 s; mean 5.6 ± 1.3 s) and a 40-s response period. A signal tone (50-ms, 1-kHz tone) indicated the beginning of a block, and another signal tone (50-ms, 75-Hz tone) signified the beginning and end of the response period. All sounds were presented via panel loudspeakers. The mean interval from the end of one response period to the beginning of the next one was 9.1 s, composed of a 2.5-s rest period between blocks, mean question duration 5.6 s and a 1-s delay before response onset.

As a control condition, we examined repeated production of the syllable /pa/ (Ruspantini *et al.*, 2012). All participants performed this task at their normal rate; additionally, 10 out of 20 subjects were randomly chosen as a control group that performed /pa/ repetition at slow (50% of normal repetition rate) and fast rates (close to maximal, ∼150% of normal repetition rate). A /pa/ repetition block consisted of a 40-s /pa/ repetition period, with a tone signal (50-ms, 75-Hz tone) indicating the beginning and end of the period.

TABLE I. Thematic questions used to elicit natural speech from the participants. Each column stands for one thematic category, each category consisting of five questions.

Own Life	Likings	People	Culture/Traditions	Society/Politics	General Knowledge
What are your plans for this day and/or the following days?	What kinds of foods do you like?	Describe a known musician, singer, or composer. Why do you find her/him interesting?	Describe what happens during a holiday at a cottage in the Finnish countryside.	What is the role of the President of Finland? Describe the Finnish presidential institution.	What do you know about skiing and snowboarding?
What kind of hobbies do you have or have had during your life?	What kinds of vacation trips do you like?	Describe a known artist, writer, or film director. Why do you find her/him interesting?	Describe a traditional Christmas holiday.	Describe the political parties of Finland.	Describe what happens during the Olympic games.
What is a typical weekend like for you?	What kinds of books or movies do you like?	Which movie, literature, or comic book character would you like to be and why?	Describe the traditional Midsummer's celebration in Finland.	Talk about public transport and private car usage in Finland.	Describe Africa's geography and nature.
Talk about your work or education.	What kinds of animals do you like?	Describe a top athlete from the present or the past.	What kinds of traditions are associated with May 1st celebration in Finland?	How does the Finnish school system operate?	Talk about what comes to mind about poker and gambling.
What is a typical weekday like for you?	What kinds of desserts do you especially enjoy?	Describe the current President of Finland.	Describe what happens during a summer festival in Finland.	What do you know about garbage and recycling policies in Finland?	What kinds of buildings can be seen in the center of Helsinki (capital of Finland)?

Repetition blocks were separated by 10 s of rest to approximate the timing of the speech conditions.

The order of the experimental conditions was randomized across participants. Prior to the first block of each condition, participants were informed of the upcoming task (speech production or /pa/ repetition) via visual input. There were six blocks per experimental condition, thus, totaling ∼4 min of data for each rate of speech production and /pa/ repetition. During the measurement, participants were instructed to keep their gaze on a fixation point projected on a screen that was placed in front of them, at a distance of ∼1 m from their sitting position.

The data reported in the present study were collected as part of a more extensive neuroimaging project in which magnetoencephalography was used to track brain dynamics in the aim to characterize the correspondence between neural patterns and behavior in natural language perception and production. The neuroimaging data will be reported separately.

C. Recordings

Acoustic signals were recorded using a portable audio recorder (FOSTEX FR-2LE, Tokyo, Japan) and sampled at 44.1 kHz. Surface EMG signals were registered with reusable circular electrodes (conductive area diameter 0.4 mm), low-pass filtered at 330 Hz, and sampled at 1.5 kHz. Two bipolar EMG channels were used to record muscular activity from the lower lip muscles (orbicularis oris), as well as muscular activity associated with tongue and jaw movements (primarily from genioglossus and mylohyoid muscles). Muscular activity from lower lip muscles was measured by placing the pair of electrodes directly under the left-hand side of the lower lip, ∼1 cm from the midline. Muscular activity associated with tongue and jaw muscles was recorded by placing the pair of electrodes on the soft tissue directly beneath the jawline (left-hand side), ∼2 cm from the midline. The exact location of the electrodes was determined individually for each participant via tactile inspection of the soft tissue beneath the jawline during repetitive production of the /n/ consonant. For both EMG channels, inter-electrode distance was 1.5 cm. Electrode resistance remained below 10 kΩ.

D. Behavioral analysis

The raw acoustic signal was analyzed both through a behavioral pipeline [Fig. 1(A), left] and a signal analysis pipeline [Fig. 1(A), right].

The audio materials from all participants, comprising both the speech production and /pa/ syllable repetition conditions, were transmitted to a transcribing company (Tutkimustie Oy, Tampere, Finland) for strict verbatim transcription, in which the audio materials are transcribed without being edited or modified. For speech audio materials, all spoken words, including utterances, false starts, repetitions, filler words, and slang were transcribed, including meaningful pauses and usual sounds (such as laughter). Any other kind of non-verbal communication was excluded. Syllable repetition materials were transcribed using the same principles. Transcription was carried out manually (i.e., without the aid of any voice-recognition system) using a transcription-specific software to play back each audio file.

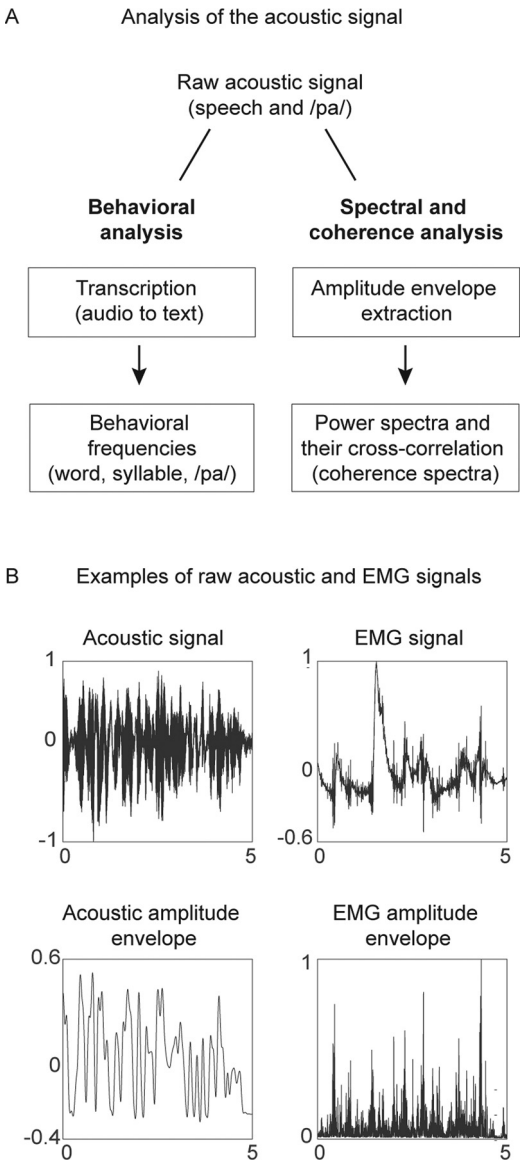

FIG. 1. Analysis procedures. (A) Flow chart of behavioral (left) and spectral (right) analysis of the acoustic signal. See Sec. II for a detailed description. (B) Examples of raw data (top) and resulting amplitude envelopes (bottom) of acoustic (left) and EMG (right) signals recorded from one participant during normal-rate speech production. Normalized amplitude (in arbitrary units; *y* axis) is plotted against time (in seconds; *x* axis). Each plot displays a 5-s chunk of data taken from a 40-s speech production block. For this particular block, mean word and syllable repetition frequencies were 2.66 Hz and 5.45 Hz, respectively.

