Comparisons of Harmony and Rhythm of Japanese and English through Signal Processing

by

Aiko Nakano

Submitted to the Department of Mechanical Engineering
in Partial Fulfillment of the Requirement for the Degree of

Bachelor of Science

at the

Massachusetts Institute of Technology

June 2009

Signature of Author .....
Department of Mechanical Engineering
May 8, 2009

Certified by ...
Barbara Hughey, PhD
Instructor
Thesis Supervisor

Accepted by .....
Professor J. Lienhard V
Collins Professor of Mechanical Engineering
Chairman, Undergraduate Thesis Committee

Comparisons of Harmony and Rhythm of Japanese and English
through Signal Processing

by

Aiko Nakano

Submitted to the Department of Mechanical Engineering on May 8, 2009
in partial fulfillment of the requirements for the Degree of Bachelor of Science in
Mechanical Engineering

ABSTRACT

Japanese and English speech structures are different in terms of harmony, rhythm, and frequency of sound. Voice samples of 5 native speakers of English and Japanese were collected and analyzed through fast Fourier transform, autocorrelation, and statistical analysis.

The harmony of language refers to the spatial frequency content of speech and is analyzed through two different measures, Harmonics-to-Noise-Ratio (HNR) developed by Boersma (1993) and a new parameter "harmonicity" which evaluates the consistency of the frequency content of a speech sample. Higher HNR values and lower harmonicity values mean that the speech is more harmonious. The HNR values are 9.6±0.6Hz and 8.9±0.4Hz and harmonicities are 27±13Hz and 41±26Hz, for Japanese and English, respectively; therefore, both parameters show that Japanese speech is more harmonious than English. A profound conclusion can be drawn from the harmonicity analysis that Japanese is a pitch-type language in which the exact pitch or tone of the voice is a critical parameter of speech, whereas in English the exact pitch is less important.

The rhythm of the language is measured by "rhythmicity", which relates to the periodic structure of speech in time and identifies the overall periodicity in continuous speech. Lower rhythmicity values indicate that the speech for one language is more rhythmic than another. The rhythmicities are 0.84±0.02 and 1.35±0.02 for Japanese and English respectively, indicating that Japanese is more rhythmic than English.

An additional parameter, the 80^{th} percentile frequency, was also determined from the data to be 1407±242 and 2021±642Hz for the two languages. They are comparable to the known values from previous research.

Thesis Supervisor: Barbara Hughey, PhD

Title: Instructor of Mechanical Engineering

Table of Contents

ACKNOWLEDGEMENT

I would like to thank my thesis advisor Barbara Hughey for helping me develop the theory for rhythmicity and also for participating in data analysis. I am grateful for her continued support since the initial research, which evolved as a 2.671 class project. I would also like to thank Ms. Miwako Hisagi of the Speech Communications Group at the Research Laboratory of Electronics at MIT for meeting with me to discuss possible improvements on the initial draft. Many thanks go to Brian Ross, Edward Cho, Hideo Nakano, Jennifer Moore, Jewel Sharpe, Koji Umeda, Noriko Nakano, Rebecca Smith, and Yuta Kuboyama for willingly agreeing to participate in this experiment. Without their participation, this experiment would not have been possible.

Finally, I would like to thank my family and friends for their continued support through this thesis and throughout my career at MIT.

BIOGRAPHICAL NOTE

Aiko Nakano is a senior at the Massachusetts Institute of Technology. She is receiving a B.S. in Mechanical Engineering with minors in Economics and Management in June 2009.

The motivations for this research were her background as a bilingual in Japanese and English and her interests in languages. She is originally from Japan and has spent five years in Seattle before high school, where she learned English. This paper originated as a class project for 2.671 Measurement and Instrumentation, which was taught by her thesis advisor.

1. Introduction

With fluencies in both Japanese and English, the author's personal experience indicates that the pitch of voice and rhythm of the two languages are significantly different. Her voice is higher in tone when speaking in English than in Japanese, and the ways that she stresses syllables are different. Linguistics studies show that in Japanese, stresses are given by lowering or raising the tone of voice. In English, stresses are given by changing the volume of speech.[1]

The present study compares Japanese and English speech by analyzing frequency spectra of 5 phrases spoken by 5 native speakers of each of the two languages. Each voice sample recording is analyzed through fast Fourier transform, correlation function techniques, and simple statistical methods. We use Praat, a signal processing program developed by Boersma, to find the harmonics-to-noise-ratio of the signal. New parameters, "harmonicity" and "rhythmicity", are introduced, which measure how musical or rhythmic the language is. Harmonicity is defined as the standard deviation of the spacing between frequency peaks in the Fourier spectrum of the voice sample. Rhythmicity is defined from the frequency spectrum of the autocorrelation function as the width encompassing 20% - 80% of the total power divided by the central frequency. Additionally, the "frequency" for each language is defined as the frequency that includes 80% of the total power in the Fourier spectrum of the voice sample.

Previous research has been performed on rhythmicity[17] and harmonicity[18, 19, 20] in different contexts such as for musical instruments or measuring the hoarseness of voice. However, the methods developed by Boersma or Kohler and Yumoto et al. have not been widely used in the context of comparing the rhythm and harmony of languages. The concepts of harmonicity and rhythmicity as used in the present work were independently developed by the author and B. Hughey and differ from Boersma's or Kohler and Yumoto's definitions.

We find that the results of this experiment agree with the linguistics research that Japanese is a pitch language and English is not. Additionally, the frequencies of the languages can be compared with the accepted values: Japanese has a frequency of 400-1200 Hz and English has a frequency of 1500-5000 Hz.[5] Further research with a variety of languages would be helpful in demonstrating the general usefulness of the two new parameters, harmonicity and rhythmicity.

2. Background

2.1.1 Frequency Analysis of Spoken Language

Any time-varying signal can be mathematically expressed as superposition of sine waves, and a particular tone can be described by a combination of a fundamental frequency and its harmonic (integer) multiples. Figure 1 shows the frequency of an A1 note played on violin in the time and frequency domain. On the top chart, we can observe the repeating pattern that looks like a mountain with three bumps. Each of these bumps corresponds to the three peaks in the bottom chart: 900, 1800, and 2700 Hz. 900Hz is the fundamental frequency of the A1 note, and the higher two frequencies are its harmonics.

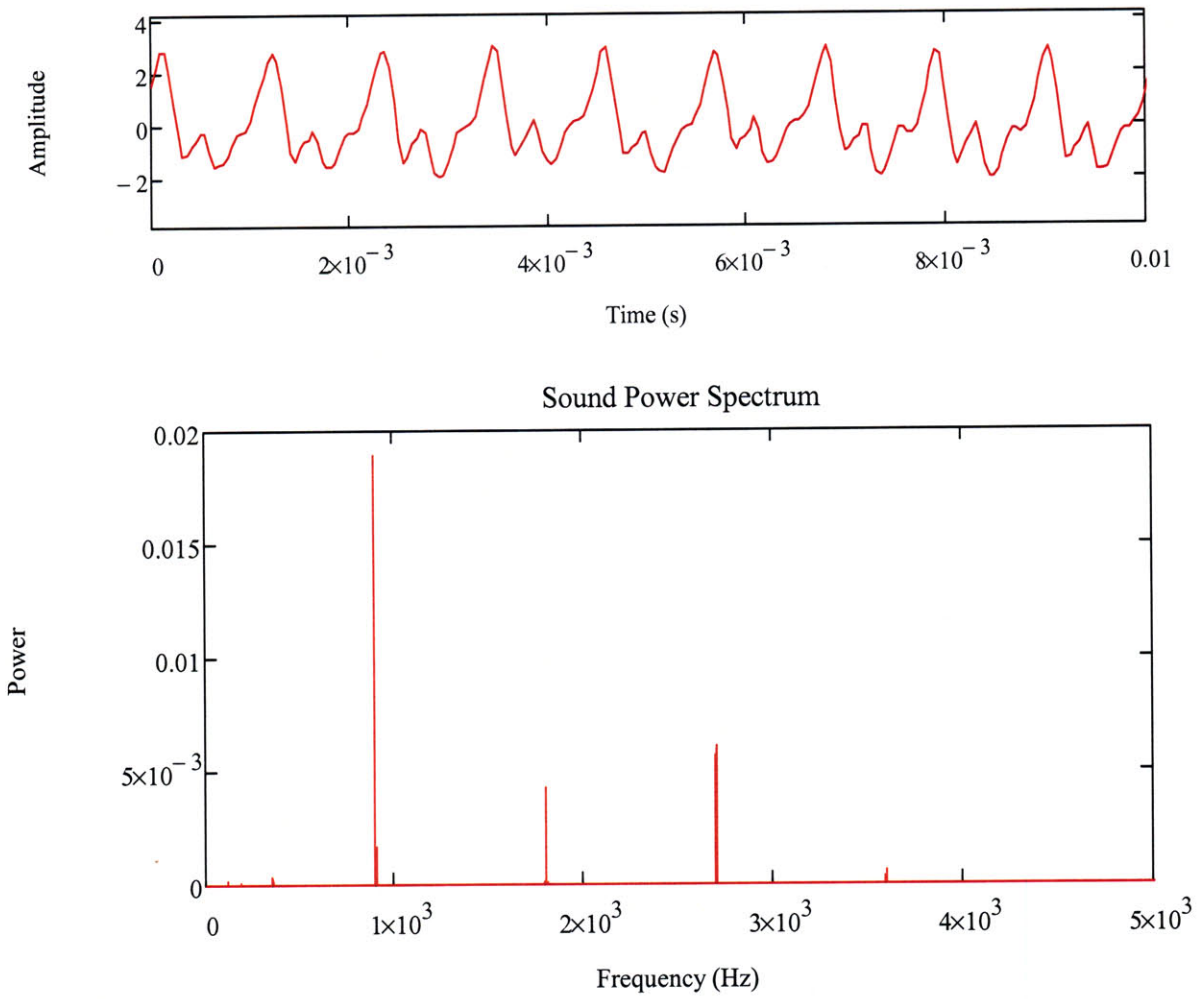

Figure 1 The A1 note played on a violin in the time (top) and frequency (bottom) domain.

