Classification of Filipino Speech Rhythm Using Computational and Perceptual Approach

Timothy Israel D. Santos [a], Rowena Cristina L. Guevara Ph.D. [b]

Digital Signal Processing Laboratory, Electrical and Electronics Engineering Institute,
University of the Philippines, Diliman

[a]**tdsantos1@up.edu.ph**
[b]**gev@eee.upd.edu.ph**

Abstract. This study incorporates computational and perceptual methods to classify Filipino speech rhythm. Speech rhythm may be described as a language's distinguishing durational sound pattern, resulting from the complexity of the language's syllable inventory.[1] Computational methods involve the correlation of rhythm-types to acoustic features such as the vocalic and consonantal intervals, one of which is the implementation of Multivariate Discriminant Analysis (MDA). Perceptual methods involve contrasting the rhythm of an unclassified language from prototype syllable-timed and stress-timed sentences. In order to isolate rhythm from speech, a data-stripping technique called *flat sasasa* resynthesis was implemented wherein the consonants are replaced with /s/ and vowels with /a/, producing a resynthesized alternating "sasasa" sounds at a constant pitch (F0). The rhythm discrimination and classification were closely examined for consistency between the data modeling and listening test results. The computational experiment was able to show that an MDA classifier trained to distinguish English and Japanese sentences tend to label Filipino sentences as Japanese 67% of the time, vis-à-vis the perceptual experiment showing that the listeners perceive Filipino to be more similar with Japanese, this study shows computational and perceptual validation that Filipino is syllable-timed, just like Japanese.

Keywords: Computational Linguistics, Natural Language Processing, Psychoacoustics, Speech Analysis, Speech Rhythm

1 Introduction

Speech prosody is an important aspect of natural speech. Pitch, duration, and intensity are some of the prosodic components that dictate the naturalness of speech and differentiate contextual meaning. The objective of this study is to present computational analyses and perceptual experiments to confirm the existence and applicability of rhythm classes to Filipino; to find their salient features for computational classification; and to find the rhythm typology of Filipino speech. Among the prosodic characteristics of speech, this study focuses on the temporal or the duration aspect.

Speech prosody or suprasegmentals refer to the features of speech that can modify the meaning and information being carried by the words not found in the lexicon. Speech intensity, duration, pitch, and rhythm are the most common features that differentiate one prosodic unit from another. Embedded in the human perception is the distinction of speech rhythm or timing to discriminate between languages. Languages are traditionally classified as stress-timed or syllable-timed where syllable-timed language were said to exhibit near-equal duration for each syllable, and stress-timed languages were said to have near-equal duration between stresses or

[1] Some languages allow a variety of complex syllables C^nV, CVC, etc., while some languages strictly adhere to simpler CV syllables (C-consonant, V-vowel).

what is called isochron*y* (Pike, 1943). Years of research on the subject has shown development in which evidences have been piling up against isochrony theory, and there emerged other rhythm classes such as Mora-timing (Abercrombie, 1967; Benadon, 2009). New models have been developed to accommodate intermediate languages, and a continuous language distribution plane was proposed (Dauer, 1983).

Instead of actually having equal duration between stresses and syllables, Dauer observed that the impression of syllable saliency for stress-timed and equal-saliency for syllable-timed is a result of the phonetic inventory, specifically syllable structures, of a language. Dauer's continuously-distributed model of rhythm postulates that a characteristic typical to stress-timed languages is a wider syllable inventory (occurrence of C^nV, CVC, etc.) resulting in heavier intervals. As a result, there are more frequent vowel reductions in the unstressed syllables, in a manner of speaking a *compensation* for the lengthening. These elements are consistent with stress-timed language's underlying concept of syllable stress that makes some syllables longer than others. For languages that seem to be more syllable-timed, it's just the opposite of the stress-timed. The simplicity of syllable (frequency of the simpler CV syllables) and fewer vowel reduction results in more *controlled* syllable durations (Fenk & Fenk-Oczlon, 2006). From these characteristics, some prefer the term *compensating* and *controlling* instead of the traditional syllable and stress timing, respectively.

Despite Japanese being classified as a distinct rhythm class called Mora (Bloch, 1950), in the course of this paper, we refer to Japanese as "syllable-timed," in adherence to the phonological model where we considered Japanese Mora as an extreme case of syllable-timing. As shown in (Ramus, Nespor, & Mehler, 1999), Japanese is located at the extreme syllable-timed side of the syllable/stress-time spectrum. Similarly, English was chosen since it has a great degree of stress-timing characteristic (Tajima , 1998).