Subsequently, based on the transcription, syllable production frequencies were calculated by syllabifying all transcribed words. Word and syllable production frequencies or the /pa/ syllable production frequency were calculated separately for each 40-s speech production block. The mean values of the word and syllable production frequencies, or the /pa/ syllable repetition frequency, for a given experimental condition and participant were obtained by averaging the values across the six blocks.

The individual mean word and syllable production frequencies were used as a behavioral reference to interpret

any peaks appearing in the acoustic and EMG amplitude envelope spectra and EMG–acoustic coherence spectra of individual subjects. Similarly, the grand average across-participants behavioral word and syllable production frequencies and mean /pa/ syllable repetition frequencies were used as reference in order to interpret any peaks emerging in the group-level acoustic and EMG amplitude envelope spectra and EMG–acoustic coherence spectra.

Speaking rate varies within a speaking turn that is comprised of multiple utterances and even within the course of a single utterance. Such variation can be described with statistical dispersion measures of syllable and word production frequencies which may help to interpret power and coherence spectra. Here, the dispersion measure of choice was the range of word and syllable production frequencies and /pa/ syllable repetition frequencies for each subject and each speaking rate. The range was computed as the difference between the minimum and maximum mean production and repetition frequency observed across the six 40-s blocks per speaking rate. To facilitate comparisons between speaking rates, normalized mean ranges for a given linguistic unit (word, syllable) and /pa/ syllable at a given speaking rate were obtained by dividing the mean range by the mean production or repetition frequency.

E. Statistical analysis

All variables were first tested for normality of distribution using a Shapiro–Wilk test of normality. The effect of speaking rate on mean word and syllable production frequencies and ranges (20 participants) was tested using a one-way within-subjects analysis of variance (ANOVA). The same ANOVA design was used to evaluate the effect of repetition rate on mean /pa/ repetition frequencies and ranges (ten participants), as well as to compare within-participant variation for speech and /pa/ syllable repetition at a given rate (ten participants). *Post hoc* pairwise comparisons were Bonferroni corrected.

The effect of speaking rate on the variance of word and syllable production frequencies (20 participants) or /pa/ repetition frequencies (10 participants) was tested using a likelihood-ratio (LR) test of equality of variances for paired samples. This test evaluates differences between two normally distributed variances by extracting two separate restricted log-likelihood values for each variable. The difference of these two log-likelihood values was computed and referred to as a chi-squared distribution.

F. Spectral analysis of acoustic and EMG signals

As summarized in Fig. 1(A) (right), the raw acoustic signal [example in Fig. 1(B), top left] was first bandpass filtered (fourth order Butterworth filter) at 80–2500 Hz to emphasize the voiced signal portions which are relevant for speech rhythm analysis (Hertrich *et al.*, 2013). Subsequently, the amplitude envelope of the bandpassed signal was extracted by full-wave rectifying the signal and low-pass filtering (fourth order Butterworth filter) at 10 Hz (Tilsen and Arvaniti, 2013). The amplitude envelope was then normalized by subtracting the mean and rescaling the envelope by its maximum absolute value, resulting in values between 1 and −1 [example in Fig. 1(B), bottom left]. The spectrum of the downsampled (by a factor of 10), Tukey-windowed ($r = 0.2$), and zero-padded envelope was calculated by taking the squared magnitude of the fast Fourier transform using an 8192-point window. Finally, a moving average operation was applied to the resulting spectrum in order to smooth out random spectral peaks and thus facilitate interpretation of the spectrum (Tilsen and Arvaniti, 2013).

The raw EMG signal [example in Fig. 1(B), top right] was first high-pass filtered at 15 Hz to remove motion artifacts (Van Boxtel, 2001). Subsequently, the EMG amplitude envelope [example in Fig. 1(B), bottom right] was extracted by full-wave rectification of the signal. EMG envelope spectrum was calculated using Welch's spectral estimator with a Hanning window (8192 points) at 75% overlap (Ruspantini *et al.*, 2012).

Group-level acoustic and EMG amplitude envelope spectra for both speech production and /pa/ syllable repetition at all three speaking rates were computed by first dividing the amplitude envelope spectrum of each individual participant by its mean value and then summing these normalized spectra across participants.

G. Coherence analysis

EMG–acoustic coherence was computed to determine possible common periodic features in the EMG and acoustic signals. Coherence quantifies the relationship between two time-series in the frequency domain. The coherence spectrum was obtained by first computing the cross spectrum of the amplitude envelopes of the two signals and subsequently dividing it by the power spectra of the amplitude envelopes of both signals (fast Fourier transform, Hanning 4096-point window). Group-level EMG–acoustic coherence spectra for both speech production and /pa/ syllable repetition at all three speaking rates were computed by first dividing each individual-participant coherence spectrum by its mean value and then summing these normalized spectra across participants.

III. RESULTS

A. Behavioral analysis of speech rate

Word and syllable production frequencies [Fig. 2(A); Table II] differed significantly between all speaking rates: the rate increased from slow through normal to fast rate [word: $F(2,3.23) = 237.32$, $P < 0.0005$; syllable: $F(2,3.23) = 281.40$, $P < 0.0005$; *post hoc* pairwise tests word and syllable, slow < normal and normal < fast, $P < 0.0005$]. The same increasing rate pattern was evident for /pa/ repetition [$F(2,0.67) = 41.47$, $P < 0.0005$; *post hoc* pairwise tests, slow < normal and normal < fast, $P < 0.0005$].

Between-participant variation [schematically illustrated by the box length in Fig. 2(A)] was smallest for slow-rate speech for both words and syllables [word: slow vs normal χ^2 (13.38,1), $P < 0.0001$; slow vs fast χ^2 (8.85,1), $P < 0.005$; syllable: slow vs normal χ^2 (2.886,1), $P < 0.05$]. For /pa/ repetition, variation increased systematically with speech

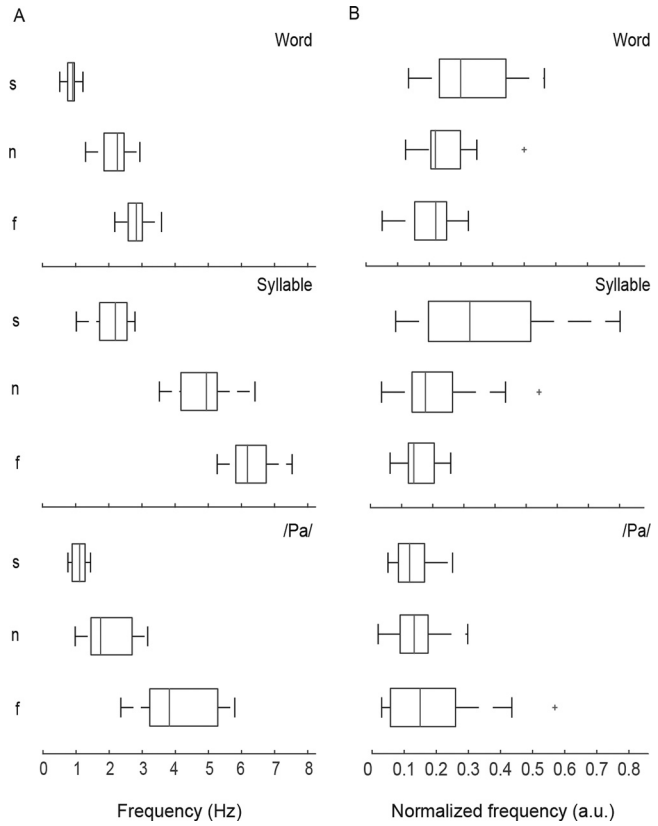

FIG. 2. Behavioral frequencies for words (top; $n = 20$), syllables (middle; $n = 20$), and /pa/ syllable (bottom; $n = 10$). (A) Distribution of mean production frequencies and (B) distribution of mean normalized range of production frequencies. Each subplot displays the three production/repetition rates (slow "s," normal "n," fast "f"). The whiskers (horizontal line) represent the maximum and minimum data points within 1.5 of the interquartile range. Interquartile range (third quadrant - first quadrant) is shown by the length of the box. The vertical line within the box represents the median. Outliers are marked by a cross.

rate: slow < normal < fast [slow vs normal χ^2 (13.67,1), $P < 0.0005$; slow vs fast χ^2 (54.95,1), $P < 10^{-8;}$ normal vs fast χ^2 (2.81,1), $P < 0.05$].