Similarly, speech can be expressed as superposition of sine waves. The frequency range for vowels is typically 300-1000Hz, whereas that for consonants is higher at 1000-3000Hz.[4] The vowels have solid tones typically expressed by multiples of fundamental frequencies, often related in frequency by simple integer ratios. In Figure 2, the frequency spectra of the five vowel sounds are shown as measured in the present study. The number of peaks varies for different sounds, but they all have peaks at consistent intervals, resulting in a clear tone. The sound power spectrum much resembles that of violin shown in Figure 1.

In contrast, consonants have higher and broader frequency components made of harmonics of many unrelated fundamental frequencies. For example, sounds such as "sh" consist of many non-harmonic components, giving a non-systematic frequency pattern.[3] This sound is shown in the Frequency domain in Figure 3. The less distinct peaks and unequal spacing between them suggest existence of more than one fundamental frequency.

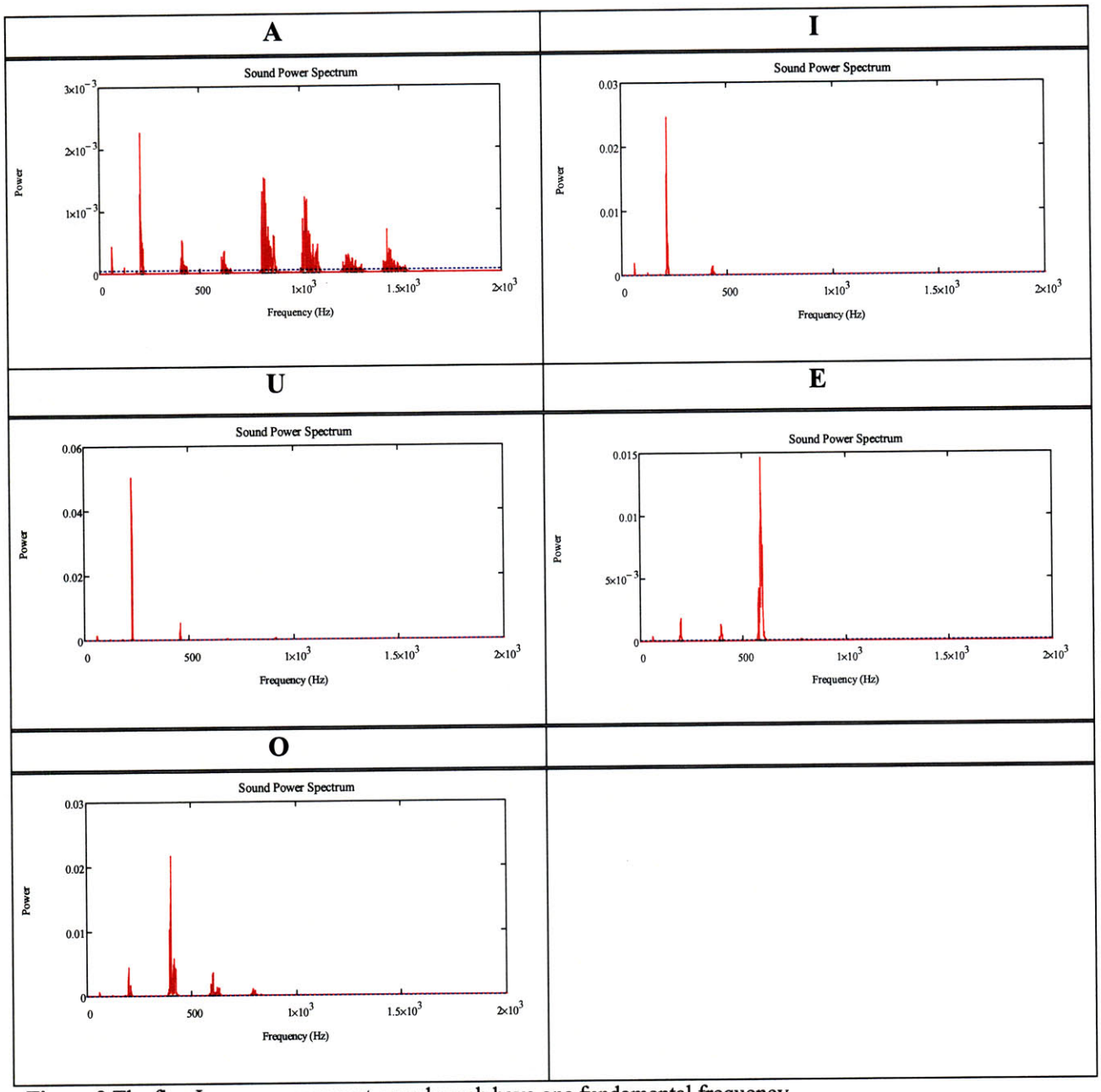

Figure 2 The five Japanese consonant sounds each have one fundamental frequency.

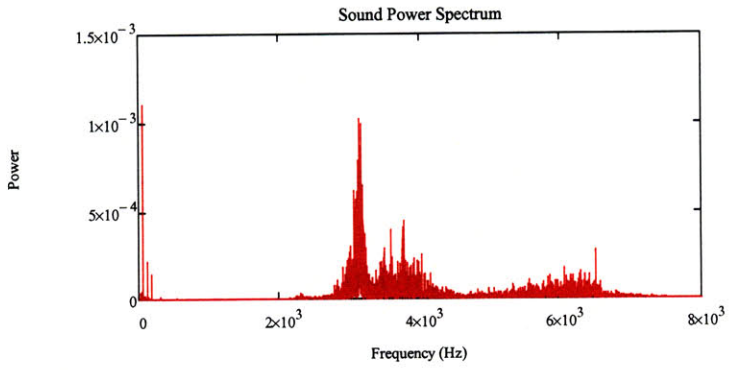

Figure 3 The sound power spectrum of a consonant sound "sh" contains more than one fundamental frequency sound as shown by the noisy spectra and unequal peak spacing.

2.1.2 Frequency Attributes of Japanese

The frequency of Japanese speech is typically 400-1200Hz. It is a low frequency language, because all Japanese sounds end with a vowel sound due to its one-to-one vowel and consonant structure. Since vowels have specific frequencies associated with them, as shown in Figure 2, we would expect that the frequency spectra of Japanese speech have distinct peaks at consistent intervals.

In linguistics term, Japanese is called a pitch-type language because the syllables are stressed by changing the pitch, or the frequency of the sound. [1]

2.1.2 Frequency Attributes of English

On the other hand, English sound structure is not very consistent when compared to Japanese and has higher frequency of 1500-5000Hz. American English has fifteen different vowel sounds, and at times the vowel sound can be reduced, such as the "a" in "about". Furthermore, the frequency structure is made more complex by the silent consonant sounds such as "sh", as shown in Figure 3. Such English consonants have frequencies as high as 8,000 or 10,000Hz.[5]

We call English a stress-type language because syllables are stressed by changing the relative prominence/loudness during pronunciation.[2]

Table 1 summarizes the characteristics of Japanese and English discussed above. The results from this experiment are compared with the data in the table.

Table 1 Key Properties of Japanese and English

	Japanese	**English**
Type	Pitch	Stress
Stress on syllables by	Tone of voice (frequency of sound)	Volume of voice (amplitude of sound)
Structure	Each sound is vowel + consonant	Many vowel sounds, silent syllables, and many consonants
Frequency	400 – 1200 Hz	1500 – 5000 Hz

2.2 Autocorrelation: Finding the Repeating Patterns in Time

Autocorrelation is the cross-correlation of a signal with itself. For a time signal x(t) that is stationary (statistics are constant), the autocorrelation $r_x(\tau)$ shifts one wave form by a time-lag τ defined as

$$r_x(\tau) = \int x(t)x(t+\tau)dt \tag{1}$$

and finds a repeating pattern within itself. It can be used to find a periodic signal which is buried under noise or to identify the missing fundamental frequency in a signal implied by its harmonic frequencies. In Figure 4, autocorrelation reveals a sinusoidal pattern within an otherwise rather

noisy signal. A noisy and less rhythmic sound appears as a vertical line at zero lag. The FFT of the original sound file and the Fast Fourier Transform of the autocorrelation function result in the same peak positions, as shown in Figure 5, but autocorrelation picks out repeating frequencies. Therefore, we can expect that the fundamental frequencies of the vowel sounds will be extracted out and the "noisy" consonant sounds such as "sh" is eliminated by the autocorrelation function.