2 Material

For this project, 120 sentences were taken each from Filipino and two other rhythmically distinct prototype languages. English was chosen as the prototype for highly-compensating rhythm or stress-timed, and Japanese for highly-controlling rhythm or syllable-timed. Most of the sentences are declarative, only less than 20 percent are interrogative. The selection was done with a couple of guidelines in mind. Syllables were kept at around 11 to 19 per sentence; as much as possible, and the average duration was kept at about 3 seconds per sentence, ranging from 2 to 4 seconds.

2.1 Corpora

The sentences came from the Filipino Speech Corpus (FSC), TIMIT English Corpus, and the Japanese Multext-J corpus. The mentioned language corpora contained phonetically transcribed sentences done by native speakers of the respective language. The following are more specific criteria since each corpus had different design and content:

TIMIT

- The sentences taken were the 5 phonetically-balanced sentences (SX sentences) for 24 speakers.
- To minimize dialectal differences within the speakers, the selected sentences came from the North Midland dialect region, considered to be the most neutral accent for American English.

MULTEXT-J

- Since there are only 6 speakers in all, 20 sentences were taken from each speaker's recordings.

- The corpus contained read-speech and emotionally performed speech. To simulate the natural language rhythm, sentences were taken from the semi-natural or emotionally performed set.

FILIPINO SPEECH CORPUS

- The 5 sentences taken were from 24 speakers.
- The selection was based on the average time of 2.5 seconds per sentence, after which sentences with 11 to 19 syllables were then chosen.

2.2 Transcription

Phoneme-labeled transcription files of the sentences were already available in the TIMIT and MULTEXT-J, transcribed by native speakers. The researcher, being a native speaker of Filipino, transcribed the sentences from the Filipino Speech Corpus using the PRAAT-Txtgrid tool. The DSP46 phoneme set in Table 1 was used, which was determined to be the unique phonetic inventory found in the Filipino Language (Guevara, et al., 2002).

Table 1: DSP 46 Phoneme Set

B	D	G	K	P	T	J	Ts	F	S	Sh	V	Z	M	N	Ng
Q	L	R	W	Y	A	E	I	O	U	Ag	At	Ak	Ad	Og	Hu
Aw	V	Ay	Oy	Iw	El	Em	En	Ha	He	Hi	Ho	Pau	Epi	Ow	

3 Experiment 1: Computational Analysis

The 120 sentence transcription in Filipino, English, and Japanese was processed to extract the duration for each phoneme. Each phone was labeled as either vowel or consonant, with the following rules to be applied: Pre-vocalic glides (as in Filipino /wa/: “wa·lâ” or /yâ/: “kan·yâ”) was treated as consonants, whereas post-vocalic glides (as in Filipino /áw/: “i·káw” or /áy/: “ka·máy”) were treated as vowels. This convention was based on the assumption that infants rely on a coarse segmentation of speech; the infant only distinguishes clusters of vowels and non-vowels. Adults have more perceptual cues taken into consideration since they have knowledge of phonotactics and have more linguistic memory. However, the rudimentary vowel-consonant coarse segmentation becomes apparent in situations where the sentences are lexically unintelligible to the adult listener. There are studies done to confirm this inherent human capability (Narayan, 2006), but a simple thought experiment could give us an idea. For example, when listening to a speaker of a completely unfamiliar language, the listener tends to just recognize an alternating pattern of the salient vowels and consonants while maintaining very minimal recognition of specific phonetic sounds. This is the reason why when one tries to mimic a foreign language, just a very small phoneme set is used whereas rhythm is actually retained; case in point, when one tries to mimic a speaker using *‘blah blah blah.’*

3.1 Acoustic Correlates

For each i-th sentence with N_i and M_i vocalic and consonantal intervals respectively, the duration of the n-th vocalic interval ($V_{i,n}$) and the duration of the m-th consonantal interval ($C_{i,m}$) were taken from the phonemic transcriptions. The interval values were then used to calculate the Proportion of Vocalic Interval (%V), and the standard deviation of the Vocalic (ΔV) and Consonantal (ΔC) Interval.