Within-participant variation, that is, the normalized frequency range of words and syllables [Fig. 2(B); Table III] was larger at the slow rate than at normal or fast rates for speech conditions [word: $F(2,38) = 7.9$, $P < 0.001$; syllable: $F(2,38) = 13.5$, $P < 0.0005$; *post hoc* pairwise tests, word: slow > fast, $P < 0.05$; slow > normal approaching significance $P = 0.07$; *post hoc* pairwise tests, syllable: slow > normal, $P < 0.01$; slow > fast, $P < 0.0005$]. For /pa/ repetition, range values did not differ significantly between the different

TABLE II. Word and syllable production frequencies ($n = 20$; mean ± standard deviation) and /pa/ syllable production frequencies ($n = 10$) at three production rates. For slow and fast rates, frequencies as percentages of normal are given in brackets.

	Word	Syllable	/pa/
Slow	0.86 ± 0.19 Hz (40%)	2.08 ± 0.5 Hz (44%)	0. 87 ± 0.27 Hz (47%)
Normal	2.17 ± 0.46 Hz	4.82 ± 0.81 Hz	1.83 ± 0.82 Hz
Fast	2.84 ± 0.39 Hz (136%)	6.26 ± 0.65 Hz (123%)	4.11 ± 1.36 Hz (225%)

TABLE III. Normalized word and syllable production range ($n = 20$; mean ± standard deviation) and /pa/ syllable repetition range ($n = 10$) at three production rates.

	Word production rate range	Syllable production rate range	/pa/ syllable repetition range
Slow	0.36 ± 0.20	0.39 ± 0.23	0.13 ± 0.06
Normal	0.25 ± 0.08	0.21 ± 0.12	0.13 ± 0.08
Fast	0.21 ± 0.07	0.15 ± 0.06	0.20 ± 0.18

rates [$F(2,18) = 1.1$; $P = 0.35$]. When comparing speech and /pa/ syllable repetition ($n = 10$; only participants that performed both tasks at all three rates), within-participant variation was smaller for /pa/ repetition than for speech at slow [$F(2,0.63) = 7.59$, $P < 0.005$; *post hoc* pairwise tests, word > /pa/, $P < 0.05$; syllable > /pa/, $P < 0.05$] and normal rates [$F(2,\ 0.18) = 3.41$, $P < 0.05$; *post hoc* pairwise tests, word > /pa/, $P < 0.05$].

B. Spectral analysis of speech rhythm

1. Power spectra

For natural speech [Fig. 3(A)], the group-level acoustic and EMG power spectra (lip and tongue) were characterized by a rather flat pattern with no discernible power maxima (beyond the lowest-frequency 1/*f* power increase) at any of the three speaking rates. The pattern was the same for the individual acoustic (Fig. 4) and EMG (Fig. 5) power spectra, with no salient local maxima.

In contrast, /pa/ syllable repetition [Fig. 3(B)] revealed salient group-level acoustic and EMG power maxima at all three rates. Furthermore, both the acoustic and EMG (lip and tongue) spectra displayed fairly similar power distribution patterns and local maxima.

2. Coherence spectra

Group-level EMG–acoustic coherence spectra (Fig. 6) demonstrated salient peaks. Contrary to the case of power spectra, local maxima were evident for both speech [Fig. 6(A)] and /pa/ conditions [Fig. 6(B)]. For /pa/ repetition, coherence of the acoustic signal with either tongue or lip EMG channels displayed a quasi-identical spectrum. For speech, however, the coherence peaks for the acoustic signal with either tongue or lip EMG were slightly apart (~1 Hz difference). The coherence peaks approximately aligned with the mean behavioral frequencies [see Fig. 2(A); Table II] for both speech and /pa/ syllable repetition.

A correspondence between behavioral frequencies and coherence maxima was also evident at the individual level for both speech production (Fig. 7) and, most strikingly, for /pa/ syllable repetition (Fig. 8).

IV. DISCUSSION

The main finding of the present study was that the temporal regularities in speech are remarkably well captured using a multimodal spectral approach. Specifically, EMG–acoustic coherence emerged as a more informative measure than

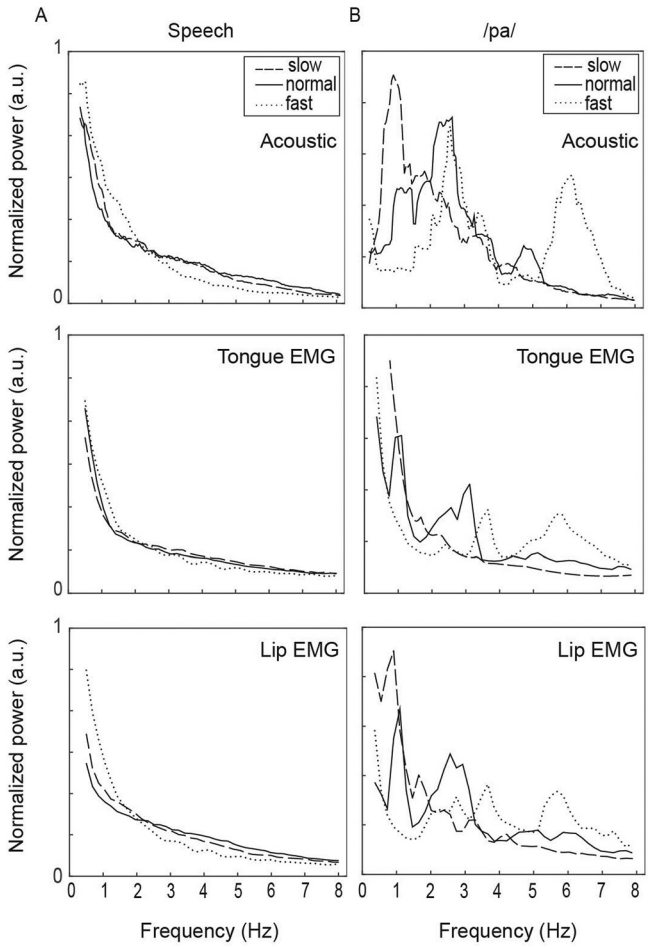

FIG. 3. Group-level amplitude envelope spectra of the acoustic (top), tongue EMG (middle), and lip EMG (bottom) signals. (A) Speech production and (B) /pa/ syllable repetition. Normalized power (in arbitrary units; *y* axis) is plotted against frequency (in Hz; *x* axis). Each plot displays data from three production/repetition rates: slow (dashed line), normal (solid line), and fast (dotted line). The *x* axis has the same scale (0–8 Hz) as for the behavioral data displayed in Fig. 1(A). Normalized power was computed using different normalizing factors for speech production and /pa/ syllable repetition; consequently, the resulting group-level amplitude envelope spectra are not directly numerically comparable between these two conditions. All visualizations of these results have been scaled so that the *y* axis has a minimum value of 0 and a maximum value of 1.