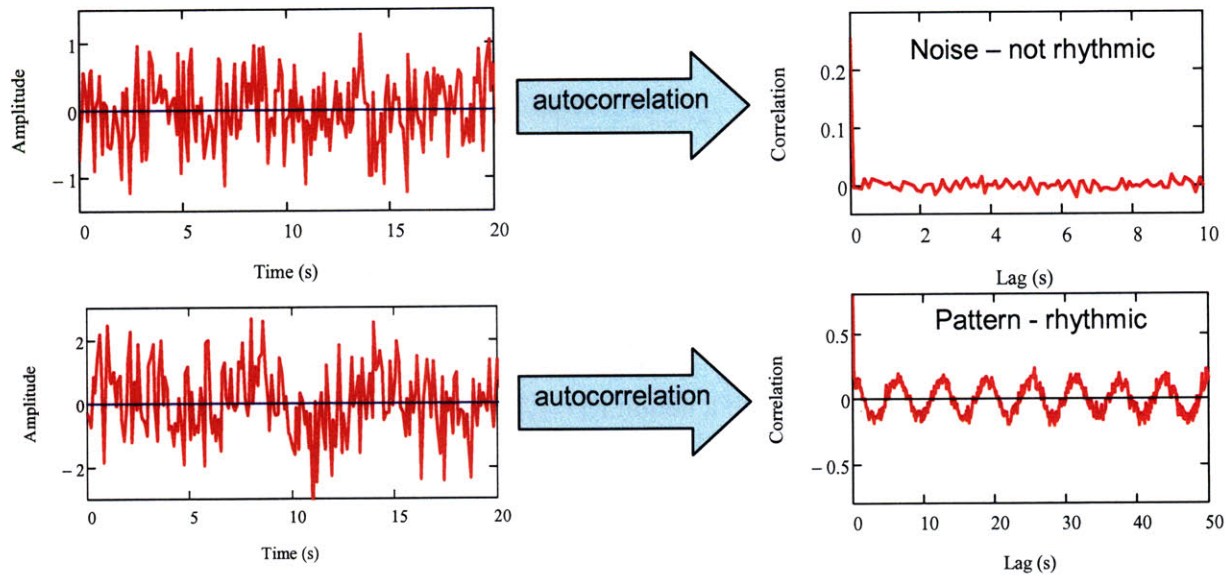

Figure 4 Autocorrelation removes the noise by collecting the non-repeating frequencies near time 0.

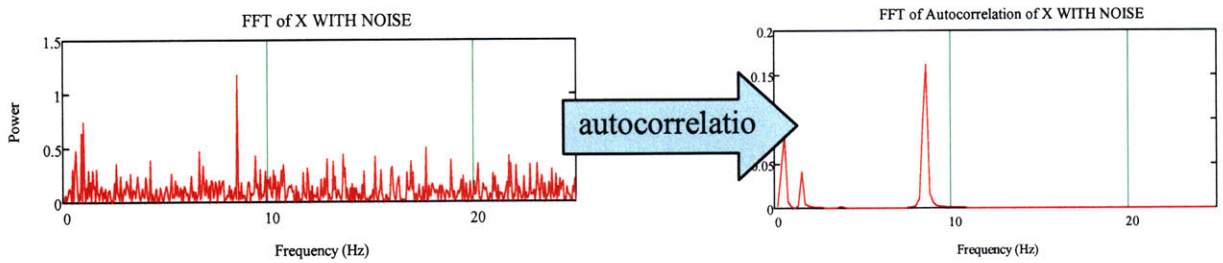

Figure 5 Autocorrelation on the FFT extracts the repeating pattern.

2.3 Harmony: Tonal Frequency Content of Spoken Language in Space

2.3.1 Harmonics-to-Noise Ratio (HNR)

The Harmonics-to-Noise ratio (HNR) was developed by Pal Boersma at the Institute of Phonetic Sciences at University of Amsterdam and can be calculated using Praat[21], a speech analysis tool for phonetics, developed by him and his colleague D. Weenink. We will use his methods as one of the ways to analyze the harmony of the speech. The following paragraphs are taken from Boersma's paper, "Accurate Short-term analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound" (1993).[20]

[As shown in Figure 4 for the non-rhythmic sound, the autocorrelation function (1)] has a global maximum for τ=0. If there are also global maxima outside 0, the signal is called periodic and there exists a lag T_0, called the period, so that all these maxima are placed at the lags nT_0, for every integer n. The fundamental frequency f_0 of this periodic signal is defined as $f_0=1/T_0$. If there are no global maxima outside 0, there can still be local maxima. If the highest of these is at a lag τ_{max}, and if its height $r_x(\tau_{max})$ is large enough, the signal is said to have a periodic part, and its harmonic strength R_0 is a number between 0 and 1, equal to the local maximum $r_x'(\tau_{max})$ of the normalized autocorrelation.

$$r'_x(\tau) \equiv \left(\frac{r_x(\tau)}{r_x(0)} \right) \tag{2}$$

We could make such a signal x(t) by taking a periodic signal H(t) with a period T_0 and adding a noise N(t) to it. We can infer from equation (1) that if these two parts are uncorrelated, the autocorrelation of the total signal equals the sum of the autocorrelations of its parts. For zero lag, we have $r_x(0) = r_H(0) + r_N(0)$, and if the noise is white (i.e., if it does not correlate with itself), we find a local maximum at a lag $\tau_{max} = T_0$ with a height $r_x(\tau_{max}) = r_H(T_0) = r_H(0)$. Because the autocorrelation of a signal at zero lag equals the power in the signal, the normalized autocorrelation at τ_{max} represents the relative power of the periodic (or harmonic) component of the signal, and its complement represents the relative power of the noise component:

$$r'_x(\tau_{max}) \equiv \left(\frac{r_H(0)}{r_x(0)} \right); \quad 1 - r'_x(\tau_{max}) \equiv \left(\frac{r_N(0)}{r_x(0)} \right) \tag{3}$$

This allows us to define the logarithmic harmonics-to-noise ratio (HNR) as

$$HNR = 10 \times \log \left(\frac{r'_x(\tau_{max})}{1 - r'_x(\tau_{max})} \right) \tag{4}$$

HNR measures "the degree of acoustic periodicity" by comparing the relative magnitudes of the harmonics and the noise. It is expressed in dB; if 99% of the energy of the signal is in the periodic part, and 1% is noise, the HNR is 10*log10(99/1) = 20 dB. A HNR of 0 dB means that there is equal energy in the harmonics and in the noise. Hence, the higher the HNR ratio, the more harmonic the sound is.

2.3.2 Harmonicity: A New Parameter of Harmony

This paper proposes another method of measuring the harmony of the language, by measuring the consistency of spacing between frequency peaks. We will call this harmonicity H, and define it as

$$H = \sigma(\Delta f_i) \quad \forall i = 1,2,...n \tag{5}$$

where σ is the standard deviation and Δf is the peak frequency spacing in the Fourier spectrum of a recorded sound, as shown in Figure 6. In this case, H is close to a value of zero,

because the frequency peaks are equally spaced and thus there is no variation in the spacing. Low value of harmonicity means that there are a small number of fundamental frequencies and therefore the speech can be called "harmonious" in terms of its frequency content. If harmonicity for 5 phrases for one language is consistently lower than the other, it can be said that that language is more harmonious than the other. The harmonicity is a more sensitive measure of the harmony than HNR, because it considers more than one periodic signal.

The parameters and their definitions are summarized in Appendix A1.