$$\%\boldsymbol{V} = \frac{\sum_{\boldsymbol{n=1}}^{\boldsymbol{N(i)}} \boldsymbol{V(i,n)}}{\sum_{\boldsymbol{n=1}}^{\boldsymbol{N(i)}} \boldsymbol{V(i,n)}+\sum_{\boldsymbol{m=1}}^{\boldsymbol{M(i)}} \boldsymbol{C(i,m)}} \qquad \textbf{(1)}$$

$$\Delta \boldsymbol{V}(\boldsymbol{i}) = \sqrt{\frac{\mathbf{1}}{\boldsymbol{N}(\boldsymbol{i})-\mathbf{1}}\sum_{\boldsymbol{n}=\mathbf{1}}^{\boldsymbol{N}(\boldsymbol{i})}(\boldsymbol{V}(\boldsymbol{i},\boldsymbol{n}) - \overline{\boldsymbol{V}}(\boldsymbol{i}))^{\mathbf{2}}} \qquad \textbf{(2)}$$

$$\Delta \boldsymbol{C}(\boldsymbol{i}) = \sqrt{\frac{\mathbf{1}}{\boldsymbol{M}(\boldsymbol{i})-\mathbf{1}}\sum_{\boldsymbol{m}=\mathbf{1}}^{\boldsymbol{M}(\boldsymbol{i})}(\boldsymbol{C}(\boldsymbol{i},\boldsymbol{m}) - \overline{\boldsymbol{C}}(\boldsymbol{i}))^{\mathbf{2}}} \qquad \textbf{(3)}$$

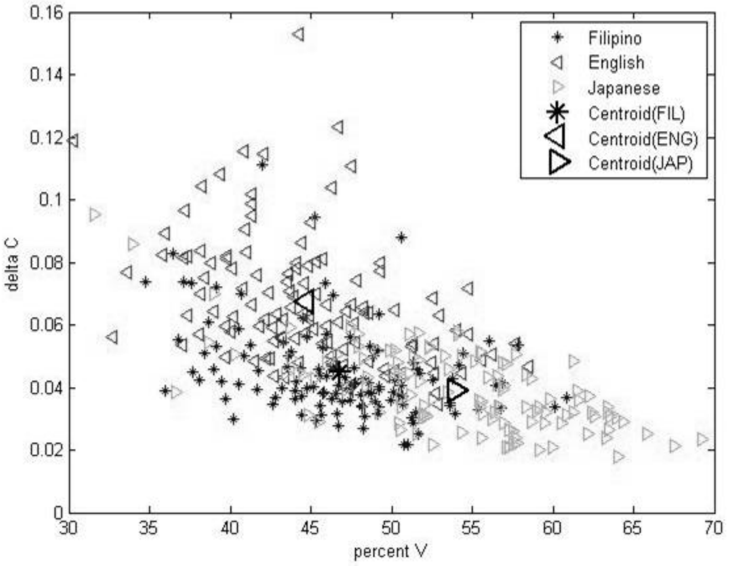


Figure 1: Rhythm distribution plane (using %V, ΔC). Each point is a representation of a sentence in the feature plane.

Table 2: Acoustic Correlates

Language	%V	ΔC	ΔV
English	44.8	6.74	5.34
Filipino	46.8	4.482	3.94
Japanese	54.0	3.91	3.90

Figure 1 is the mapping of the correlates of each sentence on a continuous rhythmic scale as proposed by Dauer (1983). Further to the right is the region for syllable-timed characteristic and further to the left is the stress-timed characteristic. The location of the Filipino sentences on the rhythmic scale suggests that Filipino may be an intermediate language, neither being extremely stress-timed nor syllable-timed as supported by Nespor (1990). With the Filipino situated in the middle of the extreme cases, it would be more suitable to describe Filipino Speech rhythm in terms of linguistic distance or similarity to the two rhythm classes.

3.2 Multivariate Discriminant Analysis (MDA)

MDA tries to predict the classification of an observation by fitting multivariate normal density using different covariance estimates. MDA was performed on a 2-group and 3-group MDA which can estimate the discriminant functions such as Linear, Diagonal-Linear, and Quadratic.