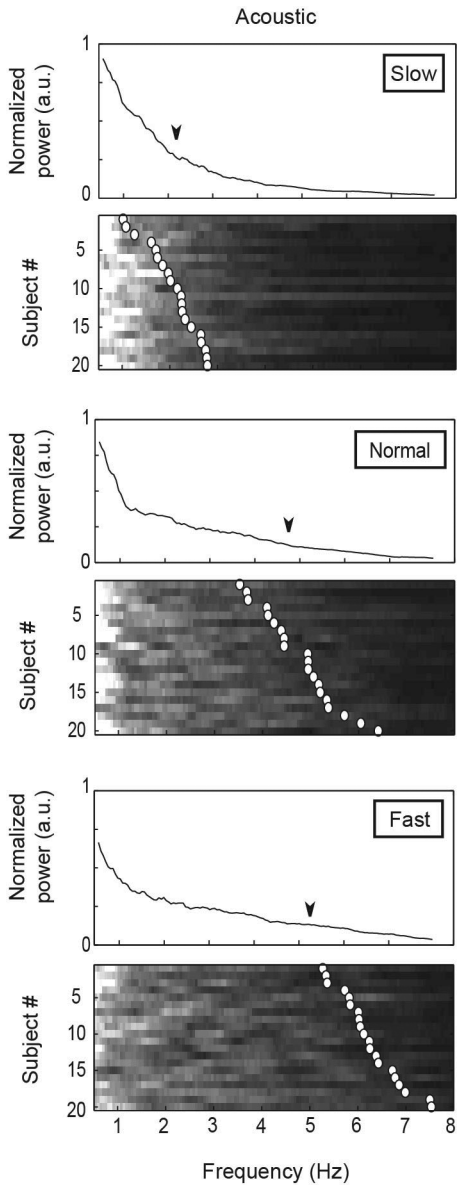

FIG. 4. Relationship of group-level and individual acoustic power spectra (speech) with mean and individual behavioral syllable production frequencies at the three speaking rates: slow (top), normal (middle), and fast (bottom). For each rate, the upper panel displays the group-level spectrum: normalized power (in arbitrary units; *y* axis) is plotted against frequency (in Hz; *x* axis), with the mean syllable production frequency indicated with an arrowhead. The lower panel compiles the power spectral distribution for the individual participants (rows), with power peaks indicated by a lighter shade of gray. The participants are ordered by their individual syllable production frequencies (circles).

spectral analysis of either EMG or acoustic amplitude envelopes alone. Coherence spectral peaks reflected behavioral frequencies, whereas no such peaks were observed in the EMG or acoustic amplitude envelope spectra.

The combined frequency-domain analysis of EMG and acoustic signals, in form of coherence, was here shown to successfully highlight behaviorally relevant temporal patterning in speech. Although both signals reflect articulatory processes—EMG as a measure of muscle activity and acoustic signal as a marker of the vocal respiratory function—they are very different in nature and origin and contain other features not necessarily directly related to articulation, including various kinds of noise [such as pink noise with a characteristic $1/f$ trend (Voss and Clarke, 1975)]. Coherence analysis helps to suppress random, uncorrelated activity in the signals and accentuate any shared oscillatory patterning. The present findings regarding EMG–acoustic coherence are consistent with the global framework of speech rhythm and oscillatory cycles as an organizing principle of speech and as such are linked, on a general level, to prominent theories in speech-acoustic research (Cummins and Port, 1998; MacNeilage, 1998; O'Dell and Nieminen, 2009; Tilsen, 2009). More specifically, the signal-processing methods presented in this paper are able to directly distinguish the oscillatory components in naturalistic speech signals associated with word and syllable production frequencies (Chandrasekaran *et al.*, 2009; Tilsen and Johnson, 2008). Furthermore, EMG–acoustic coherence may be linked to the notion of articulatory gesture as defined in the theory of articulatory phonology (Browman and Goldstein, 2000). According to this theory, events taking place during speech production

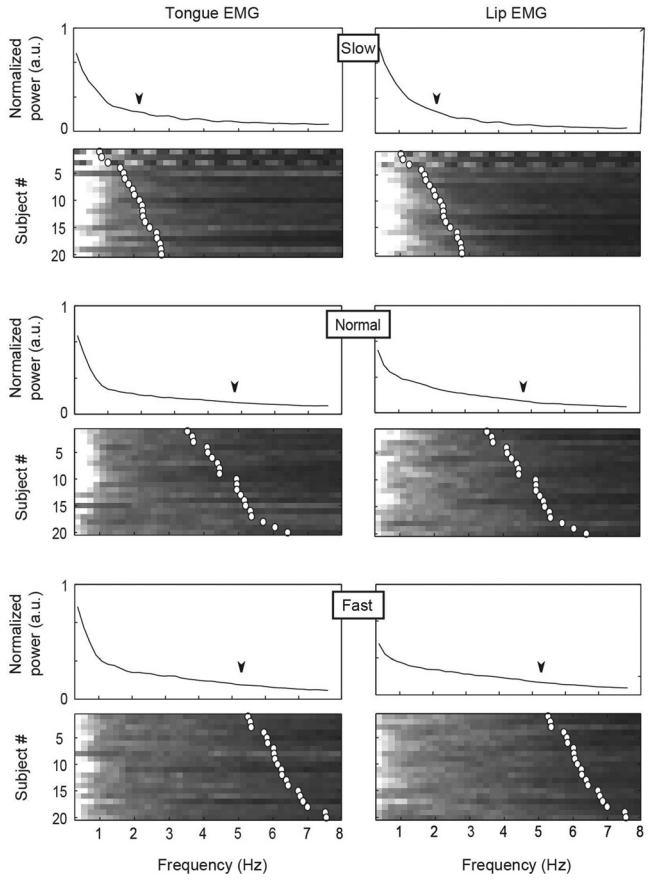

FIG. 5. Relationship of group-level and individual EMG power spectra (speech) (left, tongue; right, lip) with mean and individual behavioral syllable production frequencies, at the three speaking rates: slow (top), normal (middle), and fast (bottom). For each rate, the upper panel displays the group-level spectrum: normalized power (in arbitrary units; *y* axis) is plotted against frequency (in Hz; *x* axis), with the mean syllable production frequency indicated with an arrowhead. The lower panel compiles the power spectral distribution for the individual participants (rows), with power peaks indicated by a lighter shade of gray. The participants are ordered by their individual syllable production frequencies (circles).

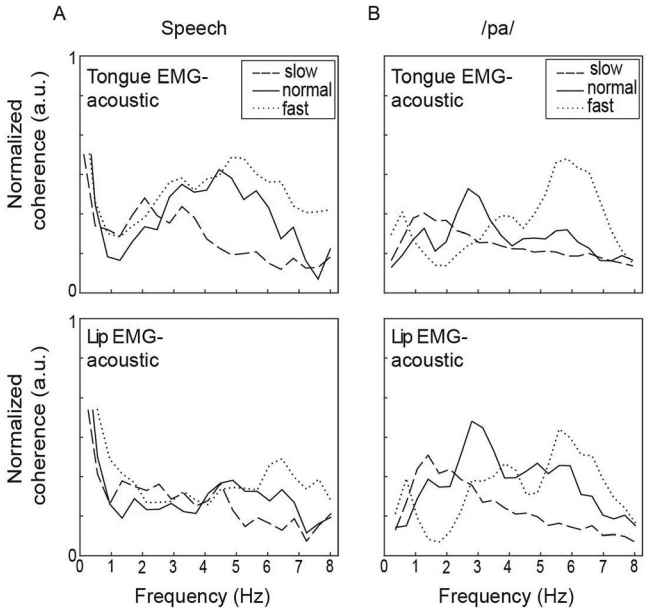

FIG. 6. Group-level EMG–acoustic coherence spectra (top, tongue; bottom, lip) for speech production (left) and /pa/ syllable repetition (right). Normalized coherence (in arbitrary units; *y* axis) is plotted against frequency (in Hz; *x* axis). The *x* axis has the same scale (0–8 Hz) as for the behavioral data [Fig. 1(A)] and power spectra [Figs. 2(A) and 2(B)]. Each plot displays data from the three production/repetition rates: slow (dashed line), normal (solid line), and fast (dotted line). Normalized coherence was computed using different normalizing factors for speech production and /pa/ syllable repetition; consequently, the resulting group-level amplitude envelope spectra are not directly numerically comparable between these two conditions. All visualizations of these results have been scaled so that the *y* axis has a minimum value of 0 and a maximum value of 1.

may be modeled by domain-general task dynamics (Saltzman and Kelso, 1987). Task dynamics in speech consist of target vocal tract configurations requiring specific action from certain articulators (e.g., lip opening). Natural speech production and syllable repetition, paradigms used in the present study, consist of dynamic spatiotemporal events involving successive gestures. Acoustic output follows a given articulatory movement after a certain fixed amount of time (Schaeffler *et al.*, 2014). Thus, coherence seems like an appropriate measure for singling out the parts of the EMG and acoustic signals that are most directly linked with articulatory dynamics.