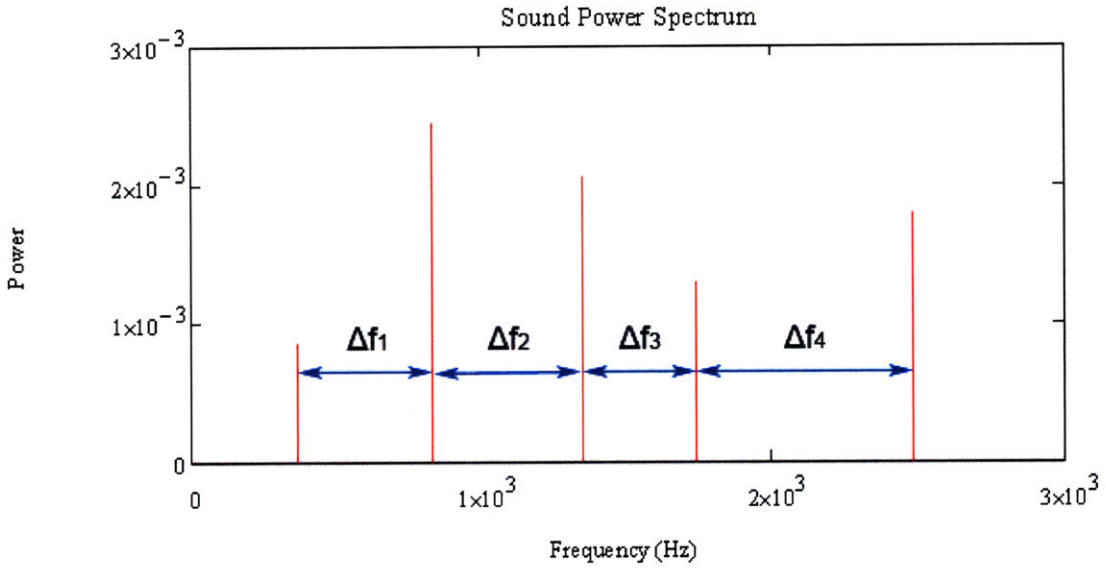

Figure 6 Harmonicity H is the standard deviation of the frequency spacing, Δf_1 through Δf_5.

2.4 Rhythmicity: Periodicity of Language in Time

The main rhythm of spoken language is determined by applying the Fourier transform to the autocorrelation function. For example, if there is only one peak at non-zero frequency, the speech has one repeating beat. Conversely, multiple frequency peaks indicate a rhythmic language with more complicated structure. The figure of merit for rhythmicity is defined in this work as

$$Rh = \frac{\delta}{f_0} \tag{6}$$

where δ is the difference in frequencies between the 20th and 80th percentile integrated power in the Fourier transform of the autocorrelation function. It is normalized by f_0, which is the 50th percentile of the integrated power. The steps to obtain the figure of merit are schematically shown in Figure 7. A lower value of Rh means that the speech is more rhythmic.

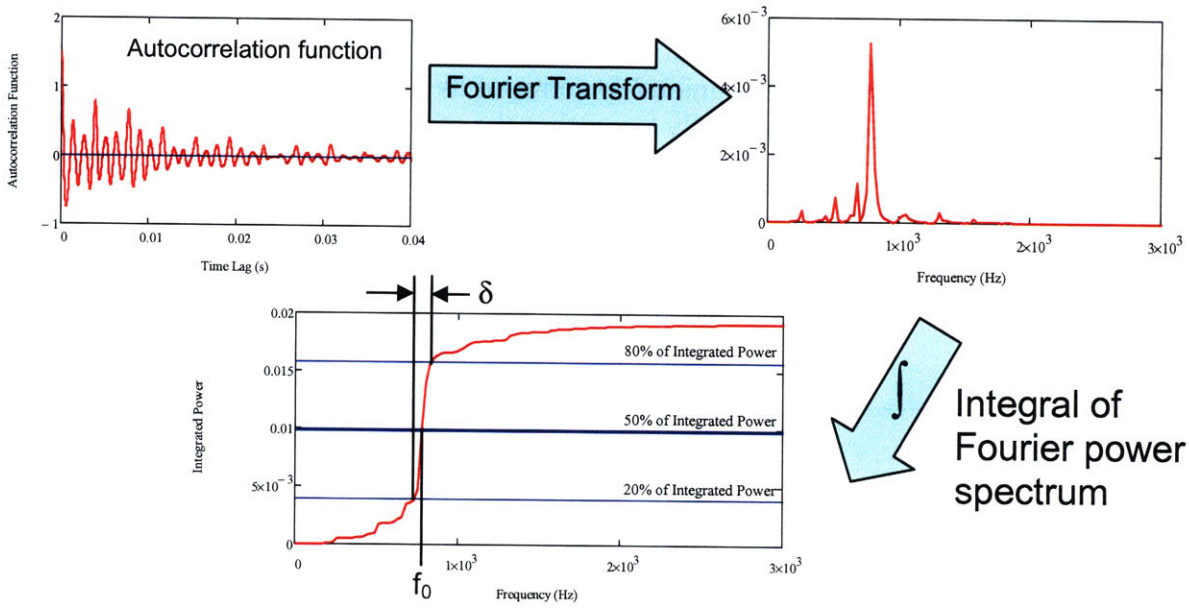

Figure 7 The rhythmicity parameter of a speech sample is calculated from the integrated frequency spectrum of autocorrelation function. Shown above are the steps used to calculate rhythmicity for Japanese speaker 1, phrase 1, trial 1.

3. Experimental Procedure

Voice samples of five native speakers of Japanese and English were collected for five phrases, which are listed in the Appendix 3. Each phrase was recorded 5 times. English speakers were asked to say only the English phrases, and Japanese speakers were asked to speak only the Japanese phrases, to control for the frequency differences intrinsic to the languages. One may expect that a female subject would have a higher tone of voice than a male subject, and a younger person would have a higher pitch than an older person. These differences do not affect our results; however, because all of our parameters are normalized. The 80th percentile frequency data is not normalized, but a general comparison of the frequencies sufficed for the purpose of this paper. Appendix A2 summarizes the characteristics of each of the eight speakers.

The collected voice samples were analyzed using Praat[21], autocorrelation, and Fourier transform techniques as described above to determine the HNR, harmonicity, rhythmicity, and the 80th percentile frequency of each sample.

3.1 Recording Voice Samples

The voice samples were recorded using a microphone (audio-technica ATR25 Stereo Microphone) and Sound Recorder application on a Windows operating system. All samples were recorded at the best quality available (PCM 48.000 kHz, 16 Bit, Stereo).

Five phrases were randomly selected for each language, famous sayings that vary in length and intonation. Each person said each phrase 5 times, resulting in a total of 25 measurements. Each phrase was recorded as one continuous recording for consistency, and then segmented into separate trials

3.2 Analyzing Voice Samples

3.2.1 HNR

Praat[21] was used to calculate the HNR values. After the voice sample is read by the program, we used Harmonicity (cc) function under Periodicity, where (cc) refers to cross-correlation. Since we want to find a repeating pattern within one recording, cross-correlation in this case is equivalent to auto-correlation. The recommended parameters of time step 0.01 seconds, maximum pitch 75 Hz, silence threshold 0.1, and periods per window 1.0 were used. Finally, we obtained the mean HNR value.

3.2.2 Harmonicity

The Sound Analysis[5] code written in MathCAD by Prof. I. W. Hunter was used to apply the Fourier transform to each voice recording. An example frequency spectrum is shown in Figure 8. The frequency of each peak was manually determined, and the value of each interval between adjacent peaks was then calculated in order to determine harmonicity, as described in Section 2.3.2. The selections of peaks are based on the two criteria:

(1) the peak was at least 10% of the maximum power

(2) if the peak was less than 10% of the maximum power, the peak was still selected if it was far enough away from and had power much greater than its neighbors

For example, the power of the fifth peak in Figure 8 is less than 10% of the power of the third (highest) peak. However, it satisfies criteria (2), so it was considered a distinct peak.

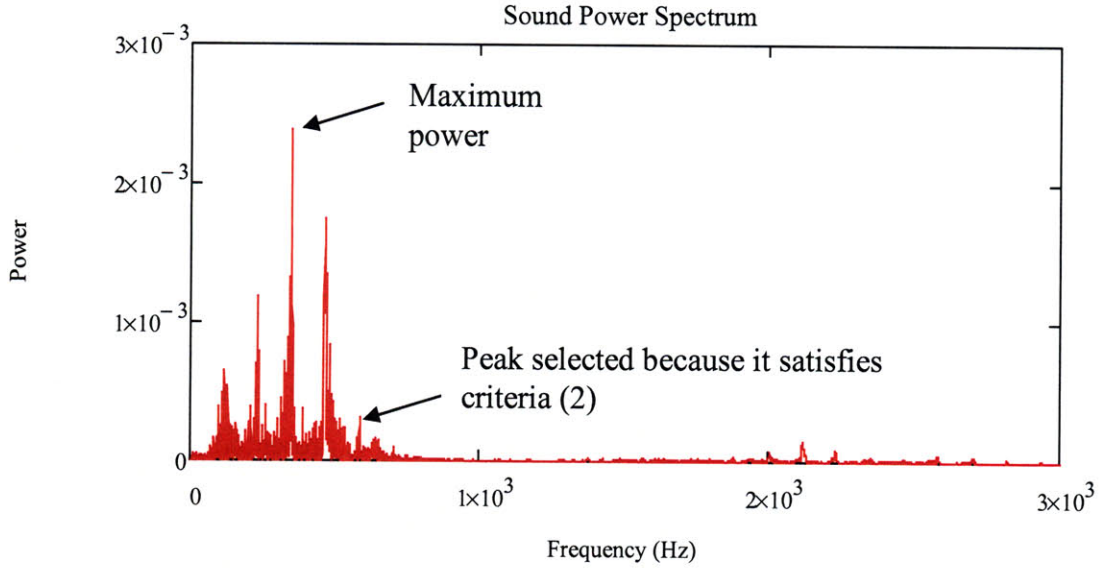

Figure 8 Fast Fourier Transform plot for English Speaker 1, phrase 1, trial 2.