Three-group MDA

To confirm the predictive significance of the acoustic correlates on the rhythm discrimination, we performed 3-group MDA beginning with previously established 2-features (%V, ΔC), adding up to 5 features (%V, ΔC, ΔV, duration, syllables). Table 3 is the summary of performance of the quadratic classifier using different feature sets. It can be observed that aside from %V and ΔC, ΔV is also a good predictor for rhythm classification which is consistent with (Ramus, Nespor, & Mehler, 1999). It can be said that the results support the suggestion of the Phonological model of rhythm that the variation of rhythm types is a product of respective phonological properties, and that these properties can be independent and cumulative in forming the distinction of a language's rhythm.

Table 3: Classifier performance & Feature Evaluation

Features	(%V ΔC)	(%V ΔC) + n_{syll}	(%V ΔC) + t_{sent}	(%V ΔC) + ΔV	(%V ΔC) + ΔV, t_{sent}	(%V ΔC) + ΔV t_{sent}, n_{syll}
$Error_{Eng}$	19.17	16.67	15.83	22.5	21.67	20.83
$Error_{Fil}$	46.67	50	50	45.83	42.5	35
$Error_{Jap}$	32.5	31.67	25.83	20.83	17.5	17.5
$P_{Correct}$	67.22	67.22	69.44	70.28	72.78	75.56

Two-group MDA

Using 2-group MDA, the probability that the rhythm of a Filipino sentence was classified as English or Japanese was evaluated. It is basically looking for the intersection of the distribution space of Japanese and English to that of Filipino. Fig. 2 shows the 2-group linear classifier trained using %V and ΔC labeling Filipino sentences as Japanese 66.7% of the time. From this result, the computational experiment suggests that Filipino intersects more with the Japanese distribution space and is therefore more similar with the Japanese syllable-timed rhythm.

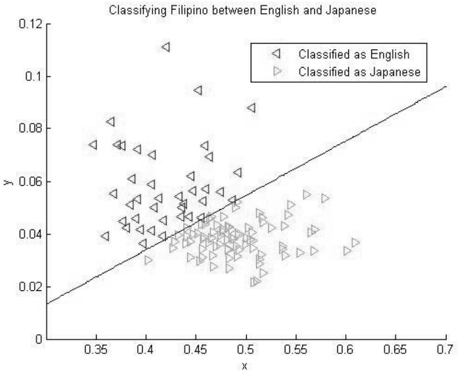


Figure 2: English-Japanese classifier with Filipino test data. The graph shows that the trained classifier identifies Filipino as Japanese 67% of the time.

4 Experiment 2: Perceptual Analysis

Auditory Signal Processing and Mathematical Psychoacoustics, the foundation of the subjective test performed in this study are widely used in pattern- classification, human auditory perception, and listener performance analyses (Leijon, 2008) (Green & Swets, 1966). Psychoacoustics tries to find the significant relationship between physical variables and its effects on psychological response. This psychoacoustic test was designed to compare the performance of a generated classifier to what is considered the ideal classifier, the human brain (Leijon, 2008), in order to validate or reject findings from the computational analysis.

4.1 Flat Sasasa Resynthesis

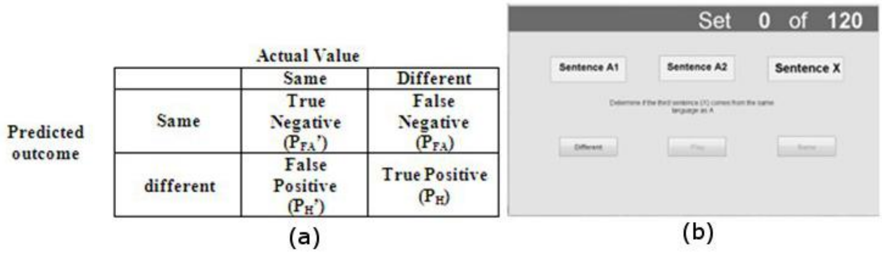
The Flat Sasasa Resynthesis is a data-stripping method to isolate the rhythmic characteristic of speech. It is composed of two phases, translation and synthesis. Firstly, the phonetic transcription was translated into unintelligible transcriptions, and then the original speech recordings were reproduced based on the transcriptions.

For the transcription translation, adjacent vowels were transformed into a long /a/ and adjacent consonants to one long /s/ sound, while maintaining the duration of the original vocalic and consonantal clusters. The outcome was a transcription of alternating vocalic and consonantal intervals, from which each vocalic interval was replaced by ‘a’ and each consonantal interval was replaced by ‘s.’ As an example, the phrase “higit sa lahat” will have the following *sasasa* translation : /h/ /i/ /g/ /i/ /ts/ /a/ /l/ /a/ /h/ /a/ /t/ - /s/ /a/ /s/ /a/ /s/ /a/ /s/ /a/ /s/ /a/ /s/.

