Check for updates

Original Article

Language and Speech

Language *and* Speech
1–21
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00238309231217689
journals.sagepub.com/home/las
Sage

The Attractiveness of Average Speech Rhythms: Revisiting the Average Effect From a Crosslinguistic Perspective

Constantijn Kaland
Institute of Linguistics, University of Cologne, Germany

Marc Swerts
Department of Communication and Cognition, Tilburg University, The Netherlands

Abstract

The current study investigates the average effect: the tendency for humans to appreciate an averaged (face, bird, wristwatch, car, and so on) over an individual instance. The effect holds across cultures, despite varying conceptualizations of attractiveness. While much research has been conducted on the average effect in visual perception, much less is known about the extent to which this effect applies to language and speech. This study investigates the attractiveness of average speech rhythms in Dutch and Mandarin Chinese, two typologically different languages. This was tested in a series of perception experiments in either language in which native listeners chose the most attractive one from a pair of acoustically manipulated rhythms. For each language, two experiments were carried out to control for the potential influence of the acoustic manipulation on the average effect. The results confirm the average effect in both languages, and they do not exclude individual variation in the listeners' perception of attractiveness. The outcomes provide a new crosslinguistic perspective and give rise to alternative explanations to the average effect.

Keywords

Rhythm, average effect, prosody, Dutch, Mandarin Chinese

1 Introduction

Research has shown repeatedly that humans perceive average faces as more attractive than the original faces that were used to generate the average (e.g., Baudouin & Tiberghien, 2004; Galton, 1879; Langlois & Roggman, 1990). This effect is often referred to as the average effect and also

Corresponding author:
Constantijn Kaland, Institute of Linguistics, University of Cologne, SFB 1252, Luxemburger Straße 299, Köln 50939, Germany.
Email: ckaland@uni-koeln.de

holds for the attractiveness of birds, fish, and cars (Halberstadt & Rhodes, 2003). Studies have sought explanations for this effect in evolutionary biology and human cognition, so far without conclusive results.

Surprisingly few works investigated the average effect beyond visual perception. It has been shown that averaging increases the attractiveness of musical pieces (Repp, 1997) and of voices (Bruckert et al., 2010; Feinberg et al., 2008). It is important to extend this line of research to language, in particular to prosodic phenomena such as intonation, stress, and rhythm. Prosody varies widely across languages (e.g., Gussenhoven & Chen, 2020) and so does the close interaction of prosody and facial expressions (e.g., Borrás-Comes et al., 2014; Crespo Sendra et al., 2013). It is therefore by no means clear whether the average effect extends to the domain of prosody in different languages.

To shed light on this issue, the current study investigates to what extent the average effect holds for speech rhythm. There are several prosodic cues contributing to perceived rhythm. The current paper mainly focuses on duration as a rhythm cue. Rhythm, although there is still a debate on what the best metric is to capture it (e.g., Nolan & Jeon, 2014), has often been argued to vary across languages. This study reports four perception experiments that presented acoustically manipulated rhythms of Dutch and Mandarin Chinese to native listeners of those respective languages. The fact that these languages are prosodically quite different could therefore shed light on the extent to which a possible average effect in speech rhythm has general validity. The experiments consisted of original rhythms as produced by individual speakers and of a rhythm in which all syllable and pause durations were averaged over the speakers.

1.1 Average effects in the visual domain

Most likely the first report of the average effect is a study in which several frontal portrait photographs of prisoners were superimposed, in order “to elicit the principal criminal types” (Galton, 1879, p. 135). However, the resulting composite image did not show stereotypical “villainous” traits in the faces of the photographed men. “All composites are better looking than their components, because the averaged portrait of many persons is free from the irregularities that variously blemish the looks of each of them” (Galton, 1879, p. 135).

Although the evaluation by the author of the first study is rather impressionistic, the effect has been replicated many times in more recent work using modern techniques to generate the average and using experimentally elicited judgments by naive participants. Again and again, it was shown that when images of faces were averaged, the average face was perceived as more attractive than the original faces used to generate the average.

One technique used in several studies (e.g., Langlois et al., 1994; Langlois & Roggman, 1990) first aligned grayscale photographs of faces by the eye pupils and the middle of the lipline. Each photograph was represented by a square of 512 × 512 gray values (pixels). The average was then generated by taking the mean gray value for each pixel. Multiple averages were generated, based on 2, 4, 8, 16, or 32 photographed faces (predominantly Caucasian). The results revealed that the rated attractiveness increased with increasing numbers of faces used to compute the average (reaching significance with 16 and 32 faces).

The explanation for this effect was initially sought in evolutionary biology, in particular in partner selection. The idea is that humans seek a partner that gives the best chances for successful reproduction, and therefore tend to be attracted by prototypical (non-outlying) features. Under this view, facial averageness and symmetry are a reflection of genetic health (e.g., Thornhill & Gangestad, 1993).

The biological account was furthermore supported by the finding that average faces are attractive cross-racially, obtaining similar attractiveness ratings from Chinese participants on average faces based on Chinese faces, Caucasian faces, and mixed faces (Rhodes et al., 2001). Studies also found the average effect based on images of dogs, wristwatches, birds, fish, and automobiles (Halberstadt & Rhodes, 2000, 2003). Crucially, familiarity with these categories explained the effect for all of them. When the effect of familiarity was statistically controlled for, the "true" effect of the average was only found for humans and animals, not for objects, which was interpreted as an indication that humans indeed have a genetic selection mechanism underlying this effect (Halberstadt & Rhodes, 2003).

The actual contribution of averageness to the average effect was also investigated in other studies by contrasting averageness with other aspects such as symmetry, health, and enhancement of particular facial features. Although facial symmetry contributes to perceived attractiveness, there is an independent effect of the average that explained the results found for human faces (Baudouin & Tiberghien, 2004; Valentine et al., 2004). These studies also argued that symmetry does not signal health, that is, weakening the evolutionary biological account of the average effect. It was, furthermore, found that specific facial features such as big eyes or a small chin made female faces more attractive, independent of the average effect (Baudouin & Tiberghien, 2004). Enhancement of specific features in an average face even increased perceived attractiveness (DeBruine et al., 2007). It therefore seems that the average cannot explain all the perceived visual attractiviness. Further evidence for the average effect was found in auditory studies, as further discussed in the following subsection.

1.2 Average effects in the auditory domain

Compared with the visual domain, much fewer studies investigated the average effect in the auditory domain. For example, the effect was investigated for average piano performances of a musical piece (Repp, 1997). In one experiment, the mathematical average of the pitches of the notes, timings of note onsets and offsets, relative intensities, and pedal onsets and offsets were taken from 10 student performances of a piano piece. This was done using a digitalized representation of the performance (MIDI; Smith & Wood, 1981). The average performance was rated second most attractive, in an overall narrow range of attractiveness scores. It was hypothesized that these resulted from the relative similarity of the student performances, that is, little individual differences in expressive timing. In a second experiment, actual waveform recordings were used, taken from student *and* expert performances. The average performance was generated based on timing only, using the note interonset intervals. Results showed that the experts' average performance was rated as most attractive, confirming that the average effect in music was stronger with more individual variation among the pieces from which the average was generated.