3.2.3 Rhythmicity

Similarly, the rhythmicity was found by manually selecting the 20th, 50th and 80th percentile frequencies of the integrated autocorrelation FFT as described in Section 2.4. An AutoCorrelation[10] code written by Prof. I. W. Hunter was used for this analysis.

3.2.4 80-Percentile Frequency

Lastly, the 80th percentile frequency was found by taking the numerical integral of the frequency spectra and selecting the frequency that intersected the horizontal 80% line (Figure 9). Sound Analysis[5] code was again used for this measurement.

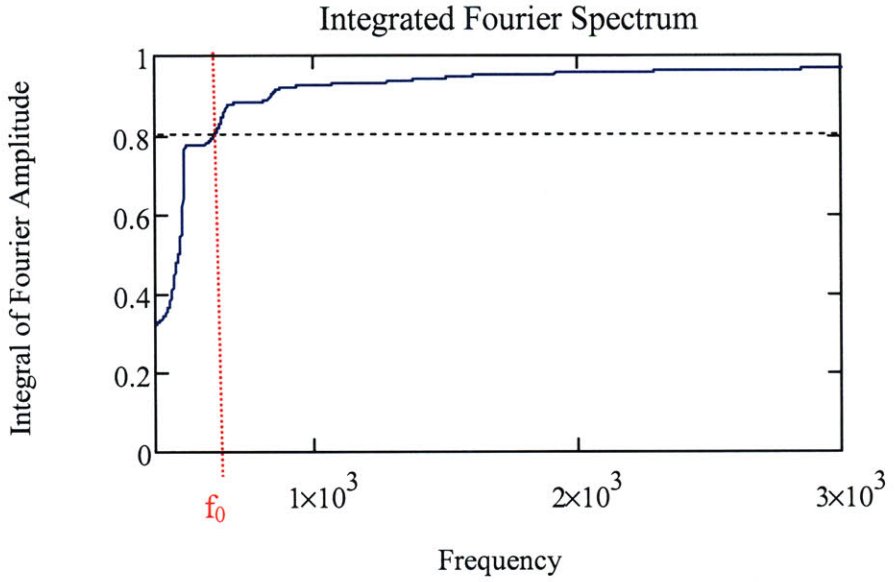

Figure 9 The 80th percentile frequency tells the frequency of the language.

4. Results and Discussion

We find that the HNR value is higher, and harmonicity and rhythmicity values are lower for Japanese than English. Hence, Japanese is more harmonious and rhythmic than English. In addition, we confirm that the frequency of the language is lower for Japanese than for English.

The qualitative analysis of harmonicity and rhythmicity is presented first, followed by the quantitative results. Qualitative analysis reveals interesting traits of each language, such as the word structure and the type of language. For the quantitative analysis, the voice samples are analyzed on the micro- and macro- scales: each trial for each word for one person and the comparison of average harmonicities for the two languages. Lastly, the 80th percentile frequency measurements are presented.

4.1 Harmony

4.1.1 Qualitative Analysis for Each Trial

Qualitative examination of the sound power spectrum can give much insight about each language. Figures 10 and 11 are representative sound power spectra of Japanese and English; other Japanese and English frequency spectra exhibit similar patterns. Figure 10 shows that Japanese has three distinct peaks below 1000 Hz, and Figure 11 shows that English has at least 6 peaks. In fact, 11 peaks are analyzed for Figure 11, since each of these peaks has power of at least 10% of the maximum power. A quick comparison reveals that not only does English have smaller spacing intervals, but these intervals are not as systematic as that of Japanese. On the other hand, Japanese has consistent intervals and thus is harmonious.

The frequency spectra also show the stronger presence of frequencies above 2000 Hz for English than for Japanese. This supports the known finding that English speech consists of many consonant sounds, since consonants have much higher frequency than vowels.

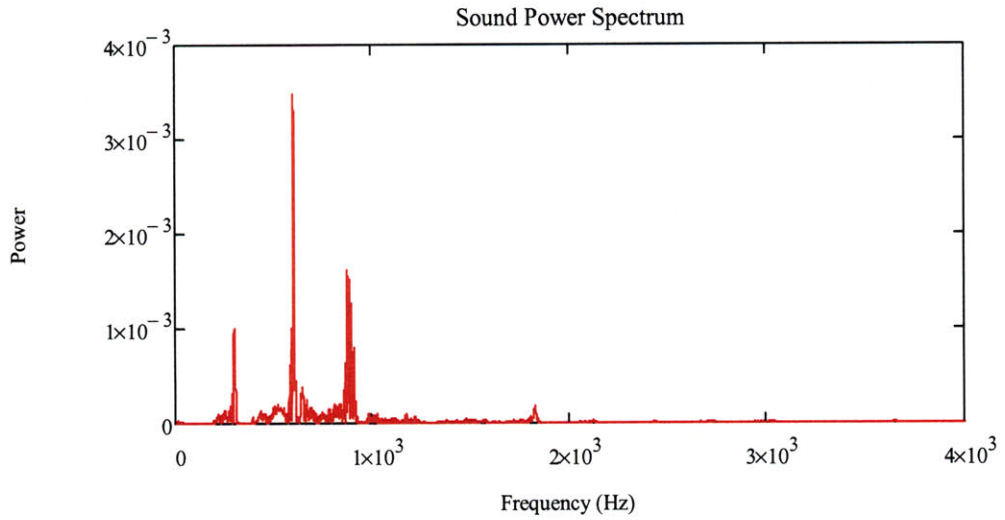

Figure 10 A sound power spectra for Japanese speaker 1, phrase 1, trial 1. The peaks have similar frequency intervals, and so Japanese is harmonious.

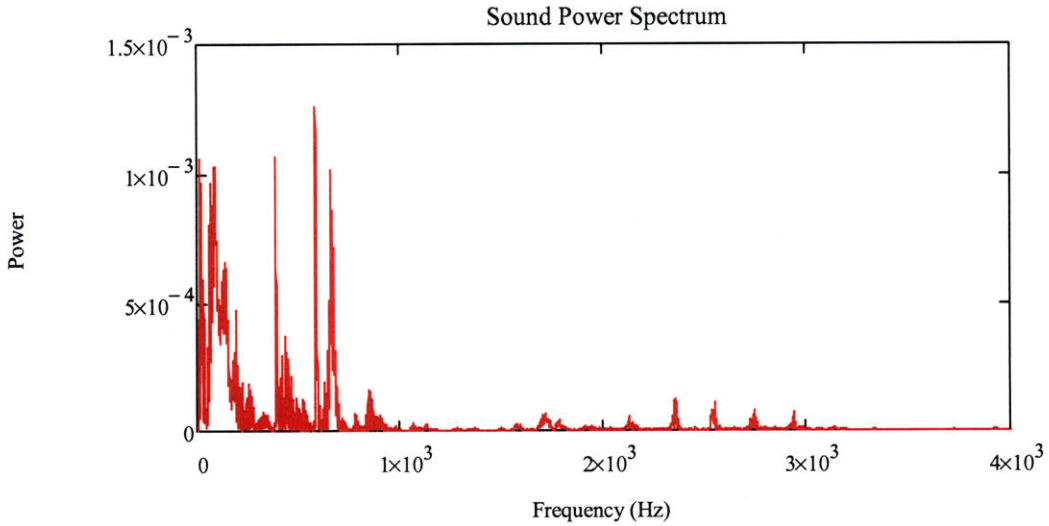

Figure 11 A Sound power spectrums for English speaker 1, phrase 1, trial 4. The frequency spacing is not consistent, so it is not harmonious.

4.1.2 Calculation of Harmonics-to-Noise Ratio (HNR)

The mean HNR values are 9.6±0.6 dB for Japanese and 8.9±0.4dB for English. The HNR values for each of the five phrases averaged over the five speakers are shown in Figure 12. The two values are within the 95% confidence intervals of each other, but Student's t-test shows that we can be 76% confident that Japanese is more harmonious than English.