The resynthesis was done by simply truncating and concatenating pre-recorded /s/ and /a/ segments, using the sasasa transcription as basis. The pitch was held constant by using only one source of /s/ and /a/ sound taken from one recording of the diphone /sa/. In order to lessen the effect of boundary mismatch, the duration of each /s/ and /a/ were controlled by dropping samples at the head and tail of the /sa/ diphone recording.

4.2 Listening Test

The listening test followed the AAX perceptual test developed by (Ramus, Dupoux, & Mehler, The psychological reality of rhythm classes: perceptual studies., 3 August 2003). It is a modified 2-Interval 2-Alternative Choice (2I2AFC) Test and Same-Different Test (SD), described by the confusion matrix Fig. 3 (a). For each language pair, the subjects were asked to listen to 3 resynthesized (*flat sasasa*) sentences per trial. The first two sentences were of the same language (AA in AAX), and functions as the context. The third was either the same or a different group (X in AAX). After listening to each AAX group, the subject was asked whether the third sentence (X) was the same language as the previous two (AA). The two alternatives being ***same*** and ***different***, makes it basically a Same-Different (SD) test despite taking the form of 3-Interval 2-Alternative Choice (3I2AFC). The inter-stimuli-interval was 500ms as prescribed by RNM which was consistent with the suggestions from (Levitt, 1971) for psychoacoustic tests. The test subjects were given visual cues on the GUI as to which sentence was playing.



		Actual Value	
		Same	Different
Predicted outcome	Same	True Negative (P_{FA}’)	False Negative (P_{FA})
	different	False Positive (P_H’)	True Positive (P_H)

Set 0 of 120

Sentence A1 Sentence A2 Sentence X

Determine if the third sentence (X) comes from the same language as A

Figure 3: AAX Listening Test (a) Confusion Matrix (b) GUI.

There were 40 AAX triplets for each language pairing of the three languages: Filipino-Japanese (FJ), Japanese-English (JE), and Filipino-English (FE). The trials were randomly distributed in the test, for a total of 120 trials.

It is a basic assumption in psychophysical experiments that there is a level of observer fallibility wherein the observer does not only ‘miss’ the presence of a response, but also can falsely assume a positive response (Green & Swets, 1966). Hit and false alarm rates are computed from the correct percent scores. Hit rate (P_H) is the proportion of correct “same” trials and false alarm rate (P_{FA}) is the proportion of incorrect “different” trials as indicated in Fig. 3-a. Without going through the intricacies of the complex internal processing procedures of perception, we simplify the analysis of the listening test results from Table 3 by establishing a baseline for discriminability between two disjoint rhythm groups in the JE experiment, and looking for a language pair that has almost a chance-level or "good-as-guess" result of 50%.

Table 3: Listening Test Results

	Fil-Eng(FE)	Fil-Jap (FJ)	Jap-Eng(JE)
P_{FA}'	64.47 %	59.47 %	73.16 %
P_H	57.11 %	52.89 %	62.63 %
P_{Cmax}	60.79 %	56.18 %	67.89 %

The listening test is comprised of 19 listeners, 12 were male and 7 were female. As expected, listeners have performed well in distinguishing Japanese from English just by rhythm alone. Higher discriminability scores for JE test than both the FE and FJ tests is consistent with the result from the objective feature-plane that Filipino rhythm is in-between Japanese and English. For the discrimination of Filipino from English and Japanese, listeners have performed relatively higher for the FE tests. This suggests that Filipino is more similar with the Japanese rhythm, causing the confusion as manifested by the 52% true-positive score (P_H), very near the 50% 'good as guess' performance.