Other work investigated the contribution of averageness to the attractiveness of women's voices (Feinberg et al., 2008). The study was in part motivated by a previous study showing that women with attractive voices tend to have attractive faces as well (Collins & Missing, 2003). The attractiveness ratings were given to different vowels, which were acoustically manipulated for F0 and had a constant duration of 500 ms and a normalized amplitude of 87.5 dB. A positive linear correlation was found between F0 level and attractiveness, that is, the higher the pitch of the voice, the higher the perceived attractiveness. In a second experiment the voices were categorized according to F0 level at vowel start: either low (200 Hz), average (220 Hz), or high (241 Hz). For each category, F0 manipulations with lowered and raised pitch were generated, to test whether raising in each category was preferred equally, or whether averageness affects attractiveness in any way. Again, it was found that raising F0 leads to more attractiveness in all categories, which was interpreted as an effect of femininity. That is, a higher pitch is preferred in women, even if it exceeds the average level. Thus, these findings are in line with the ones showing that enhancing specific facial features (above the average) leads to even higher perceived attractiveness (DeBruine et al., 2007).

Using a different auditory averaging method, attractiveness ratings were elicited for male and female voices producing the syllable "had" (Bruckert et al., 2010). The averaging was done using auditory morphing (Kawahara & Matsui, 2003), in which five parameters were independently manipulated to generate the average: F0, formant frequency, duration, spectrotemporal density (harmonics-to-noise ratio [HNR]), and aperiodicity. The first three are voice shape related, that is, mainly prosodic, whereas the latter two are voice texture related, that is, how smooth a voice sounds with more periodicity and higher HNR leading to a smoother voice. In three subsequent experiments, it was shown that these categories of parameters have independent contributions to the average effect. That is, averaging led to voices that were closer to the population mean as well as to smoother voices (with reduced aperiodic noise). These auditory findings were interpreted as analogous to the ones for faces. That is, faces were also found to be be more similar to the prototypical one and have a smoother skin texture when averaged. The separation of these two independent factors was also the basis for a study that investigated the neural correlates of attractiveness in voices (Weiss et al., 2021), that is, distance-to-mean and smoothness (HNR). In particular, it was investigated whether these two parameters explained all measured auditory brain activity. When controlling for the variance explained by distance-to-mean and HNR, remaining activity in an area related to auditory working memory was still found. It was tentatively concluded that unattractive voices might demand more neural processing effort than attractive voices. These findings suggest that overall intelligibility might play a role in attractiveness.

1.3 The current study: rhythm

The auditory studies on the average effect investigated music and voice characteristics independently of language. It has been shown that languages differ substantially in their use of "musical" aspects, such as the ones found in prosody. These relate to all acoustic aspects that are not traditionally analyzed as part of the segments (i.e., vowels and consonants), such as intonation, stress, and rhythm. These aspects not only contribute to the attractiveness of the speaker's voice, but also to the attractiveness of the speaker. It has been shown that prosodic characteristics contribute to the likability, charisma, and sex-appeal of speakers (e.g., Belin, 2021). While intonation and stress differ widely across languages, there are different views on the extent to which their rhythm differs.

Traditionally, rhythm in language has been studied under the assumption that there are clear crosslinguistic rhythmic differences. Much work has been devoted to the search for equally timed intervals in speech (isochrony; i.e., Lehiste, 1977). The original idea was that languages could be grouped into rhythm classes, that is, the distinction between languages that show a tendency to space syllables equally in time (syllable-timed), such as Spanish, and languages that show a tendency to space stress-feet equally in time (stress-timed), such as English (Pike, 1945). However, little consistent support was found for the metrics proposed to quantify these rhythm classes (e.g., Arvaniti, 2012). It has, furthermore, been argued that rhythm in language is not strictly isochronic in the sense that it shows regular timing (Nolan & Jeon, 2014). Other typological accounts have indeed backgrounded the importance of temporal cues to rhythm and focused more on tonal patterns. Recent proposals were done to classify languages based on an overall tonal rhythm (macro-rhythm; Jun, 2014b). Macro-rhythm constitutes the global perceived rhythmicity of phrases, which is the product of prosodic events at the phrase level as well as the word level (e.g., pitch accents and word stress, respectively). The attempts to quantify macro-rhythm are few and require more research (e.g., Kaland, 2022; Polyanskaya et al., 2020; Prechtel, 2020).

Other studies on speech rhythm have focused on separating rhythm aspects that are similar from ones that are different crosslinguistically. It appears that languages are quite similar when their rate of intensity fluctuations is calculated in a modulation spectrum (Ding et al., 2017; Varnet et al., 2017). A modulation spectrum correlates with the way the auditory nerve processes speech (i.e., decomposed into narrow frequency bands and low-pass filtered), in particular with the processing of syllables. Research has shown that modulation spectra of speech play an important role in speech intelligibility. It was demonstrated that heavily (digitally) compressed speech became significantly more intelligible when it was manipulated by adding an artificial rhythm of regular amplitude modulations to it (e.g., Bosker & Ghitza, 2018). Quantifying speech rhythmicity using amplitude modulations, furthermore, showed useful to assess speaker charisma (Bosker, 2021), with more rhythmical speech being perceived as more attractive.

Modulation spectra are also able to reveal rhythmic differences across languages (Varnet et al., 2017), which are undisputed (e.g., Gussenhoven & Chen, 2020; Jun, 2014a). The two languages investigated in the current study are Dutch (iso: NLD) and Mandarin Chinese (iso: CMN). They are rhythmically different, regardless of which rhythm classification one follows (e.g., Arvaniti, 2009; Jun, 2014b for overviews). That is, Dutch is traditionally analyzed as stress-timed (e.g., Grabe & Low, 2002) and is medium macro-rhythmic due to a variety of pitch accents used to mark phrases (e.g., ToDI Collective, 2019). Mandarin Chinese is traditionally analyzed as syllable-timed (e.g., Mok & Dellwo, 2008) and is weakly macro-rhythmic as its stresses are acoustically not prominently marked and F0 is mainly used to distinguish lexical tones (e.g., Beckman & Venditti, 2010).

1.4 Research question

Summing up the discussed literature, research has shown that the average effect is found in visual and auditory perception. The mechanisms underlying this effect are not fully understood, but it seems clear that the generated average has a unique effect on human perception of attractiveness. In particular in the auditory domain, more research is needed that takes into account linguistically representative speech. That is, in the studies conducted so far, either vowels or single syllables were used as stimulus material (Bruckert et al., 2010; Feinberg et al., 2008). They provide little insight into how humans speak. It is therefore important to use linguistically more complex units, such as phrases. The question arises whether the average effect applies to speech rhythm in phrases and it is not a priori clear whether this result holds for typologically different languages. On one hand, rhythm tends to signal speaker charisma and likeability, which contribute to perceived attractiveness, potentially in similar ways across languages (and cultures). On the other hand, languages differ rhythmically depending on their prosodic structure and conceptualizations of attractiveness could differ among culturally different groups of listeners. Given these differences, the average effect might not hold across languages. The research question investigated in this study is therefore:

Is there a difference in the average effect between two rhythmically different languages such as Dutch and Mandarin Chinese?

The research question is investigated using perception experiments in which rhythmically different versions of the first sentence of the Dutch and Mandarin Chinese versions of the fable *The North Wind and The Sun* were presented to native listeners of either language. The rhythms were either the ones originally produced by native speakers, or a generated average rhythm based on the rhythms of the individual speakers. The methodology is outlined in further detail in the following section.

2 Methodology

The methodological choices were identical for Dutch (NLD) and Mandarin Chinese (CMN). Where the methodologies are different between the languages, this is stated explicitly in the text.

2.1 Recordings and processing

For each language, eight speakers were asked to produce the first sentence of the Aesop fable *The North Wind and the Sun* (Jacobs, 1894); see Example 1. The sentences were recorded from five female and three male speakers in each language. Although they were native speakers of the language, they had varying linguistic backgrounds and some regional variation could therefore not be excluded.