Coherence analysis may be further extended to include kinematic data, which may be combined with existing evidence of time-domain correlations of articulator velocities (Gracco, 1988) and articulator apertures (Chandrasekaran *et al.*, 2009) with EMG and acoustic signals. Kinematic data would also allow the examination of how dynamic coupling of articulatory gestural kinematics with linguistic units, such as words and syllables (Tilsen, 2009), is manifested in the frequency domain. Other uses of coherence analysis in speech research involve the quantification of the functional coupling (EMG-EMG coherence) between perioral and mandibular muscles involved in articulation, enabling the investigation of motor control patterns during speech production in both adults (e.g., Moore *et al.*, 1998) and children (e.g., Moore and Ruark, 1996). Furthermore, coherence analysis has been used to reveal functional links between the motor cortex and EMG or kinematic signals in rhythmic tasks (e.g., Jerbi *et al.*, 2007; Piitulainen *et al.*, 2013), including a rudimentary language production task (e.g., Ruspantini *et al.*, 2012).

In the present study, salient EMG–acoustic coherence spectral peaks were found in both speech and /pa/ repetition tasks, suggesting an operational synergy that does not depend on the degree of linguistic complexity. The frequency of the coherence peaks largely aligned with the behaviorally estimated production frequencies, both at group-level and in individual participants. The coherence peaks reflect an aspect of speech rhythm closely associated with oscillatory properties of speech-related signals, as opposed to approaches focusing on linguistic aspects of rhythm, such as stress patterns. For /pa/ syllable repetition, use of either lip or tongue EMG resulted in a very similar coherence spectrum. For speech, however, the local maximum of the EMG–acoustic coherence spectrum occurred at a slightly lower frequency for tongue than lip muscles. This suggests that for a simple oromotor task, muscles of the jaw area and lip muscles are tightly coordinated, whereas speech production relies on a different mode of operation, with some degree of desynchronization between the articulators (Smith, 1992). Hence, it may be concluded that

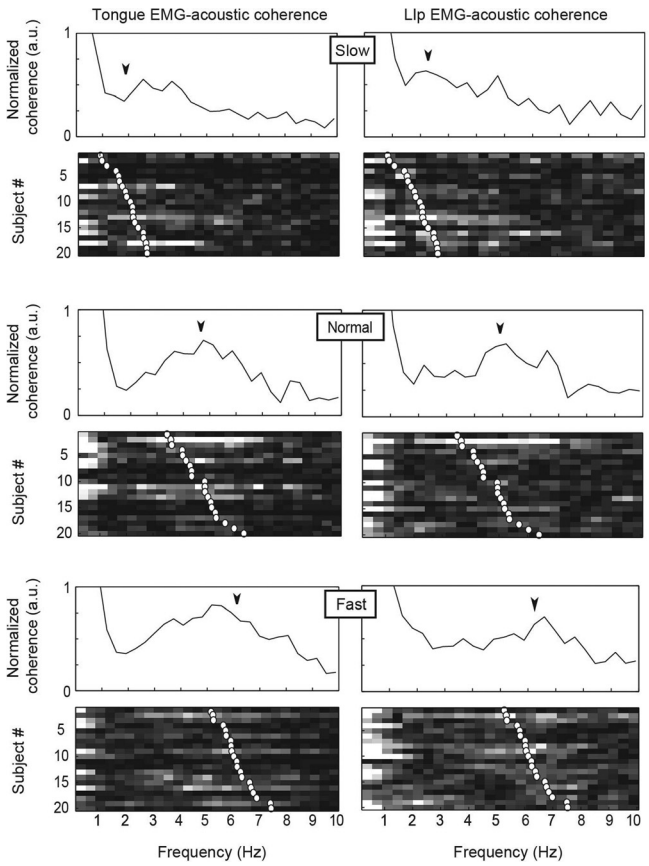

FIG. 7. Relationship of group-level and individual EMG–acoustic coherence spectra for speech production (left, tongue; right, lip) with mean and individual behavioral syllable production frequencies, at the three speaking rates: slow (top), normal (middle), and fast (bottom). For each rate, the upper panel displays the group-level coherence spectrum: normalized coherence (in arbitrary units; *y* axis) is plotted against frequency (in Hz; *x* axis), with the mean syllable production frequency indicated with an arrowhead. The lower panel compiles the power spectral distribution for the individual participants (rows), with coherence peaks indicated by a lighter shade of gray. The participants are ordered by their individual syllable production frequencies (circles).

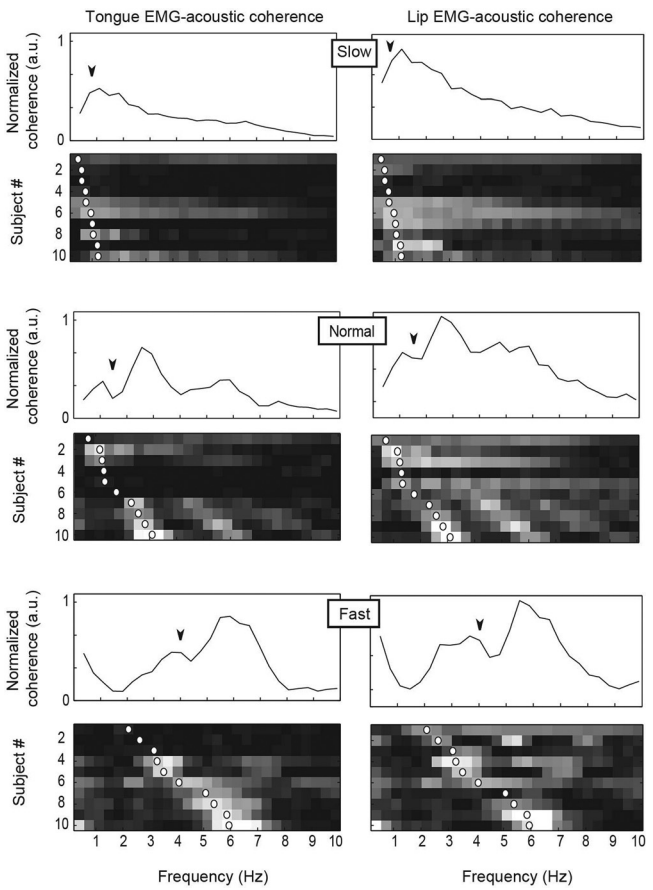

FIG. 8. Relationship of group-level and individual EMG–acoustic coherence spectra for /pa/ syllable repetition (left, tongue; right, lip) with mean and individual behavioral /pa/ syllable production frequencies, at the three repetition rates: slow (top), normal (middle), and fast (bottom). For each rate, the upper panel displays the group-level coherence spectrum: normalized coherence *x* (in arbitrary units; *y* axis) is plotted against frequency (in Hz; *x* axis), with the mean syllable production frequency indicated with an arrowhead. The lower panel compiles the power spectral distribution for the individual participants (rows), with coherence peaks indicated by a lighter shade of gray. The participants are ordered by their individual /pa/ syllable production frequencies (circles).

EMG–acoustic coherence not only highlights periodic components in speech but can also provide insights into the role of various articulators.