5 Conclusions

Placing the results on the feature plane, Filipino sentences lie in between Japanese and English, consistent with the suggestion that instead of discrete rhythm classes, rhythm is continuously distributed where opposite sides represent extreme and ideal conditions of syllable and stress timing. On a computational perspective, the acoustic correlates were shown to contribute cumulatively to the distinction of the rhythm of a language. A 3-group, 2-feature to 5-feature MDA classifier was able to identify the languages without the aid of lexical or semantic information at a rate of 67% to 75% accuracy. This characteristic of rhythm, and the significance of the features selected is valuable for Language Identification (LID) systems at the front-end of multilingual speech recognition systems. We have shown that the features for speech rhythm investigated may be used for objective speech rhythm and accent discrimination; hence, the speech rhythm features extraction may be automated to be used in modern automatic speech systems. It can be further expanded to not just discriminate between language, but to be able to distinguish whether a certain sentence is spoken in either its natural rhythm or in a different 'accent' due to the interference of another language (probably a speaker's native tongue). This is especially useful in the context of second-language learning. Language development, historical linguistics, and other linguistic issues can benefit from similarly-done studies to observe the changes in the natural rhythm of Filipino from recordings a couple of years ago to the modern globally-influenced Filipino language.

The computational and perceptual experiments yielded supporting results where the MDA predictive model classified the Filipino sentences as Japanese 67% of the time, despite Filipino being closer to English in the sense of Eucledian-mean distance. The perceptual experiment showed that human listeners have a hard time of telling the difference between Japanese and Filipino when presented with the rhythm-isolated sentence, where the correct distinction 52% was very near the chance-level of 50%. Both experiments done in respective objective and subjective manner were showing compelling evidence that Filipino is rhythmically similar to the rhythm of Japanese which is syllable-timed.

Table 4: Computational and Perceptual Results Analysis

	Filipino is Similar to English therefore **Stress-Timed if:**	Filipino is Similar to Japanese therefore **Syllable-Timed if:**
Objective Test: Multivariate Discriminant Analysis (MDA)	P (Eng|x=Fil) > P(Jap|x=Fil)	P(Jap|x=Fil)> P (Eng|x=Fil)
Subjective Test: Same-Different Test	P(correct|Fil-Jap) > P(correct|Fil-Eng)	P(correct|Fil-Eng) > P(correct|Fil-Jap)

The combination of computational and perceptual analysis, as summarized in Table 4, can be used as a good interdisciplinary template for resolving some of the issues in Filipino language standardization, and for performing similar tasks for other Philippine-type languages.

Acknowledgement

We would like to thank the Department of Science and Technology's Philippine Council on Industry, Energy, and Emerging Technology Research & Development (PCIEERD) and Engineering Research and Development for Technology (ERDT) for supporting this research endeavor.

References

Abercrombie, D. (1967). Elements of General Phonetics. Chicago: Aldine.

Benadon, F. (2009). Speech Rhythms and Metric Frames. *Communications in Computer and Information Science, 38*, 22-31.

Bloch, B. (1950). Studies in colloquial Japanese IV: Phonemics. *Language 26*, 86-125.

Dauer, R. (1983). Stress-timing and Syllable-timing Reanalyzed. *Journal of Phonetics 11*, 51-69.

Fenk, A., & Fenk-Oczlon, G. (2006). Crosslinguistic Computation and a rhythm-based classification of languages. Data Information Analysis to Knowledge Engineering. *Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation*, (pp. 350-357). Berlin/Heidelberg:Springer.

Green, D., & Swets, J. (1966). *Signal Detection Theory and Psychophysics*. Cambridge: John Wiley & Sons, Inc.

Guevara, R., Co, M., Tan, E., Garcia, I., Espina, E., Ensomo, R., et al. (2002). Development of a Filipino Speech Corpus. *3rd National ECE Conference.*

Leijon, A. (2008, May 14). *Auditory Signal Processing-Mathematical Psychoacoustics.* Retrieved September 29, 2010, from KTH Electrical Engineering: www.ee.kth.se/sip/courses/FEN3100/

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *J. Acoustic Society Am., 49(2)*, 479-483.

Narayan, C. (2006). Acoustic-perceptual salience and developmental speech perception.

Nespor, M. (1990). On the rhythm parameter in phonology. *Logical issues in language acquisition*, (pp. 157-175). Dordrecht: Foris.

Pike, K. L. (1943). *Phonetics: a critical analysis of phonetic theory and a technic for the practical description of sounds.* Ann Arbor: University of Michigan.

Ramus, F., Dupoux, E., & Mehler, J. (3 August 2003). The psychological reality of rhythm classes: perceptual studies. *15th International Congress of Phonetic Sciences*, (pp. 337-342). Barcelona.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of Linguistic Rhythm in the Speech Signal. *Cognition*.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of Linguistic Rhythm in the Speech Signal. *Cognition*.