Example 1:

NLD De Noordenwind en de Zon waren erover aan het redetwisten wie de
The North-Wind and the Sun were about on it argue who the
Sterkste was van hun beiden.
Strongest was of them both.

CMN 有一回,北风和太阳正在争论谁的本事大。
you yi hui bei feng he tai yang zheng zai zheng lun shui de ben shi da.

The North Wind and the Sun were arguing which of them was stronger.

The recordings were segmented at the syllable level in Praat (Boersma & Weenink, 2022) by native speakers of the respective languages who received phonetic training to determine (syllable) boundaries. The resulting segmentations consisted of both syllables and pauses. Pauses were segmented if they were produced by at least one speaker of the respective language. For speakers who did not produce a pause at a specific location in the sentence, an interval approximating zero length was inserted (e.g., 0.000001 ms), such that the total number of intervals was identical across all speakers of a language. The identical number of intervals across speakers allowed a uniform calculation of the average rhythm, that is, by dividing the total duration by the number of speakers for each syllable and pause (see below). In total, the segmented sentences consisted of 27 syllables and 4 pause locations for Dutch, and of 17 syllables and 7 pause locations for Mandarin Chinese.

For each of the eight sentences per language all interval durations (syllables and pauses) in milliseconds (ms) were measured. The average rhythm per language was computed by taking the mean duration of all eight instances of each interval. Thus, pause durations in the average rhythm were calculated by summing all individual pause interval durations (even if they were zero) and dividing them by eight.

2.2 Speaker selection

To avoid that the attractiveness of the rhythm was influenced by the voice characteristics of the different speakers, both the average rhythm and the individual speaker rhythms were generated based on the sentence (Example 1) of a single speaker per language (henceforth, carrier phrase). The carrier phrase was manipulated such that each individual rhythm and the average rhythm were obtained from it by means of acoustic (duration) manipulation. The speaker of the carrier phrase was chosen for

Table 1. Total Durations, Mean Interval Durations (Syllables and Pauses), and Mean Absolute Deviations re. Average Rhythm (all in ms) Per Rhythm (Speaker) and Per Language.

Language	Speaker/rhythm	Total dur.	M interval dur.	M |deviation|	
NLD	1	6,152.65	198.47	47.97	f
	2	5,635.74	181.80	30.05	
	3	5,030.50	162.27	26.82	
	4	4,793.47	154.63	32.49	
	5	5,700.75	183.90	24.58	c
	6	5,988.25	193.17	34.35	
	7	6,351.70	204.89	37.85	
	8	5,354.13	172.71	23.08	
	AVG	5,615.21	181.14	0	
CMN	1	3,272.38	136.35	28.27	
	2	4,306.49	179.44	43.19	
	3	3,044.98	126.87	27.72	
	4	3,498.43	145.77	21.05	
	5	3,598.31	149.93	17.96	c
	6	3,518.55	146.61	44.11	f
	7	2,896.41	120.68	36.94	
	8	3,961.48	165.06	27.13	
	AVG	3,512.13	146.34	0	

Note. c = speaker chosen as closest to the average, f = speaker chosen as furthest from the average, AVG = average, NLD = Dutch, CMN = Mandarin Chinese.

each language from the group of eight speakers. This was done based on the amount of regional variation (as assessed informally by the experimenters) and how close that speaker's rhythm was to the average rhythm (based on acoustic measures). The closeness of the individual rhythms to the average was taken as a quantification of how much acoustic manipulation would be needed to convert that speaker's rhythm into any of the target rhythms. This was computed by taking the mean absolute deviation (in ms) of the individual speakers' intervals relative to the intervals of the average rhythm (see Table 1). Dutch Speaker #8 deviated the least from the average; however, that speaker had a stronger regional accent compared with the second smallest deviating speaker (#5). The latter was therefore chosen as the speaker whose rhythm was closest to the average. Speaker #1 was chosen as the one whose rhythm was furthest away from the average. Mandarin Chinese Speakers #5 and #6 had the rhythms with the least (closest) and most (furthest) deviation from the average, respectively.

Two separate experiments per language were designed; one using stimuli based on the "closest" speaker and one using stimuli based on the "furthest" speaker (NLD.closest, NLD.furthest, CMN.closest, CMN.furthest). This was done to control for any effects of the amount of acoustic manipulation on the attractiveness results. Thus, choosing a speaker whose rhythm lies close to the average is intuitive from the perspective of preserving the naturalness of the individual rhythms, that is, with overall the least amount of acoustic manipulation. However, with such a choice the amount of acoustic manipulation needed for each rhythm correlates negatively with the hypothesized attractiveness (more manipulation—less attractive). Thus, the results of such a setup alone would not allow to disentangle whether listeners based their attractiveness ratings on the degree of acoustic manipulation or on actual attractiveness due to the average effect. Therefore, a second version of the experiment was designed using the speaker whose rhythm was the furthest away from the average.

With the latter setup the amount of acoustic manipulation needed would be overall more; however, there would be no confound of the amount of manipulation and the hypothesized attractiveness.

2.3 Participants

In total, 80 participants took part (20 per experiment). Note that this number is a close match to the required sample size in a two-group design (Kenny, 1987; p. 214; Table 13.1; Cohen's $d = .5$, $\alpha = .05$, power=.90). The design of the current study and the availability of participants also affected the pragmatic choice for this number of participants. They were all native speakers of the respective languages and did not have hearing problems. No participants took part in more than one experiment. NLD.closest: 11 F/9 M, *M* age=46, age range 23–78; NLD.furthest: 9 F/11 M, *M* age=30, age range 19–56; CMN.closest: 17 F/3 M, *M* age=22, age range 19–27; CMN.furthest: 11 F/9 M, *M* age=24, age range 19–28.

2.4 Stimulus material

On the basis of each carrier phrase, eight rhythms were generated for each language (NLD and CMN) and each experiment (closest and furthest). That is, using acoustic manipulation in Praat (Boersma & Weenink, 2022; To manipulation . . .) the interval durations of the carrier phrase were shortened or lengthened such that they matched the interval durations of each of the other speakers (i.e., not being the ones chosen in that carrier phrase, *N*=7) and the average rhythm. F0 was kept unmanipulated during the duration manipulations. The average rhythm was recalculated based on the other seven speakers. In this way, the original rhythm of the speaker of the carrier phrase was not represented in the stimulus material and did not affect the calculation of the average rhythm. Note that inclusion of the rhythm of the speaker of the carrier phrase would not have required any acoustic manipulation and could have therefore undesiredly promoted that rhythm as more attractive than the other (manipulated) ones.

In each experiment (NLD.closest, NLD.furthest, CMN.closest, CMN.furthest) the eight stimuli were combined into pairs. The pairs consisted of all possible combinations of two members of a single set of eight stimuli (8 × 8=64 stimulus pairs per experiment). This was done to obtain ratings based on which of the two rhythms in a pair was chosen as the most attractive. Note that the pairs for which both members were identical were included as dummy stimuli and discarded before analysis (see below).

2.5 Procedure

The stimulus pairs were presented to participants in an online fashion using Qualtrics (Qualtrics International Inc., 2020). For each stimulus pair, an html-page was generated with the question, "Which of the following two utterances sounds the best rhythmically?" and a written version of the phrase in Example 1. Buttons were presented underneath the text to play each of the stimuli in the pair. Participants could then select the most attractive rhythm of the pair using an html radio button. They could listen to the stimuli as often as needed and change their response. After proceeding to the next stimulus pair, their choice could not be altered any longer. The stimulus pairs were presented in a different random order for each participant.