Previous studies have used extensive speech corpora to show that the production frequencies of linguistic units and, hence, the periodicity of the speech signal could be mapped as peaks in the acoustic amplitude envelope power spectra (Das *et al.*, 2008; Tilsen and Johnson 2008; Chandrasekaran *et al.*, 2009; Tilsen and Arvaniti, 2013). However, even for larger data sets, 1/*f* trend removal has been deemed necessary for saliently distinguishing spectral peaks related to speech rhythm (Chandrasekaran *et al.*, 2009; see also Ruspantini *et al.*, 2012). In the present study, EMG and acoustic spectra as such were not particularly informative for assessing periodicity in speech production, although salient peaks in power spectra were observed for the more rudimentary /pa/ syllable production. One likely reason is the relatively concise data set (speech material from 20 participants, 4 minutes of data per speaking rate per participant). The rhythmic pattern of speech is characterized by inherent irregularities; these irregularities tend to become amplified with lesser amounts of data. In the present study, irregularities in rhythm are manifested as notable intra-individual variations (within-subject variation) in word and syllable production frequencies. According to the coupled oscillator model, the lack of a continuous periodic patterning in the speech signal may be theoretically represented by the introduction of uncertainty (O'Dell *et al.*, 2007). In contrast, /pa/ syllable repetition, a simple rhythmic task, was shown to feature much less variation compared with speech, especially for slow and normal rates. Indeed, such a simple oromotor task displays a nearly perfect oscillatory pattern with little variation in frequency (O'Dell and Nieminen, 2009). Importantly, EMG–acoustic coherence analysis succeeded in extracting salient spectral peaks even in the more irregular natural speech and for this relatively concise experimental data set. The high behavioral relevance of the resulting spectral peaks is emphasized by their alignment with the behavioral syllable/word production frequencies as a function of speaking rate.

EMG–acoustic coherence analysis could be especially beneficial in a clinical context, where the amount of data

tends to be limited. Specifically, several disorders of speech and language, including aphasia (e.g., Hadar *et al.*, 1998; Patel, 2005), Parkinson's disease (Fox *et al.*, 1996; Giraud *et al.*, 2008; Lu *et al.*, 2010), and stuttering (e.g., Alm, 2004), involve impairments of speech rhythm. Although the etiology of these pathologies is quite diverse, dysfunctions of rhythm-generating structures in the brain, such as the basal ganglia, seem to be a common causal factor (e.g., Brunner *et al.*, 1982). Hence, it may be suggested that a description of pathological manifestations of speech rhythm in patient groups could possibly advance the understanding of the underlying causes and, consequently, contribute to the development of appropriate treatments for a given pathology. EMG–acoustic coherence could be introduced as an efficient tool for such purposes.

Speaking rate as an integral part of speech rhythm is a multifaceted variable. There is considerable intra-individual variation in speaking rate, manifested as changes in the production frequency and duration of linguistic units (Janse, 2004), due to a variety of both linguistic and extra-lingustic factors, such as gender, neuromuscular constraints, and presence of noise in the environment and language (e.g., Byrd, 1992; Tsao *et al.*, 1997; Jacewicz et al., 2009). The present findings link speaking rate to variation in behavioral frequencies and, thus, in temporal structure of speech. The observed mean word (∼2 Hz) and syllable (∼5 Hz) production frequencies at normal rate fall within the range previously reported in the literature for a variety of languages (e.g., Levelt, 1999; Poeppel *et al.*, 2008; Ruspantini *et al.*, 2012). Similar to speech, mean /pa/ syllable repetition frequency at normal rate (∼2 Hz) is also consistent with previous reports (Ruspantini *et al.*, 2012). However, languages have different rhythmic properties (e.g., Ramus and Mehler, 1999). As an example, habitual speaking rates demonstrate language-specific variations (Dellwo, 2008), with lower syllabic frequencies in Spanish, German, and English than Italian (Clopper and Smiljanic, 2011; Tilsen and Arvaniti, 2013). These differences would potentially be reflected as spectral shifts in coherence peaks relative to our present findings for Finnish.

Despite these cross-linguistic variations, the syllabic rate of ∼5 Hz is regarded as an important structural element in terms of binding and integration in speech across languages (MacNeilage, 1998; Giraud and Poeppel, 2012) and plays a central role in coupled oscillator models of speech rhythm (e.g., O'Dell and Nieminen, 2009; Tilsen, 2009). Intriguingly, the preference for certain frequencies does not seem to be confined solely to speech-related tasks, but rather appears to be a cross-modal phenomenon encompassing multiple human motor behaviors, such as finger-tapping and walking (MacDougall and Moore, 2005; Jerbi *et al.*, 2007). The apparent preference for a specific rhythm may be a domain-general phenomenon related to both optimized neural processing and mechanical efficiency of task performance (Lindblom, 1983; Sparrow, 1983; Tsao *et al.*, 1997). Furthermore, such behavioral motor rhythms seem to find counterparts in the neural dynamics of the motor cortex (Jerbi *et al.*, 2007; Ruspantini *et al.*, 2012), as well as the basal ganglia and the cerebellum (Buhusi and Meck, 2005).

Speaking rate variations were duly carried out by the participants; this observation was consistent with previous reports of on-demand speaking rate modulations (e.g., Tsao *et al.*, 2006). However, for all three speaking rates, between-participant variation was small. This suggests that within a certain speech production tempo, the speaking rates of different individuals were rather similar. In contrast, /pa/ syllable repetition rates varied considerably between individuals for all three speaking rates. It may be suggested that speech is an over-learned, albeit complex, construct (Smith, 1992), unlike the rather more artificial /pa/ syllable repetition. Hence, it may be proposed that speech features a relatively tight control of all its constituent parameters, ensuring that there is little variation between individual speech production frequencies. An alignment of normal speaking rates across individuals may importantly serve a communicative purpose by ensuring optimal coupling between interlocutors.

The present findings contribute to the recently initiated cross-disciplinary discussion on the definition of rhythm that seeks to bring together the fields of neurophysiology and behavior (Smith *et al.*, 2014). A description of speech rhythm, such as the one provided here, would afford valuable information when considering the functional role of rhythm in speech comprehension. The existence of a hierarchical, rhythmic internal structure in speech has been advanced as the key element in initiating the process of transformation of an incoming physical signal into comprehensible lexical units (Poeppel *et al.*, 2008). In line with the notion that speech production and speech perception are functionally intertwined (e.g., Pulvermüller and Fadiga, 2010; Giraud and Poeppel, 2012), speech rhythm has been suggested to serve as the "bridging" element between these two processes. Furthermore, correspondence of the syllabic rate (∼5 Hz) with the timescales of spontaneous oscillatory activity in cortical neuron populations has led to a view that the existence of temporal regularities in both speech and cortical signals is paramount for successful cortical processing of spoken language (for a review, see Peelle and Davis, 2012). A number of studies from both a behavioral (Drullman *et al.*, 1994a,b; Shannon *et al.*, 1995; Smith *et al.*, 2002) and a neurophysiological perspective (e.g., Ghitza and Greenberg, 2009; Peelle and Davis, 2012) have provided additional evidence that speech rhythm, in addition to spectral detail, contains crucial information employed by the listener to extract meaning from an utterance. For instance, a disruption of acoustic cues corresponding to the syllabic rate (∼5 Hz) has proven detrimental to comprehension (e.g., Drullman *et al.*, 1994a; Shannon *et al.*, 1995).

V. CONCLUSIONS

The present findings demonstrate that coherence analysis, a spectral analysis tool linking different measurement modalities, is far more informative than a unimodal spectral approach in quantifying periodicity of speech signals. Local maxima in EMG–acoustic coherence spectra signify the existence of functional synergy between articulatory systems and phonatory systems, and the frequency of maximum coherence aligns with speaking rate. Future studies on natural

speech production could utilize these same approaches to examine the relationship between recordings from oral articulatory systems and phonatory systems, as well as from motor cortical areas.