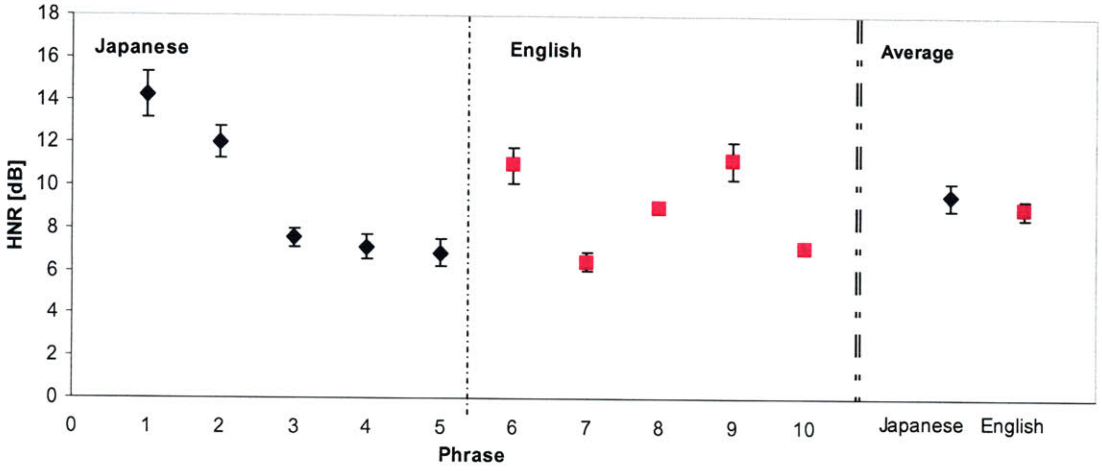

Figure 12 The HNR for the two languages. T-test analysis gives a 76% chance that Japanese HNR is higher than that of English.

Figure 12 shows the results for all five phrases for a single Japanese speaker and a different English speaker. The five different Japanese phrases are shown as trials 1-5, and the English phrases are shown as trials 6-10. All error bars display the 95% confidence interval calculated using standard statistical techniques for small numbers of samples.[9] The average values of the 5 data points for Japanese and English on the left are plotted to the right of the double dashed line. Their 95% confidence intervals are computed using all values used to calculate the averages, so they are not merely the average 95% confidence interval values of the five data points. The same chart format is used throughout the Results section.

4.1.3 Harmonicity

4.1.3.1 Building Blocks for Harmonicity: 1 Phrase for 1 Speaker

The analysis of the harmonicity of a particular language requires averaging over multiple speakers and multiple phrases, but useful qualitative information can be gained by comparing results for a single phrase and a single speaker. Only in this section, we present the results as "average frequency spacing ± harmonicity" to emphasize the large variance of the English speakers.

The first phrase for one Japanese speaker results in average frequency spacing of 273±10 Hz, and that for the English speaker is 168±31 Hz. The lower average value for English shows that the English phrase has more prominent peaks close to each other. More importantly, the large variance for English shows that the frequencies are not equally spaced and that there is more than one fundamental frequency, as expected from the complex English speech structure. On the other

hand, the 95% confidence interval for Japanese is roughly a third of that of English. This means that the variance of the spacing is small, so Japanese is more harmonious than English.

Figure 13 demonstrates that Japanese has consistent frequency spacing and small variance for each trial. This implies that Japanese speech can be controlled purely by changing the frequency or the pitch. From this we can infer that Japanese is a pitch language and new learners of Japanese can acquire a native-like speech by practicing to say each syllable consistently. On the other hand, English frequency intervals vary significantly for the same phrase, even though each trial sounded consistent to our ears. Hence, English uses different pitches (or frequencies) to say the same phrases and is not a pitch language.

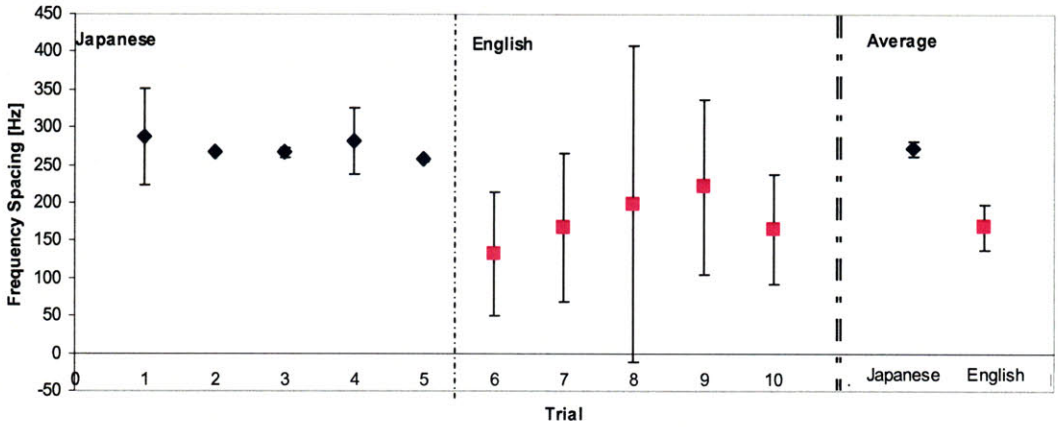

Figure 13 The averages and 95% confidence intervals of frequency spacing for 5 trials of one phrase for one Japanese and one English speaker. The average of the 5 trials is the average frequency spacing for the phrase for one speaker.

4.1.3.2 Harmonicity of the Language

The harmonicity of each language is found by averaging the harmonicities of each phrase for each speaker. The average of this value is the average harmonicity, shown next to the double dashed lines. Japanese has harmonicity of 27±13 Hz, and English has harmonicity of 41±26 Hz, as shown in Figure 14. We are 90% confident that Japanese is more harmonious than English from performing Student's t-test.

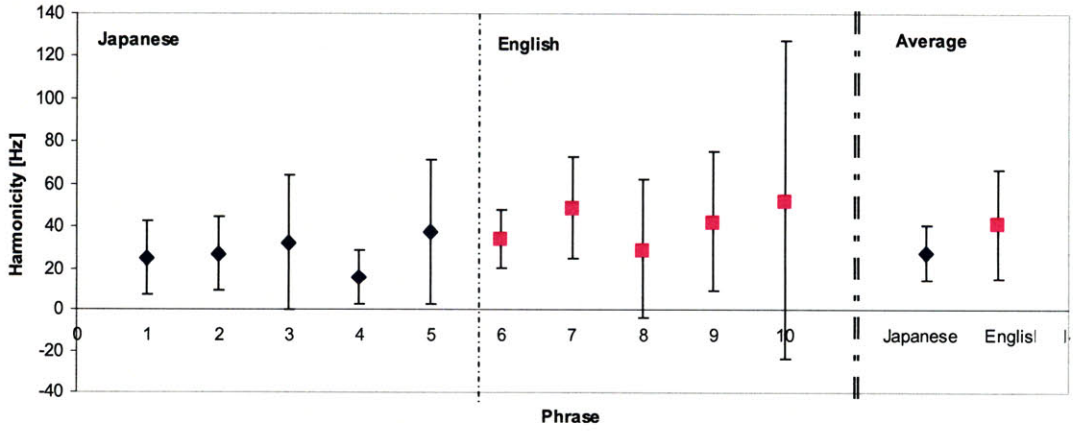

Figure 14 The harmonicity for each phrase averaged over all speakers.

4.2 Rhythmicity

4.2.1 Qualitative Analysis for Each Trial

Before the rhythmicity figure of merit is presented, it is useful to examine the frequency spectra of the autocorrelation functions for each language as shown in Figure 15. Japanese has only one main frequency peak, meaning that the language has one beat. On the other hand, English has several prominent peaks. It has a mixture of beats in its rhythm, and so it is not as rhythmic as Japanese. This information is imbedded in the figure of merit, since more peaks lead to a broader δ as shown in Figure 15, and thus higher value of Rh.