Participants received instructions and were presented an example stimulus before the actual experiment. They were instructed to do the experiment in a quiet room with headphones. A single experiment lasted approximately 20 minutes per participant. For each stimulus pair the participants' choice was saved on a webserver.

2.6 Data analysis

Participants' choices were converted into an attractiveness score per rhythm. This was done for the stimulus pairs in which the members were non-identical, that is, pairs for which participants had to choose the most attractive rhythm from two identical ones were discarded. In this way, 56 pairs per experiment were left for processing (64 pairs − 8 identicals, see Appendix A). The attractiveness score was computed based on the number of times a certain rhythm occurred as a member in a pair, that is, the number of times a rhythm could have been chosen as the most attractive one in a pair (14 times). The attractiveness score per rhythm was then calculated by dividing these counts by 14, such that the scale ran from 0 (never chosen as attractive) to 1 (always chosen as attractive).

The attractiveness scores were statistically analyzed using linear mixed effects models (LMM) in R (R Core Team, 2022; R Studio Team, 2022). One LMM (using the package lmerTest; Kuznetsova et al., 2017) per language was run with *attractiveness score* as response (320 observations per language), with *rhythm* (8 levels: 1–7 and average as reference level) and *experiment* (2 levels: closest, furthest) as fixed sum-coded factors (Brehm & Alday, 2022), and with *participant* as random intercept. Pairwise comparisons (Holm–Bonferroni corrected; Holm, 1979) between the average rhythm and all other individual rhythms were computed using the package emmeans (Lenth, 2023; Searle et al., 1980). Interactions were computed using a different model structure. That is, the random intercept was removed to avoid overfitting and rank deficiency. The LM thus had *attractiveness score* as response, and the interaction between *rhythm* and *experiment* as fixed sum-coded factors.

In addition to the attractiveness scores, the individual variation in the participants' choices was assessed by counting which rhythm was chosen most often per participant. For all participants, there was only one most often chosen rhythm (no shared "first places"). Then it was counted per rhythm how many participants chose that rhythm most often.

3 Results

The mean attractiveness scores for each language and each experiment as well as the number of participants that chose a particular rhythm most often are given in Figure 1 and Table 2. The results of the pairwise comparisons per language are given in Table 3.

For Dutch, the average rhythm showed the highest attractiveness scores in both the "closest" and the "furthest" experiments. The pairwise comparisons revealed that the average rhythm was rated higher than all other individual rhythms. This difference was signifant for all rhythm comparisons, except the ones with Rhythms 2 and 8 (Table 3). The factor experiment showed a trend in that overall lower scores were obtained in the "furthest" version than in the "closest" version ($\beta = .02$, $SE = 0.01$, $df = 310$, $t = 1.71$, $p = .09$). Figure 1 and Table 2, furthermore, show that the scores obtained in the "furthest" experiment lie closer to the average rhythm and have overall larger standard deviations. The interactions with experiment were significant for Rhythms 3, 6, and 8, in that attractiveness scores were lower for those rhythms in the "furthest" experiment. The interaction with experiment was also significant for Rhythm 7, showing an opposite effect (higher scores in the "furthest" experiment). The Dutch results also show that the average rhythm was chosen most often by 19 out of 40 participants, whereas the other most often chosen rhythms counted clearly less participants. Rhythms 4 and 6 were never the most often chosen ones.

For Mandarin Chinese, the average rhythm had the highest attractiveness score in the "closest" experiment, whereas it was thirdmost attractive in the "furthest" experiment (after Rhythms 5 and 1). The pairwise comparisons showed that, except for Rhythm 5, the average rhythm was rated

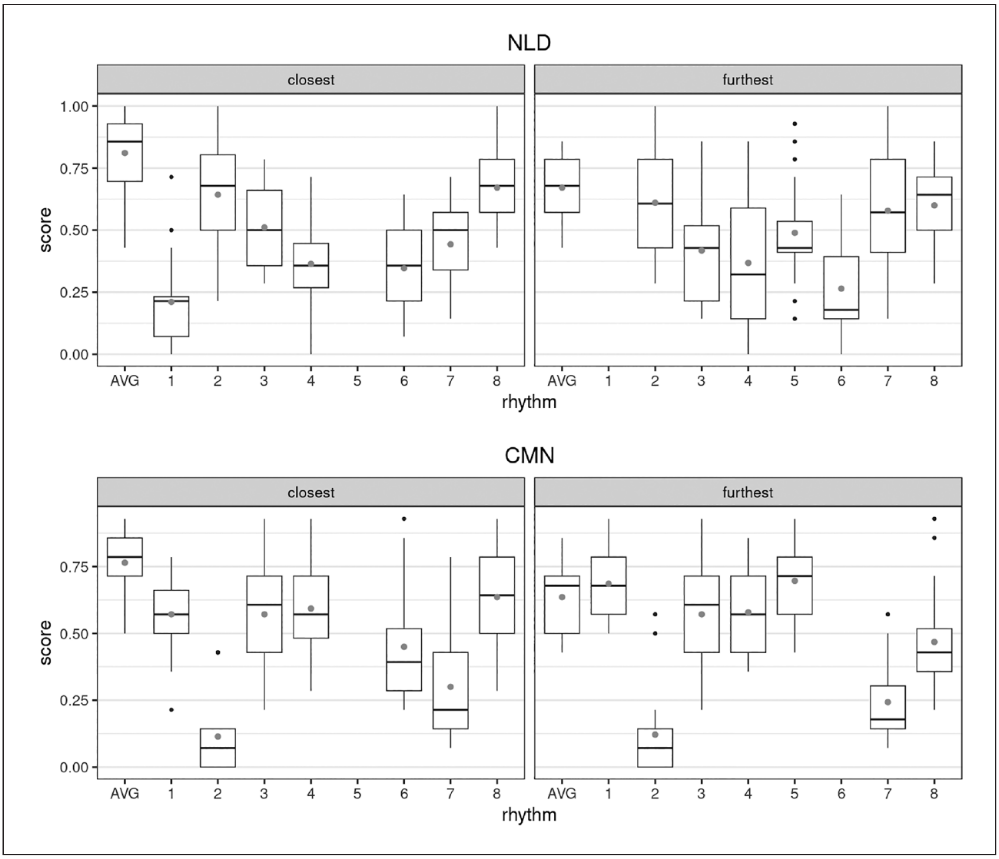

Figure 1. Boxplots for Dutch (top) and Mandarin Chinese (bottom) attractiveness scores split by experiment (closest/furthest). Gray dots indicate mean scores (Table 1). NLD = Dutch, CMN = Mandarin Chinese.

higher than any of the individual rhythms. This difference was significant, except for Rhythms 1 and 5 (Table 3). The factor experiment showed a trend in that overall lower scores were obtained in the "furthest" version than in the "closest" version (β = .02, *SE* = 0.01, *df* = 310, *t* = 1.75, *p* = .08). The interaction with experiment was significant for Rhythm 1, showing an opposite effect: lower scores in the "closest" experiment. The average rhythm was the most often chosen rhythm for 12 out of 40 participants. Except for Rhythm 2 (never the most often chosen one) the other rhythms were most often chosen by 6 or less participants.

4 Discussion and conclusion

Although the results seem to strongly support the average effect in speech rhythm, they were not unanimous when comparing the experimental outcomes. The rhythm in which the average interval duration was taken from seven speakers was perceived as the most attractive one in three out of four experiments and was significantly higher rated compared with six other rhythms in each language. In Dutch the average was the most attractive, regardless of whether the original rhythm of the speaker of the carrier phrase was closest or furthest away from the average rhythm. For

Table 2. Mean Attractiveness Scores (and *SD*s) Per Rhythm Split for Language (NLD/CMN) and Experiment (Closest/Furthest) and Number of Participants Choosing a Rhythm Most Often.