ACKNOWLEDGMENTS

This work was financially supported by the Academy of Finland (personal grants to J.K. and R.S.) and the Sigrid Jusélius Foundation. The authors declare no competing financial interests.

Abercrombie, D. (**1967**). *Elements of General Phonetics* (Edinburgh University Press, Edinburgh), Chap. 6, pp. 89–110.
Alfonso, P. J., and Baer, T. (**1982**). "Dynamics of vowel articulation," Lang. Speech **25**, 151–173.
Alm, P. A. (**2004**). "Stuttering and the basal ganglia circuits: A critical review of possible relations," J. Commun. Disord. **37**, 325–369.
Barbosa, P. N. A. (**2007**). "From syntax to acoustic duration: A dynamical model of speech rhythm production," Speech Commun. **49**, 725–742.
Browman, C. P., and Goldstein, L. (**2000**). "Competing constraints on intergestural coordination and self-organization of phonological structures," Les Cahiers de l'ICP. Bull. commun. parlée **5**, 25–34.
Brunner, R. J., Kornhuber, H. H., Seemüller, E., Suger, G., and Wallesch, C. (**1982**). "Basal ganglia participation in language pathology," Brain Lang. **16**, 281–299.
Buhusi, C. V., and Meck, W. H. (**2005**). "What makes us tick? Functional and neural mechanisms of interval timing," Nat. Rev. Neurosci. **6**, 755–765.
Byrd, D. (**1992**). "Preliminary results on speaker-dependent variation in the TIMIT database," J. Acoust. Soc. Am. **92**, 593–596.
Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A. A. (**2009**). "The natural statistics of audiovisual speech," PLoS Comput. Biol. **5**, e1000436.
Clopper, C. G., and Smiljanic, R. (**2011**). "Effects of gender and regional dialect on prosodic patterns in American English," J. Phonetics **39**, 237–245.
Cummins, F. (**2009**). "Rhythm as entrainment: The case of synchronous speech," J. Phonetics **37**, 16–28.
Cummins, F. (**2012**). "Oscillators and syllables: A cautionary note," Front. Psychol. **3**, 1–2.
Cummins, F., and Port, R. (**1998**). "Rhythmic constraints on stress timing in English," J. Phonetics **26**, 145–171.
Cutler, A., and Butterfield, S. (**1992**). "Rhythmic cues to speech segmentation: Evidence from juncture misperception," J. Mem. Lang. **31**, 218–236.
Das, T., Singh, L., and Singh, N. C. (**2008**). "Rhythmic structure of Hindi and English: New insights from a computational analysis," Prog. Brain Res. **168**, 207–272.
Davis, B. L., and MacNeilage, P. F. (**2002**). "The internal structure of the syllable: An ontogenetic perspective on origins," in *The Evolution of Language out of Pre-Language*, edited by T. Givon and Bertram F. Malle (Benjamins, Amsterdam), Chap. 5, pp. 133–151.
Dellwo, V. (**2008**). "The role of speech rate in perceiving speech rhythm," in *Proceedings of the 4th Conference on Speech Prosody*, Campinas, pp. 375–378.
Dellwo, V., and Wagner, P. (**2003**). "Relationships between rhythm and speech rate," in *Proceedings of the 15th International Congress of the Phonetic Sciences*, Barcelona, pp. 471–474.
Dolata, J. K., Davis, B. L., and MacNeilage, P. F. (**2008**). "Characteristics of the rhythmic organization of vocal babbling: Implications for an amodal linguistic rhythm," Infant Behav. Dev. **31**, 422–431.
Drullman, R., Festen, J. M., and Plomp, R. (**1994a**). "Effect of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Am. **95**, 2670–2680.
Drullman, R., Festen, J. M., and Plomp, R. (**1994b**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**, 1053–1064.
Flaugnacco, E., Lopez, L., Terribili, C., Zoia, S., Buda, S., Tilli, S., Monasta, L., Montico, M., Sila, A., and Ronfani, L. (**2014**). "Rhythm perception and production predict reading abilities in developmental dyslexia," Front. Hum. Neurosci. **8**, 1–14.
Fox, P. T., Ingham, R. J., Ingham, J. C., Hirsch, T. B., Downs, J. H., Martin, C., Jerabek, P., Glass, T., and Lancaster, J. L. (**1996**). "A PET study of the neural systems of stuttering," Nature **382**, 158–162.
Fraisse, P. (**1974**). *Psychologie du Rythme* (*Rhythm Psychology*) (Presses Universitaires de France, Paris).
Ghazanfar, A. A. (**2013**). "Multisensory vocal communication in primates and the evolution of rhythmic speech," Behav. Ecol. Sociobiol. **67**, 1441–1448.
Ghitza, O., and Greenberg, S. (**2009**). "On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence," Phonetica **66**, 113–126.
Giraud, A., Neumann, K., Bachoud-Levi, A., von Gudenberg, A. W., Euler, H. A., Lanfermann, H., and Preibisch, C. (**2008**). "Severity of dysfluency correlates with basal ganglia activity in persistent developmental stuttering," Brain Lang. **104**, 190–199.
Giraud, A., and Poeppel, D. (**2012**). "Speech perception from a neurophysiological perspective," in *The Human Auditory Cortex* (Springer, New York), Chap. 9, pp. 225–260.
Gracco, V. L. (**1988**). "Timing factors in the coordination of speech movements," J. Neurosci. **12**, 4629–4639.
Greenberg, S. (**2006**). "A multi-tier framework for understanding spoken language," in *Listening to Speech: An Auditory Perspective* (Laurence Erlbaum Associates, Mahwah, NJ), Chap. 25, pp. 411–433.
Grosjean, F., and Deschamps, A. (**1975**). "Analyse contrastive des variables temporelles de l'anglais et du français: Vitesse de parole et variables composantes, phénomènes d'hésitation" ("Contrastive analysis of temporal variables in English and French: Speaking rate and composing variables, hesitation phenomena"), Phonetica **31**, 144–184.
Grosjean, F., and Lane, H. (**1976**). "How the listener integrates the components of speaking rate," J. Exp. Psychol.-Hum. Percept. Perform. **2**, 538–543.
Grosse, P., Cassidy, M., and Brown, P. (**2002**). "EEG–EMG, MEG–EMG and EMG–EMG frequency analysis: Physiological principles and clinical applications," Clin. Neurophysiol. **113**, 1523–1531.
Hadar, U., Wenkert-Olenik, D., Krauss, R., and Soroker, N. (**1998**). "Gesture and the processing of speech: Neuropsychological evidence," Brain Lang. **62**, 107–126.
Hertrich, I., Dietrich, S., and Ackermann, H. (**2013**). "Tracking the speech signal—Time-locked MEG signals during perception of ultra-fast and moderately fast speech in blind and in sighted listeners," Brain Lang. **124**, 9–21.
Jacewicz, E., Fox, R. A., O'Neill, C., and Salmons, J. (**2009**). "Articulation rate across dialect, age, and gender," Lang. Var. Change **21**, 233–256.
Janse, E. (**2004**). "Word perception in fast speech: Artificially time-compressed vs. naturally produced fast speech," Speech Commun. **42**, 155–173.
Jerbi, K., Lachaux, J. P., N'Diaye, K., Pantazis, D., Leahy, R. M., Garnero, L., and Baillet, S. (**2007**). "Coherent neural representation of hand speed in humans revealed by MEG imaging," Proc. Natl. Acad. Sci. U.S.A. **104**, 7676–7681.
Kelso, J. A., Saltzman, E. L., and Tuller, B. (**1986**). "The dynamical perspective on speech production: Data and theory," J. Phonetics **14**, 29–59.
Kohler, K. J. (**2009**). "Rhythm in speech and language: A new research paradigm," Phonetica **66**, 29–45.
Levelt, W. J. (**1999**). "Models of word production," Trends Cogn. Sci. **3**, 223–232.
Lindblom, B. (**1983**). "Economy of speech gestures," in *The Production of Speech* (Springer, New York), Chap. 10, pp. 217–245.
Lu, C., Peng, D., Chen, C., Ning, N., Ding, G., Li, K., Yang, Y., and Lin, C. (**2010**). "Altered effective connectivity and anomalous anatomy in the basal ganglia-thalamocortical circuit of stuttering speakers," Cortex **46**, 49–67.
MacDougall, H. G., and Moore, S. T. (**2005**). "Marching to the beat of the same drummer: The spontaneous tempo of human locomotion," J. Appl. Physiol. **99**, 1164–1173.
MacNeilage, P. F. (**1998**). "The frame/content theory of evolution of speech production," Behav. Brain Sci. **21**, 499–511.
Marcus, S. M. (**1981**). "Acoustic determinants of perceptual center (P-center) location," Percept. Psychophys. **30**, 247–256.
Martin, J. G. (**1972**). "Rhythmic (hierarchical) versus serial structure in speech and other behavior," Psychol. Rev. **79**, 487–509.
Meireles, A. R., and Barbosa, P. N. A. (**2008**). "Speech rate effects on speech rhythm," in *Proceedings of the 4th Conference on Speech Prosody*, Campinas, pp. 327–330.