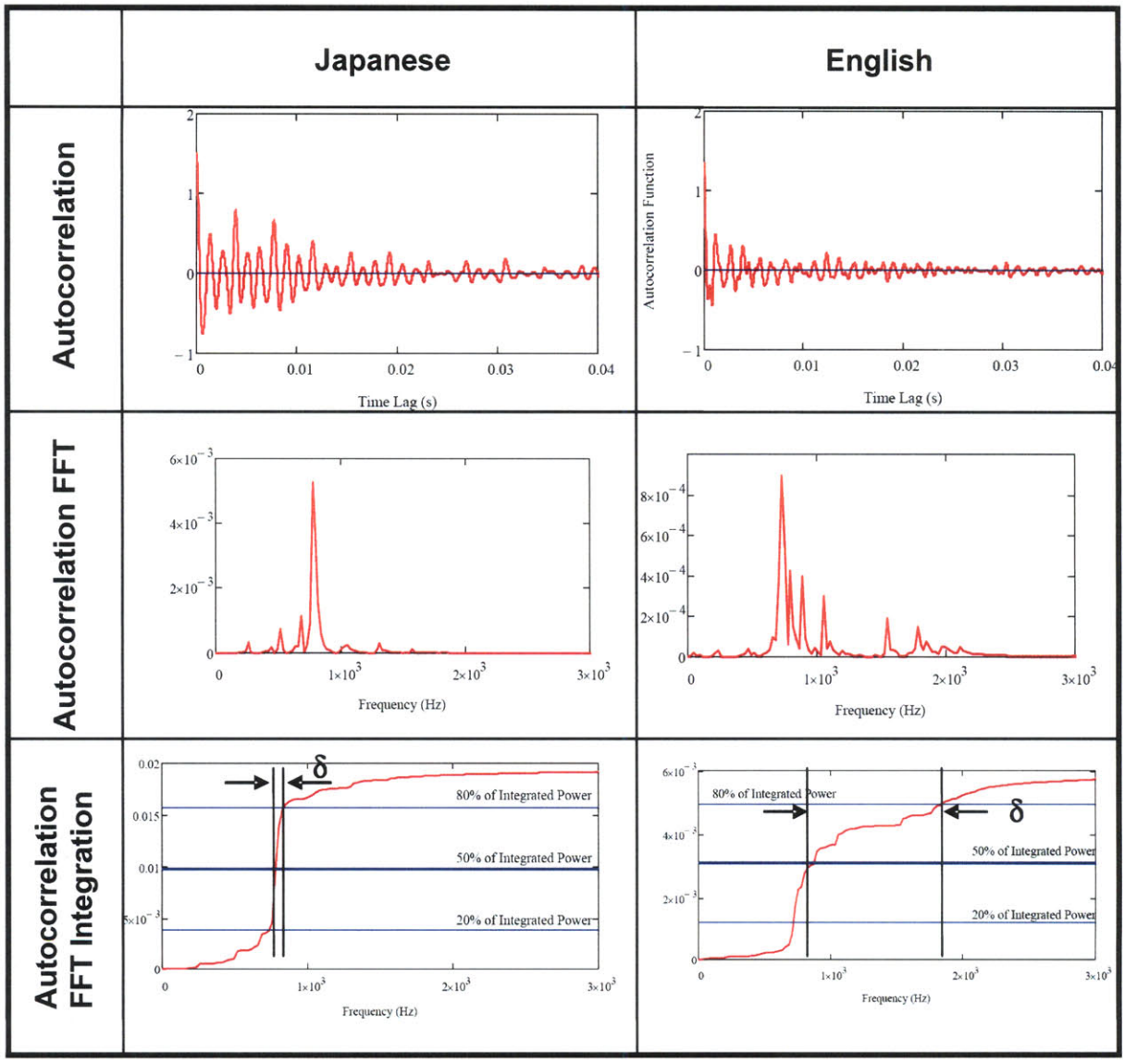

Figure 15 A typical Japanese and English autocorrelation function, its Fourier spectrum, and the integrated Fourier spectrum. As described in the text, English has more peaks because of inconsistent rhythm.

4.2.2 Building Blocks for Rhythmicity: 1 Phrase for 1 Speaker

Again, the analysis of rhythmicity of a particular language requires averaging over multiple speakers and multiple phrases, but useful qualitative information can be gained by comparing results for a single phrase and a single speaker.

The individual trial results in Figure 16 show that the average rhythmicity is consistent for each trial for Japanese. This is perhaps because every syllable is enunciated in Japanese and is given equal length of pronunciation. Since every Japanese sound is a combination of a vowel and a constant sound, the vowel sound is repeated in each sound, and the speech can be very rhythmic. On the other hand, both the average and variance varied for English, even though the speaker was saying the same phrases repeatedly. For Trial 6, the 95% confidence interval is wide, meaning that the speaker did not say the phrase very rhythmically. On the other hand, Trial 7 has a very tight confidence interval. This shows that the speaker can easily change her speech to become rhythmic or arrhythmic because English speech is controlled by the volume of speech.

The two numbers to the right of the double dashed lines in Figure 16 show the rhythmicities of Japanese and English: 0.454±0.086 and 1.44±0.52 respectively. The low rhythmicity value means that Japanese is more rhythmic, and the tight 95% confidence interval shows that Japanese speech is consistently rhythmical over the 5 phrases. On the other hand, English has mixed beats. Since the 95% confidence intervals do not overlap for the two languages, the rhythmicities of Japanese and English are significantly different. In fact, Student's t-test shows that the rhythmicity value is lower for Japanese than English with 98% confidence.

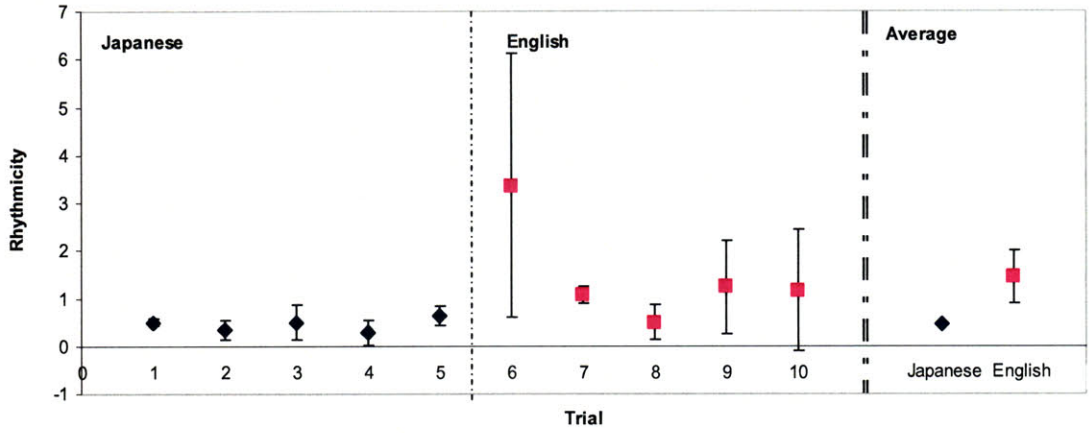

Figure 16 The rhythmicity of the two languages for one speaker for one phrase. Japanese is more rhythmic than English.

4.2.3 Rhythm of the Languages

The rhythmicity of Japanese is 0.84±0.02 and that of English is 1.35±0.02. Thus, Japanese is more rhythmic than English. Furthermore, the 95% confidence interval does not overlap at all, hence the rhythms of the two languages are significantly different. Surprisingly, the standard deviation of the average rhythmicity for English resulted in a very small variance, when the confidence interval was calculated from all 125 sets of data used in this experiment, as shown in Figure 17.

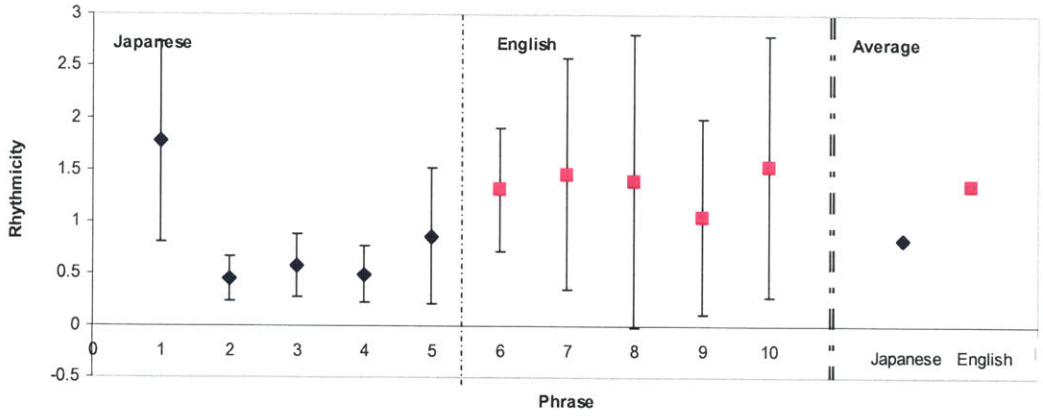

Figure 17 Rhythmicities of each language for all phrases. Japanese is more rhythmic.

4.3 Frequency of the Languages

The average 80th percentile frequency of all five phrases across all speakers is 1407±242 Hz for Japanese and 2021±642Hz for English, as shown in Figure 18. Thus, Japanese has lower frequency than English, perhaps due to its one-to-one vowel and consonant structure. Also, as mentioned in the Background section, English has many consonant sounds that make the frequency of the language high. The large uncertainties for both languages are expected because the 80th percentile frequencies are not normalized. A person with a higher tone of voice naturally has a higher frequency of speech. From Student's t-test, we are 99.5% confident that the frequency of English is higher than that of Japanese.

These results can be compared to literature values: Japanese typically has frequencies between 400-1200 Hz and American English has frequencies between 1500-5000Hz.[5] For English, the experimental frequency value lies within the generally accepted range, but for Japanese, it is 207Hz higher than the upper bound of the accepted value range. This could be caused due to some bias in the sample or the phrases that were chosen for this experiment.