	Rhythm	Closest	Furthest	Most chosen by
NLD	1	0.21 (0.18)	▨	1
	2	0.64 (0.22)	0.61 (0.21)	7
	3	0.51 (0.17)	0.42 (0.21)	2
	4	0.36 (0.19)	0.37 (0.25)	0
	5	▨	0.49 (0.21)	4
	6	0.35 (0.16)	0.26 (0.20)	0
	7	0.44 (0.15)	0.58 (0.24)	4
	8	0.67 (0.16)	0.60 (0.15)	3
	AVG	0.81 (0.19)	0.67 (0.13)	19
CMN	1	0.57 (0.14)	0.69 (0.14)	6
	2	0.11 (0.15)	0.12 (0.16)	0
	3	0.57 (0.19)	0.57 (0.19)	6
	4	0.59 (0.16)	0.58 (0.14)	5
	5	▨	0.70 (0.13)	4
	6	0.45 (0.21)	▨	2
	7	0.30 (0.19)	0.24 (0.15)	1
	8	0.64 (0.18)	0.47 (0.19)	4
	AVG	0.76 (0.14)	0.64 (0.13)	12

Note. Shaded cells represent the rhythms from the speaker of the carrier phrase (absent in the design). Visual representation in Figure 1. AVG = average, NLD = Dutch, CMN = Mandarin Chinese.

Mandarin Chinese, however, the average rhythm was the most attractive one only when the original rhythm of the speaker of the carrier phrase was closest to the average. When the original rhythm of the speaker of the carrier phrase was furthest away from the average, the average rhythm was chosen as the thirdmost attractive (Table 2).

The question arises, to what extent are the outcomes of the “furthest” experiment from Mandarin Chinese due to a genuine linguistic difference? The answer to this question is not a priori clear. It is important to observe that there were notable differences in the stimulus materials, such as overall shorter phrases, shorter interval durations, and faster speech rates in Mandarin Chinese than in Dutch. The mean absolute deviations from the average rhythm did not differ much between the languages. Thus, it can be reasoned that similar amounts of acoustic manipulation per interval have a higher impact on shorter phrases (Mandarin Chinese) than on longer ones (Dutch). The direction of such an effect on the averaging, however, remains open to speculation.

It is, furthermore, clear from the stimulus materials that the “closest” and “furthest” rhythms did not share the same characteristics in both languages. Although the main criterion for selection as stimulus material—absolute deviation from the mean—was met in both languages, the total and interval durations differed. That is, the Dutch “closest” speaker (Speaker #5) was not only close to the average rhythm in terms of absolute deviation, that speaker also had a total phrase duration and interval duration that were both relatively close to those measures taken from the average rhythm. The same is true for the Dutch “furthest” speaker (Speaker #1). This speaker’s total and interval duration is relatively far away from those measures taken from the average rhythm. For Mandarin Chinese, however, the “closest” and “furthest” speaker both have total and interval durations that are close to those of the average rhythm. Thus, the rhythms in the Mandarin Chinese group of

Table 3. Results of the Pairwise Comparisons Between the Average Rhythm (AVG) and all other Individual Rhythms (1–8) Per Language.

Language	Comparison	β	*SE*	*df*	*t*	*p*
NLD	AVG—1 (f)	.55	0.05	272	10.10	<.001
	AVG—2	.11	0.04	272	2.63	.11
	AVG—3	.28	0.04	272	6.37	<.001
	AVG—4	.38	0.04	272	8.63	<.001
	AVG—5 (c)	.23	0.05	272	4.26	<.001
	AVG—6	.44	0.04	272	10.03	<.001
	AVG—7	.23	0.04	272	5.30	<.001
	AVG—8	.11	0.04	272	2.43	.16
CMN	AVG—1	.07	0.04	272	1.90	.46
	AVG—2	.58	0.04	272	15.51	<.001
	AVG—3	.13	0.04	272	3.43	<.05
	AVG—4	.11	0.04	272	3.05	<.05
	AVG—5 (c)	−.01	0.05	272	-0.30	1.00
	AVG—6 (f)	.27	0.05	272	5.69	<.001
	AVG—7	.43	0.04	272	11.42	<.001
	AVG—8	.15	0.04	272	3.95	<.01

Note. AVG = average, NLD = Dutch, CMN = Mandarin Chinese.

speakers did not vary as much as the Dutch ones. This difference is likely to affect the attractiveness ratings in that they lie closer to each other, that is, with smaller differences between the average rhythm and the other rhythms. This is exactly what can be seen in Table 1; the difference in attractiveness score between the average rhythm and the closest other rhythm is 0.14 and 0.07 for the Dutch experiments, respectively, whereas the difference is 0.12 and 0.05 for the Mandarin Chinese experiments, respectively. Note that this observation is also affected by the trend that the factor experiment showed for both languages. That is, the rating differences were smaller in part due to the choice of speaker for the carrier phrase, indicating that the amount of manipulation played a (marginal) role in the attractiveness ratings. In addition, it cannot be excluded that other acoustic differences in the stimulus materials affected the attractiveness scores. For example, intelligibility might have co-varied with the temporal characteristics reported in Table 1; that is, fast rhythms being less intelligible than slow rhythms. In addition, the specific F0 patterns in the rhythms were left uncontrolled. Although informal assessment did not reveal major differences between the speakers, we cannot exclude that even subtle F0 contour differences affected the rhythm perception. It remains to be investigated how F0 might contribute to the average effect in speech rhythm.

Related to the differences in the stimulus material is the question of whether the applied acoustic manipulations indeed targeted speech rhythm and not speech rate. This question is challenging to answer due to the fact that the two are intertwined. That is, speech rates differ across languages and rhythmic differences determine the extent to which speakers can vary their speech rate (e.g., Dellwo, 2008; Dellwo & Wagner, 2003). The acoustic manipulations applied in the current study were such that the speech rate was averaged across the speakers. However, the manipulations were done at the syllable level, a unit that plays an important role in all of the traditional rhythm classes (e.g., Arvaniti, 2012) and its analysis as a meaningful unit in speech is supported by phonetic research (e.g., Albert et al., 2018). We therefore believe to have not only manipulated speech rate, but also the potentially intricate relative timings of syllables that contribute to language-specific rhythm. It could, nevertheless, be the case that the observed differences between the Dutch and Mandarin Chinese results are indirectly the result of rhythm differences between the languages, that is, making the average effect come out stronger from the Dutch stimuli than from the Mandarin Chinese ones. Note that from this observation it can be concluded that the current results do not fit in a traditional stress- (Dutch) versus syllable-timing (Mandarin Chinese) classification. The acoustic manipulation of rhythm, as operationalized in this study by varying syllable durations, would be expected to have a larger impact in Mandarin Chinese than in Dutch. Subsequently, it could be expected that the speaker selection difference in the experiments in either language (closest vs. furthest) showed clearer effects on the rated attractiveness for Mandarin Chinese than Dutch, as syllable-timing in the former language was potentially more affected by the acoustic manipulations than in the latter language. This expectation could not be confirmed by the current results. The outcomes of this study rather show an opposite effect with larger differences in attractiveness scores between the experiments in Dutch than in Mandarin Chinese (Table 2).