Meireles, A. R., and Gambarini, V. de P. (**2012**). "Rhythm typology of Brazilian Portuguese dialects," in *Proceedings of the 6th International Conference on Speech Prosody (Volume II)*, Shangai, pp. 474–477.
Moore, C. A., and Ruark, J. L. (**1996**). "Does speech emerge from earlier appearing oral motor behaviors?," J. Speech Lang. Hear. Res. **39**, 1034–1047.
Moore, C. A., Smith, A., and Ringel, R. L. (**1988**). "Task-specific organization of activity in human jaw muscles," J. Speech Lang. Hear. Res. **31**, 670–680.
Nazzi, T., and Ramus, F. (**2003**). "Perception and acquisition of linguistic rhythm by infants," Speech Commun. **41**, 233–243.
O'Dell, M., Lennes, M., Werner, S., and Nieminen, T. (**2007**). "Looking for rhythms in conversational speech," in *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrucken, pp. 1201–1204.
O'Dell, M. L., and Nieminen, T. (**2009**). "Coupled oscillator model for speech timing: Overview and examples," in *Nordic Prosody: Proceedings of the 10th conference*, Helsinki, pp. 179–190.
Patel, A. D. (**2005**). "The relationship of music to the melody of speech and to syntactic processing disorders in aphasia," Ann. N. Y. Acad. Sci. **1060**, 59–70.
Peelle, J. E., and Davis, M. H. (**2012**). "Neural oscillations carry speech rhythm through to comprehension," Front. Psychol. **3**, 1–17.
Petitto, L. A., Holowka, S., Sergio, L. E., and Ostry, D. (**2001**). "Language rhythms in baby hand movements," Nature **413**, 35–36.
Piitulainen, H., Bourguignon, M., De Tiege, X., Hari, R., and Jousmäki, V. (**2013**). "Corticokinematic coherence during active and passive finger movements," Neuroscience **238**, 361–370.
Poeppel, D., Idsardi, W. J., and van Wassenhove, V. (**2008**). "Speech perception at the interface of neurobiology and linguistics," Philos. Trans. R. Soc. Lond. B. Biol. Sci. **363**, 1071–1086.
Pulvermüller, F., and Fadiga, L. (**2010**). "Active perception: Sensorimotor circuits as a cortical basis for language," Nat. Rev. Neurosci. **11**, 351–360.
Ramus, F., Nespor, M., and Mehler, J. (**1999**). "Correlates of linguistic rhythm in the speech signal," Cognition **73**, 265–292.
Rosen, S. (**1992**). "Temporal information in speech: Acoustic, auditory and linguistic aspects," Philos. Trans. R. Soc. Lond. B. Biol. Sci. **336**, 367–373.
Ruspantini, I., Saarinen, T., Belardinelli, P., Jalava, A., Parviainen, T., Kujala, J., and Salmelin, R. (**2012**). "Corticomuscular coherence is tuned to the spontaneous rhythmicity of speech at 2–3 Hz," J. Neurosci. **32**, 3786–3790.
Saltzman, E., and Kelso, J. (**1987**). "Skilled actions: A task-dynamic approach," Psychol. Rev. **94**, 84–106.
Schaeffler, S., Scobbie, J. M., and Schaeffler, F. (**2014**). "Measuring reaction times: Vocalisation vs. articulation," in *Proceedings of the 10th International Seminar in Speech Production (ISSP 10)*, Cologne, pp. 379–382.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Shepherd, S. V., Lanzilotto, M., and Ghazanfar, A. A. (**2012**). "Facial muscle coordination in monkeys during rhythmic facial expressions and ingestive movements," J. Neurosci. **32**, 6105–6116.
Smith, A. (**1992**). "The control of orofacial movements in speech," Crit. Rev. Oral Biol. Med. **3**, 233–267.
Smith, A., Goffman, L., Zelaznik, H. N., Ying, G., and McGillem, C. (**1995**). "Spatiotemporal stability and patterning of speech movement sequences," Exp. Brain Res. **104**, 493–501.
Smith, A., Luschei, E., Denny, M., Wood, J., Hirano, M., and Badylak, S. (**1993**). "Spectral analyses of activity of laryngeal and orofacial muscles in stutterers," J. Neurol. Neurosurg. Psychiatry **56**, 1303–1311.
Smith, R., Rathcke, T., Cummins, F., Overy, K., and Scott, S. (**2014**). "Communicative rhythms in brain and behaviour," Philos. Trans. R. Soc. Lond. B. Biol. Sci. **369**, 20130389.
Smith, Z. M., Delgutte, B., and Oxenham, A. J. (**2002**). "Chimaeric sounds reveal dichotomies in auditory perception," Nature **416**, 87–90.
Sparrow, W. (**1983**). "The efficiency of skilled performance," J. Mot. Behav. **15**, 237–261.
Tilsen, S. (**2008**). "Relations between speech rhythm and segmental deletion," in *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, Chicago, Vol. 44, pp. 211–223.
Tilsen, S. (**2009**). "Multitimescale dynamical interactions between speech rhythm and gesture," Cogn. Sci. **33**, 839–879.
Tilsen, S., and Arvaniti, A. (**2013**). "Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages," J. Acoust. Soc. Am. **134**, 628–639.
Tilsen, S., and Johnson, K. (**2008**). "Low-frequency Fourier analysis of speech rhythm," J. Acoust. Soc. Am. **124**, EL34–EL39.
Tsao, Y., and Weismer, G. (**1997**). "Interspeaker variation in habitual speaking rate: Evidence for a neuromuscular component," J. Speech Lang. Hear. Res. **40**, 858–866.
Tsao, Y., Weismer, G., and Iqbal, K. (**2006**). "Interspeaker variation in habitual speaking rate: Additional evidence," J. Speech Lang. Hear. Res. **49**, 1156–1164.
Van Boxtel, A. (**2001**). "Optimal signal bandwidth for the recording of surface EMG activity of facial, jaw, oral, and neck muscles," Psychophysiology **38**, 22–34.
Voss, R. F., and Clarke, J. (**1975**). "'1/*f* noise' in music and speech," Nature **258**, 317–318.
Wohlert, A. B., and Hammen, V. L. (**2000**). "Lip muscle activity related to speech rate and loudness," J. Speech Lang. Hear. Res. **43**, 1229–1239.
Woodruff, C. K., White-Schwoch, T., Tierney, A. T., Strait, D. L., and Kraus, N. (**2014**). "Beat synchronization predicts neural speech encoding and reading readiness in preschoolers," Proc. Natl. Acad. Sci. U.S.A. **111**, 14559–14564.