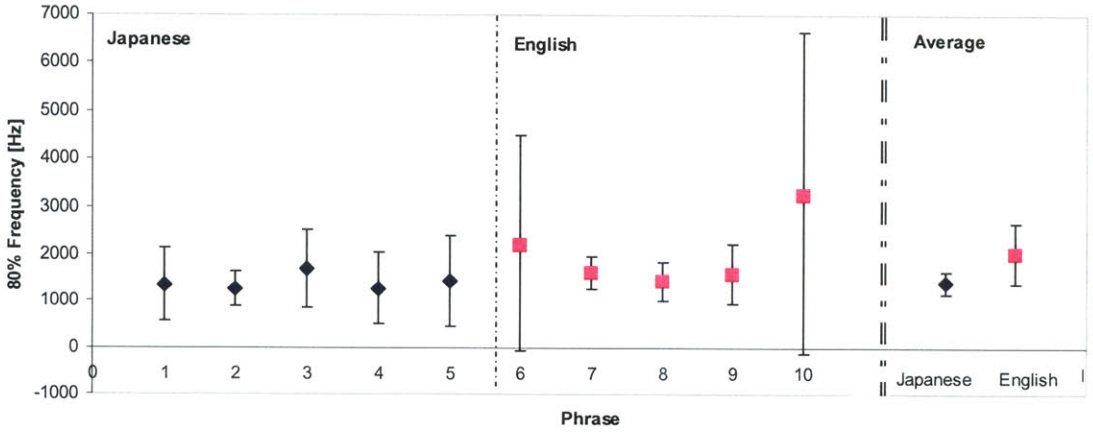

Figure 18 The average 80th percentile frequency of each phrase averaged over 5 speakers for each language. The overall average for each language is plotted at right. This tells the average frequency of the two languages.

4.4 Summary

The results are summarized in Table 3. They show that Japanese is more harmonious and rhythmic, and has lower frequency than English.

Table 2 Summary of Key Parameters

	Japanese		English		t-test
	Experimental	**Expected**	**Experimental**	**Expected**	
HNR [dB]	9.6±0.6	-	8.9±0.4	-	76%
Harmonicity [Hz]	27±13	-	41±26	-	90%
Rhythmicity	0.84±0.02	-	1.35±0.02	-	~100%
80% Frequency [Hz]	1407±242	400 – 1200	2021±642	1500 – 5000	99.5%

5. Conclusions and Recommendations

Japanese is a more harmonious (musical) and rhythmic language than English due to its vowel and consonant structure. We can also conclude, from the controlled frequency spacing of Japanese speech that Japanese is a pitch type language and English is not. Finally, the 80^{th} percentile frequencies of the two languages show that the frequency of Japanese is on average lower than that of English, because of the one-to-one vowel and consonant structure in Japanese and the presence of many consonant sounds in English.

The usefulness of the new parameters, harmonicity and rhythmicity, should be tested by applying them to variety of languages. The effect of dialects of each speaker should also be investigated, since some dialects in Japanese are stress-based, like English speech. For future experiments, we recommend collecting larger sample of subjects and recording more trials for each phrase to reduce uncertainty.

Furthermore, this experiment is able to suggest that English is not a pitch type language, but it is not able to confirm that English is a stress type language. We would need to analyze the intensity of volume in the time domain to confirm this result. The volume intensity may be implied in its non-harmonious and arrhythmic nature, but further investigation must be done to confirm this.

The methods of this paper may be applied to gaining understanding of the second language acquisition by bilingual speakers. Both primary and secondary language voice samples of people with varying fluencies in each language can be collected to reveal the effect of the native language's pitch, rhythm, and frequency characteristics on the second language. Perhaps a tool based on these findings can be built to help the learners of new languages correct their accents and learn to speak the second language better. This would be especially useful for learners of Japanese, as native-like speech can be acquired by practicing to pronounce the sounds consistently.

References

[1] Iwate University, "Acoustic Analysis of Japanese Spoken Language", http://sp.cis.iwate-u.ac.jp/sp/lesson/j/doc/accent.html

[2] University of Washington, Symbols for American English Vowel Sounds, http://faculty.washington.edu/dillon/PhonResources/newstart.html

[3] M. Ross, Rehabilitation Engineering Research Center on Hearing Enhancement, http://www.hearingresearch.org/Dr.Ross/Audiogram/Audiogram.htm

[4] R. J. Baken, "Clinical Measurement of Speech and Voice", London: Taylor and Francis Ltd, 1987.

[5]K. Murase, "Most Effective Method to Learn a Foreign Language", Nihon Zitsugyou Shuoppansha.

[6] P. Bourke, University of Western Australia, http://local.wasp.uwa.edu.au/~pbourke/other/dft/

[7] I. W. Hunter, "Fourier Analysis," 2.671 Measurement and Instrumentation, MIT, Spring 2008 (unpublished).

[8] I. W. Hunter, "Sound Analysis," 2.671 Measurement and Instrumentation, MIT, Spring 2008 (unpublished).

[9] I. W. Hunter and B. J. Hughey, "Uncertainty Analysis," 2.671 Measurement and Instrumentation, MIT, Spring 2008 (unpublished).

[10] I. W. Hunter and B. J. Hughey, "AutoCorrelation Analysis of WAV Files," 2.671 Measurement and Instrumentation, MIT, Spring 2008 (unpublished).

[11] M. Ross, Rehabilitation Engineering Research Center on Hearing Enhancement, http://www.hearingresearch.org/Dr.Ross/Audiogram/Audiogram.htm

[12] P. Ladefoged, "Phonetic Data Analysis", Malden: Blackwell Publishing, 2003.

[14] L. R. Waugh and C.H. van Schooneveld, "The Melody of Language", Baltimore: University Park Press, 1980.

[15] D. Gibbon and H. Richter, "Intonation, Accent, and Rhythm", Berlin: Walter de Gruyter & Co., 1984.

[16] P. Warren, "Prosody and Parsing", Erlbaum: Psychology Press, 1996.

[17] K. J. Kohler, "Rhythm in Speech and Language, A New Research Paradigm", Phonetica 2009; 66:29–45

[18] E. Yumoto, W. J. Gould, and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness", Journal of Acoustic Society of America 1982: 71(6)

[19] J. C. Brown, "Frequency ratios of spectral components of musical sounds", Journal of Acoustic Society of America 1996: 99(2)

[20] Paul Boersmia, "Accurate Short-term analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound", Institute of Phonetic Sciences, University of Amsterdam, Proceedings 17 (1993), 97-110.

[21] Paul Boersma & David Weenink (2009): Praat: doing phonetics by computer (Version 5.1.05) [Computer program]. Retrieved May 1, 2009, from http://www.praat.org/

Appendices

A1 Summary of the Parameters Used in This Paper

HNR [dB]	$HNR = 10 \times \log\left(\frac{r'_x(\tau_{max})}{1 - r'_x(\tau_{max})}\right)$ where $r'_x(\tau_{max}) \equiv \left(\frac{r_H(0)}{r_x(0)}\right)$; $1 - r'_x(\tau_{max}) \equiv \left(\frac{r_N(0)}{r_x(0)}\right)$ from autocorrelation $r_x(\tau) = \int x(t)x(t+\tau)dt$ with time-lag τ of harmonious H(t) and noise N(t) of voice samples. HNR measures "the degree of acoustic periodicity" by comparing the relative magnitudes of the harmonics and the noise in the signal. A higher value of HNR means more harmonious sound.
H[Hz]	$H = \sigma(\Delta f_i) \quad \forall i = 1,2,...n$ where σ is the standard deviation and Δf is the peak frequency spacing in the Fourier spectrum of a recorded sound. Lower value of H means more harmonious sound.
Rh	$Rh = \frac{\delta}{f_0}$ where δ is the difference in frequencies between the 20th and 80th percentile integrated power in the Fourier transform of the autocorrelation function. It is normalized by f_0, which is the 50th percentile of the integrated power. Lower value of Rh means more rhythmic sound.

A2 Characteristics of Speakers

The following table summarizes the characteristics of the test subjects when their voice samples were collected:

Subject	Age	Gender	Home Town	First Language	Note
1	21	F	Osaka, Japan	Japanese	Osaka dialect, characterized by its "melodic" nature compared to the standard Japanese
2	24	M	Nagoya, Japan	Japanese	
3	50	M	Fukui, Japan	Japanese	Fukui dialect, spoken with up-and-down, sing-song like

					manner
4	49	F	Kyoto, Japan	Japanese	Kyoto Dialect, very similar to Osaka accent
5	21	M	Tokyo, Japan	Japanese	Born in Tokyo, moved to Singapore at the age of 7 until he came to MIT three years ago
6	20	F	Albany, NY	English	
7	18	M	Los Angeles, CA	English	Actor
8	28	M	San Francisco, CA	English	
9	21	F	Overland Park, KS	English	
10	20	M	Seattle, WA	English	

A3 Phrases Used for This Research

The phrases used in this study are famous saying for both languages.

Japanese

1	猿も木から落ちる (Saru mo ki kara ochiru) — Even monkeys fall from a tree. (Even experts make mistakes.)
2	河童の川流れ (Kappa no kawa nagare) – Even Kappa (legendary animal) drown in water. (Even experts make mistakes.)
3	猫に小判 (Neko ni koban) – A coin to a cat (Don't offer things to people who are incapable of appreciating them)
4	三日坊主 (Mikka bouzu) – "Three-day monk". It describes people who start things with big promises and great enthusiasm, but never see them through to the end.
5	類は友を呼ぶ (Rui ha tomo wo yobu) – Birds of a feather flock together.

English

1	A penny saved is a penny earned.
2	A cat has nine lives.
3	Every cloud has a silver lining.
4	The pen is mightier than the sword.
5	Three strikes and you're out.