Thus, it can be concluded that the average effect is present in the results of both languages. This conclusion is furthermore supported by Table 2 showing that the number of participants most often choosing a particular rhythm is the highest for the average and clearly lower for all other rhythms in both languages. This observation indicates that the average rhythm was most often the rhythm that was chosen over all other rhythms. To come back to the research question in Section 1.4; the average effect applies to speech rhythm and holds across Dutch and Mandarin Chinese. The fact that these languages are different for their prosody, including their rhythm, did not seem to prevent the average rhythm to be rated as more attractive than the other rhythms.

To sum up, this study has shown that the perceived attractiveness of speech rhythm is driven by similar mechanisms as found for faces and voices. Following the evolutionary account of the average effect, how could the current results be explained? One important aspect of speech rhythm is timing, which can be taken as a direct reflection of articulatory motor control and thus of brain–muscle coordination in the speaker. Fluent, prototypical (i.e., close-to-mean) rhythms could therefore signal a physically healthy speaker. As shown for faces, health signaling is, however, not the only explanation to offer in this respect.

A puzzling aspect of the average effect concerns its relation to individual variation in the perception of beauty. This holds in particular for nonaverage features. For example, how can the average effect explain why extraordinary bright eyes might be found attractive by a considerable part of the population? The current results show that the average effect does not exclude individual preferences. This becomes particularly clear from the number of participants preferring a certain rhythm most often (Table 1, rightmost column). These counts showed that in both languages more than half of the participants preferred a non-average rhythm most often, despite the fact that the average received (one of) the highest ratings throughout. There were only a few rhythms that were never the most often chosen one (two in Dutch, one in Mandarin Chinese), indicating that the rhythm preferences were distributed over the overall majority of the available rhythms. The latter observation is taken as an indication of individual variation among the participants in either language.

Given that individual variation coexists with the average effect, it is difficult to claim that there is an intrinsic attractiveness to averaged features. If averaged features are more attractive than any outlying (individual-specific) features, the average, or anything close to it, would be rated as the most attractive by the majority of the perceivers. This is not observed and hints at the possibility that the average effect does not primarily originate from the intrinsic attractiveness of the average, but rather from the mathematics of calculating the average. This possibility has received some attention in work on facial attractiveness (DeBruine et al., 2007; Rhodes et al., 2003) and deserves a brief final discussion in the following.

It can be demonstrated that the average has the highest probability of being perceived as attractive, as it is—on average—the closest to all individual observations, regardless of the amount of variation in the observations. Support for this reasoning is given in an analogous example (Appendix B). It is thus crucial to note that the mathematical average by nature has a high probability of being selected. Intrinsic attractiveness should therefore be kept separate from the mathematical advantage of central tendency measures. It is left for future work to investigate which rhythmic features might exhibit intrinsic attractiveness in the way that faces were shown to have (e.g., big eyes or a small chin; Baudouin & Tiberghien, 2004). It is plausible that enhancement of specific rhythmical features could make the average rhythm even more attractive to listeners (cf. DeBruine et al., 2007). The question remains whether enhancement of these features has similar effects across languages, or whether these enhancements are the ones in which languages differ substantially. Acoustic manipulation of regular amplitude modulations (cf. Bosker & Ghitza, 2018) could be one method of tackling these issues. Related to these variables is the overall intelligibility of speech that could be expected to co-vary with rhythm (faster rhythms leading to less intelligible speech). The current study did not explicitly focus on intelligibility and this factor is therefore left for future work. It is particularly important to understand the extent to which the effects hypothesized above can be explained by evolutionary mechanisms, as hinted at in the visual literature (e.g., Halberstadt & Rhodes, 2003; Thornhill & Gangestad, 1993) or whether attractiveness in speech rhythm can entirely be accounted for by a mathematical advantage of central tendency measures, as suggested as alternative explanation in the current study.

Acknowledgements

The authors thank Lennard van de Laar, Stijn Paludanus, Chun Wang, and Gyongmin Oh for their effort in setting up and running the experiments. The Dutch results are also reported in a separate study (Swerts & Kaland, submitted).

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Research for this paper was funded by the German Research Foundation (DFG)—Project-ID 281511265—SFB 1252.

Ethical approval

The experiments reported in this paper have been conducted following protocols and informed consent practices in compliance with the Helsinki Declaration, with approval of the Research Ethics and Data Management Committee of the Tilburg School of Humanities and Digital Sciences and the Faculty of Arts and Humanities of the University of Cologne.

Informed consent

Informed consent was obtained from each individual participant prior to participation.

ORCID iDs

Constantijn Kaland https://orcid.org/0000-0002-1813-5902
Marc Swerts https://orcid.org/0000-0002-4367-641X

Note

1. It is beyond the scope of this demonstration to discuss the different ways in which attractiveness could be modeled, for example, whether a normal distribution indeed applies or whether number selection reflects human perception of beauty.

References

Albert, A., Cangemi, F., & Grice, M. (2018). Using periodic energy to enrich acoustic representations of pitch in speech: A demonstration. *Speech Prosody*, *2018*, 804–808. https://doi.org/10.21437/SpeechProsody.2018-162

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica*, *66*(1–2), 46–63. https://doi.org/10.1159/000208930

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, *40*(3), 351–373. https://doi.org/10.1016/j.wocn.2012.02.003

Baudouin, J.-Y., & Tiberghien, G. (2004). Symmetry, averageness, and feature size in the facial attractiveness of women. *Acta Psychologica*, *117*(3), 313–332. https://doi.org/10.1016/j.actpsy.2004.07.002

Beckman, M. E., & Venditti, J. J. (2010). Tone and intonation. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The handbook of phonetic sciences* (1st ed., pp. 603–652). Wiley. https://doi.org/10.1002/9781444317251.ch16

Belin, P. (2021). On voice averaging and attractiveness. In B. Weiss, J. Trouvain, M. Barkat-Defradas, & J. J. Ohala (Eds.), *Voice attractiveness: Prosody, phonology and phonetics* (pp. 139–149). Springer. https://doi.org/10.1007/978-981-15-6627-1_8

Boersma, P., & Weenink, D. (2022). *Praat: Doing phonetics by computer*. http://www.praat.org/

Borrás-Comes, J., Kaland, C., Prieto, P., & Swerts, M. (2014). Audiovisual correlates of interrogativity: A comparative analysis of Catalan and Dutch. *Journal of Nonverbal Behavior*, *38*(1), 53–66. https://doi.org/10.1007/s10919-013-0162-0

Bosker, H. R. (2021). The contribution of amplitude modulations in speech to perceived charisma. In B. Weiss, J. Trouvain, M. Barkat-Defradas, & J. J. Ohala (Eds.), *Voice attractiveness: Prosody, phonology and phonetics* (pp. 165–181). Springer. https://doi.org/10.1007/978-981-15-6627-1_10

Bosker, H. R., & Ghitza, O. (2018). Entrained theta oscillations guide perception of subsequent speech: Behavioural evidence from rate normalisation. *Language, Cognition and Neuroscience*, *33*(8), 955–967. https://doi.org/10.1080/23273798.2018.1439179

Brehm, L., & Alday, P. M. (2022). Contrast coding choices in a decade of mixed models. *Journal of Memory and Language*, *125*, 104334. https://doi.org/10.1016/j.jml.2022.104334

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G. A., Kawahara, H., & Belin, P. (2010). Vocal attractiveness increases by averaging. *Current Biology*, *20*(2), 116–120. https://doi.org/10.1016/j.cub.2009.11.034

Collins, S. A., & Missing, C. (2003). Vocal and visual attractiveness are related in women. *Animal Behaviour*, *65*(5), 997–1004. https://doi.org/10.1006/anbe.2003.2123

Crespo Sendra, V., Kaland, C., Swerts, M., & Prieto, P. (2013). Perceiving incredulity: The role of intonation and facial gestures. *Journal of Pragmatics*, *47*(1), 1–13. https://doi.org/10.1016/j.pragma.2012.08.008

DeBruine, L. M., Jones, B. C., Unger, L., Little, A. C., & Feinberg, D. R. (2007). Dissociating averageness and attractiveness: Attractive faces are not always average. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1420–1430. https://doi.org/10.1037/0096-1523.33.6.1420

Dellwo, V. (2008). *The role of speech rate in perceiving speech rhythm*. ISCA. https://doi.org/10.5167/UZH-111799

Dellwo, V., & Wagner, P. (2003). *Relations between language rhythm and speech rate*. https://doi.org/10.5167/UZH-111779

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, *81*, 181–187. https://doi.org/10.1016/j.neubiorev.2017.02.011

Feinberg, D. R., DeBruine, L. M., Jones, B. C., & Perrett, D. I. (2008). The role of femininity and averageness of voice pitch in aesthetic judgments of women's voices. *Perception*, *37*(4), 615–623. https://doi.org/10.1068/p5514

Galton, F. (1879). composite portraits, made by combining those of many different persons into a single resultant figure. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *8*, 132. https://doi.org/10.2307/2841021

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the Rhythm Class Hypothesis. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology 7* (pp. 515–546). De Gruyter. https://doi.org/10.1515/9783110197105.515

Gussenhoven, C., & Chen, A. (Eds.). (2020). *The Oxford handbook of language prosody* (1st ed.). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198832232.001.0001

Halberstadt, J., & Rhodes, G. (2000). The attractiveness of nonface averages: Implications for an evolutionary explanation of the attractiveness of average faces. *Psychological Science*, *11*(4), 285–289. https://doi.org/10.1111/1467-9280.00257

Halberstadt, J., & Rhodes, G. (2003). It's not just average faces that are attractive: Computer-manipulated averageness makes birds, fish, and automobiles attractive. *Psychonomic Bulletin & Review*, *10*(1), 149–156. https://doi.org/10.3758/BF03196479

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Jacobs, J. (1894). *The fables of Aesop*. Macmillan.

Jun, S.-A. (Ed.). (2014a). *Prosodic typology II: The phonology of intonation and phrasing*. Oxford University Press.

Jun, S.-A. (2014b). Prosodic typology: By prominence type, word prosody, and macro-rhythm. In S.-A. Jun (Ed.), *Prosodic typology II* (pp. 520–539). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199567300.003.0017

Kaland, C. (2022). Bending the string: Intonation contour length as a correlate of macro-rhythm. *Interspeech*, *2022*, 5233–5237. https://doi.org/10.21437/Interspeech.2022-185

Kawahara, H., & Matsui, H. (2003). Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings (ICASSP '03)*. IEEE. https://doi.org/10.1109/ICASSP.2003.1198766

Kenny, D. A. (1987). *Statistics for the social and behavioral sciences*. Little, Brown and Company.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science*, *1*(2), 115–121. https://doi.org/10.1111/j.1467-9280.1990.tb00079.x

Langlois, J. H., Roggman, L. A., & Musselman, L. (1994). What is average and what is not average about attractive faces? *Psychological Science*, *5*(4), 214–220. https://doi.org/10.1111/j.1467-9280.1994.tb00503.x

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, *5*(3), 253–263. https://doi.org/10.1016/S0095-4470(19)31139-8

Lenth, R. V. (2023). *emmeans: Estimated marginal means, aka least-squares means* [Manual]. https://CRAN.R-project.org/package=emmeans

Mok, P., & Dellwo, V. (2008). Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English. In *Proceedings of the 4th International Conference on Speech Prosody, SP 2008* (pp. 423–426). International Speech Communication Association.

Nolan, F., & Jeon, H.-S. (2014). Speech rhythm: A metaphor? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1658), 20130396. https://doi.org/10.1098/rstb.2013.0396

Pike, K. L. (1945). *The intonation of American English*. University of Michigan Press.

Polyanskaya, L., Busá, M. G., & Ordin, M. (2020). Capturing cross-linguistic differences in macro-rhythm: The case of Italian and English. *Language and Speech*, *63*(2), 242–263. https://doi.org/10.1177/0023830919835849

Prechtel, C. (2020). Macro-rhythm in English and Spanish: Evidence from Radio Newscaster Speech. In *Speech Prosody 2020* (pp. 675–679). International Speech Communication Association. https://doi.org/10.21437/SpeechProsody.2020-138

Qualtrics International Inc. (2020). *Qualtrics*. https://www.qualtrics.com/

R Core Team. (2022). *R: The R project for statistical computing*. https://www.r-project.org/

Repp, B. H. (1997). The aesthetic quality of a quantitatively average music performance: Two preliminary experiments. *Music Perception*, *14*(4), 419–444. https://doi.org/10.2307/40285732

Rhodes, G., Jeffery, L., Watson, T. L., Clifford, C. W., & Nakayama, K. (2003). Fitting the mind to the world: Face Adaptation and Attractiveness Aftereffects. *Psychological Science*, *14*(6), 558–566. https://doi.org/10.1046/j.0956-7976.2003.psci1465.x

Rhodes, G., Yoshikawa, S., Clark, A., Lee, K., McKay, R., & Akamatsu, S. (2001). Attractiveness of facial averageness and symmetry in Non-Western cultures: In search of biologically based standards of beauty. *Perception*, *30*(5), 611–625. https://doi.org/10.1068/p3123

R Studio Team. (2022). *RStudio: Integrated development for R*. https://www.rstudio.com/

Searle, S. R., Speed, F. M., & Milliken, G. A. (1980). Population marginal means in the linear model: An alternative to least squares means. *The American Statistician*, *34*(4), 216–221. https://doi.org/10.1080/00031305.1980.10483031

Smith, D., & Wood, C. (1981, October). *The "USI," or universal synthesizer interface*. Presented at the 70th Audio Engineering Society Convention, Paper 1845, New York, NY, United States.

Swerts, M., & Kaland, C. (submitted). *Mean rhythm. Listeners prefer speech with quantitatively average syllable durations* [Preprint]. SSRN. https://doi.org/10.2139/ssrn.4472289

Thornhill, R., & Gangestad, S. W. (1993). Human facial beauty: Averageness, symmetry, and parasite resistance. *Human Nature*, *4*(3), 237–269. https://doi.org/10.1007/BF02692201

ToDI Collective. (2019). *ToDI second edition–Transcription of Dutch intonation* (2nd ed., Release 2.3). http://todi.let.kun.nl/ToDI/home.htm

Valentine, T., Darling, S., & Donnelly, M. (2004). Why are average faces attractive? The effect of view and averageness on the attractiveness of female faces. *Psychonomic Bulletin & Review*, *11*(3), 482–487. https://doi.org/10.3758/BF03196599

Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J., & Lorenzi, C. (2017). A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America*, *142*(4), 1976–1989. https://doi.org/10.1121/1.5006179J2

Weiss, B., Trouvain, J., Barkat-Defradas, M., & Ohala, J. J. (Eds.). (2021). *Voice attractiveness: Studies on sexy, likable, and charismatic speakers*. Springer. https://doi.org/10.1007/978-981-15-6627-1

Appendix A

Table A1. Stimulus Pairs as Used in the Experiments.

Item #	Member A	Member B
1	Rhythm 1	Rhythm 1
2	Rhythm 1	Rhythm 2
3	Rhythm 1	Rhythm 3
4	Rhythm 1	Rhythm 4
5	Rhythm 1	Rhythm 5
6	Rhythm 1	Rhythm 6
7	Rhythm 1	Rhythm 7
8	Rhythm 1	Rhythm avg
9	Rhythm 2	Rhythm 1
10	Rhythm 2	Rhythm 2
11	Rhythm 2	Rhythm 3
12	Rhythm 2	Rhythm 4
13	Rhythm 2	Rhythm 5
14	Rhythm 2	Rhythm 6
15	Rhythm 2	Rhythm 7
16	Rhythm 2	Rhythm avg
17	Rhythm 3	Rhythm 1
18	Rhythm 3	Rhythm 2
19	Rhythm 3	Rhythm 3
20	Rhythm 3	Rhythm 4
21	Rhythm 3	Rhythm 5
22	Rhythm 3	Rhythm 6
23	Rhythm 3	Rhythm 7
24	Rhythm 3	Rhythm avg
25	Rhythm 4	Rhythm 1
26	Rhythm 4	Rhythm 2
27	Rhythm 4	Rhythm 3
28	Rhythm 4	Rhythm 4
29	Rhythm 4	Rhythm 5
30	Rhythm 4	Rhythm 6
31	Rhythm 4	Rhythm 7
32	Rhythm 4	Rhythm avg
33	Rhythm 5	Rhythm 1
34	Rhythm 5	Rhythm 2
35	Rhythm 5	Rhythm 3
36	Rhythm 5	Rhythm 4
37	Rhythm 5	Rhythm 5
38	Rhythm 5	Rhythm 6
39	Rhythm 5	Rhythm 7

(Continued)

Table A1. (Continued)

Item #	Member A	Member B
40	Rhythm 5	Rhythm avg
41	Rhythm 6	Rhythm 1
42	Rhythm 6	Rhythm 2
43	Rhythm 6	Rhythm 3
44	Rhythm 6	Rhythm 4
45	Rhythm 6	Rhythm 5
46	Rhythm 6	Rhythm 6
47	Rhythm 6	Rhythm 7
48	Rhythm 6	Rhythm avg
49	Rhythm 7	Rhythm 1
50	Rhythm 7	Rhythm 2
51	Rhythm 7	Rhythm 3
52	Rhythm 7	Rhythm 4
53	Rhythm 7	Rhythm 5
54	Rhythm 7	Rhythm 6
55	Rhythm 7	Rhythm 7
56	Rhythm 7	Rhythm avg
57	Rhythm avg	Rhythm 1
58	Rhythm avg	Rhythm 2
59	Rhythm avg	Rhythm 3
60	Rhythm avg	Rhythm 4
61	Rhythm avg	Rhythm 5
62	Rhythm avg	Rhythm 6
63	Rhythm avg	Rhythm 7
64	Rhythm avg	Rhythm avg

Note. Results obtained from pairs indicated in shaded rows were not taken into account (identical members). Avg: average.

Appendix B

Demonstration: population mean ≈ number with highest selection probability

Imagine 10 people (a–j) selecting 10 numbers they find most attractive on a scale from 1 to 12 (including decimals). Some might choose higher numbers, some might choose lower numbers, and others might choose numbers that are spread over a large part of the scale. The numbers thus show a large amount of individual variation that could be expected from participants with different individual preferences.[1] Note that in this simplified example, 10 normally distributed values with a randomly chosen mean between 1 and 10 and a randomly chosen standard deviation between 0 and 2 were generated. See Figure B1 for one example generated for each participant using `rnorm(10,sample(1:10),sample(0:2))` in R (R Core Team, 2022). When the probabilities of each possible attractive number are estimated using kernel density estimation (using the `density()` function in R; R Core Team, 2022), the highest probability is obtained for 5.94, which lies close to the mean of 6.18 (Figure B2).

The generation of the random data was repeated 10,000 times. For each repetition the difference between the population mean and the value with the highest probability was calculated, which indicated that the population mean closely approximates the value with the highest probability (i.e., of being found attractive: $M_{diff} < 0.001$).

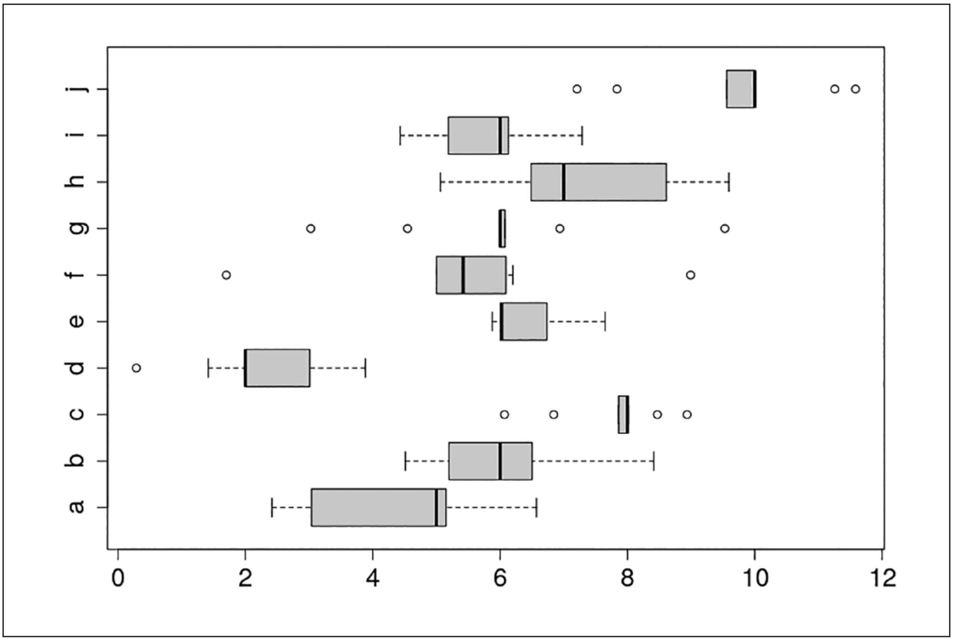

Figure B1. Boxplots of 10 normal distributions (a–j) containing 10 values with a random mean between 1 and 10 and a random standard deviation between 0 and 2. $M = 6.18$, $min = 0.29$, $max = 11.58$.

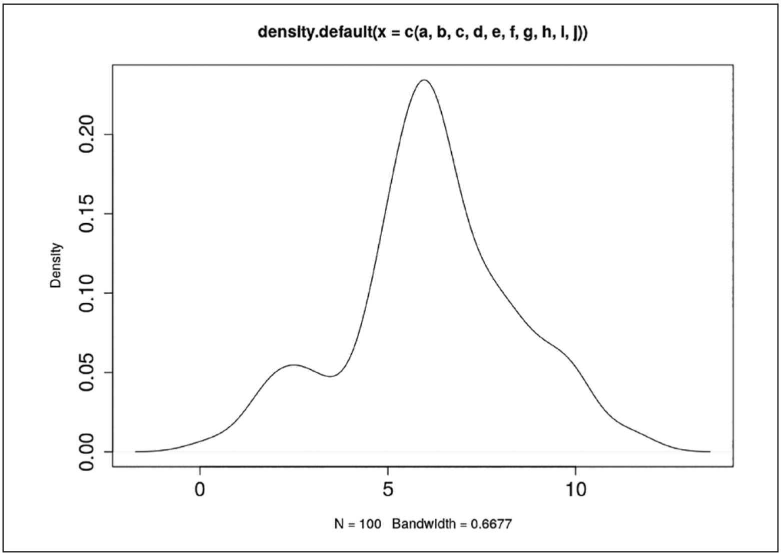

Figure B2. Probability density curve for the observations in Figure B1. $P_{max} = 5.94$.