強さ曲線と変調スペクトルの日本語音節との関係

小松 雅彦(北海道医療大学心理科学部) 荒井 隆行(上智大学理工学部)
koma2@hoku-iryo-u.ac.jp, arai@sophia.ac.jp

1. はじめに

「音節」は、リズム分析の重要な概念であり、2 通りの見方ができる。1 つは、音節を強さの大小の変化パターンであると見なす方法である(非分節的アプローチ)。すなわち、振幅が小さな値から大きくなってまた小さくなるのが音節であると考える。もう 1 つは、音節を子音と母音から構成される構造物であると見なす方法である(分節的アプローチ)。このアプローチでは、リズムを音節構造の複雑さ等と関連付け、子音区間や母音区間の長さを測定してリズムを計算するといったことを行う(Ramus et al., 1999; Grabe & Low, 2002)。もともと音声というのは口の開閉の繰り返しであって、口が大きく開いている時には振幅が大きくなり母音が生成され、口が閉じてくると振幅が小さくなり子音が生成されるわけなので、上述の 2 つのアプローチは同じ音声の性質を異なった視点から見ているに過ぎない。しかし、両者の実際の分析手法は大きく異なる。

本研究では、これら 2 つのアプローチの橋渡しをすることを目的として、強さ曲線の音響的分析結果(非分節的アプローチ)と音素により定義される単位(分節的アプローチ)の関係を論じる。具体的には、コーパス(日本語 MULTEXT)の音声データの強さ曲線の音響分析結果を、コーパス付属の音素ラベルに対して照合するという手法を用いる。

まず 2 節では、強さ曲線がどの程度正確に音節の位置を表しているのかを調べる。一定のアルゴリズム(Komatsu & Arai, 2003; Komatsu & Miyakoda, in press; Aoki et al., 2002)により強さ曲線の局所ピークを抽出し、抽出された局所ピークの位置をコーパスの音素ラベルと照合する。その結果、局所ピークが非常に正確に音節の位置を示していることを示す。(音節境界についての議論はしない。)

次に 3 節では、強さ曲線から算出された変調スペクトルについて検討する。変調スペクトルの発話了解度における重要性やリズムとの関係は過去に議論されてきている(Houtgast & Steeneken, 1985; Arai & Greenberg, 1997; Greenberg & Arai, 2004)。その中で、日本語と英語の変調スペクトルの類似性が指摘されており、英語の変調スペクトルと音節長の関係は詳細に調べられている。日本語の変調スペクトルと音節の関係も検討されているが、本研究では、それを更に詳細に推し進め、4 Hz のピークが 2 モーラまたは 2 音節の脚と関係があるのではないかという示唆をする。

本研究では、まず、強さ曲線の局所ピーク抽出を様々な設定で行うことにより、強さ曲線が音節位置を示していることを示し、次に、強さ曲線から得られた変調スペクトルと音素ラベルから得られた音節長の検討により、変調スペクトルのピークと脚の関係を示唆する。これらの検討により、音節を一見異なるように分析している非分節的アプローチと分節的アプローチの関連性を示し、音節・脚の重要性を指摘する。

2. 強さ曲線による音節の検出

2.1 実験方法

本実験では、強さ曲線から音節がどの程度正確に検出できるかを調べた。音節の核(典型的には母音、V)の振幅は通常は両端(典型的には子音、C)より大きいので、強さ曲線の局所ピークは音節核の位置を示すことが予想される。実験では 2 種類のアルゴリズムを用いてピークを抽出し、

その位置(時間)をコーパス付属の音素ラベルと照合した。抽出されたピークが、音素ラベルの V の位置にあれば、音節が正しく検出されたと判定した。

実験では、日本語 MULTEXT(北澤, 2004)の全音声データを分析した。日本語 MULTEXT には、合計で480個の日本語の短い発話の音声とその音素ラベルが収録されている。話者は男女各 3 名で、各自が朗読と模擬自発発話の 2 つの発話スタイルで読んでいる。ただし、発話スタイルによる実験結果の違いはほとんどなく、本節では紙幅の都合で朗読の結果のみ報告する。

2.2 局所ピーク抽出アルゴリズム

強さ曲線から音節位置を表しているであろう局所ピークを抽出する方法として、次の 2 通りのアルゴリズムを用いた。

(1)RMS ピークによる方法:音声信号の値を 2 乗し、ハニング窓を乗じて、RMS を求めた。分析窓は 8 ms ごとに移動し、窓長は 32, 64, 128, 256, 512 ms の 5 種類とした。(ハニング窓の乗算は緩やかな低域通過フィルタの効果を持つ。各窓長での−3dB カットオフは、22.5, 11.3, 5.6, 2.8, 1.4 Hz、メインローブ上限は、62.5, 31.3, 15.6, 7.8, 3.9 Hz。)上記の結果得られた RMS 曲線のすべての極大値のうち、閾値以上のものを抽出した。閾値は、RMS 曲線の最大値の 0, 0.2, 0.4, 0.6, 0.8 倍とした(閾値 0 ではすべての極大値が抽出される)。窓長を長くすると、RMS 曲線をなだらかにする効果があり、閾値を大きくすると、振幅の小さな要素(望むべくは子音と雑音)を捨てる効果がある。

(2)相関係数のピークによる方法:音声信号の値を 2 乗し、基準信号(余弦波 1 周期、ハニング窓と同じ形)との相関係数を計算した。音節の中心では振幅が大きく両端では小さいので、中心の値が大きく両端の値が小さい基準信号との相関係数は、両者が時間的に重なった位置で大きくなる。分析窓(基準信号)は 8 ms ごとに移動し、窓長(基準信号長)は 32, 64, 128, 256, 512 ms とした。典型的な音節が4 Hz で生起するとすれば、256 ms (3.9 Hz)の基準信号がもっとも高い相関を示すと予想される。8 ms ごとに求めた相関係数の値から、すべての極大値のうち閾値以上のものを抽出した。閾値は、最大値の 0, 0.2, 0.4, 0.6, 0.8 倍とした(常に正の値のもののみ抽出)。窓長と閾値を増すと、RMS ピークによるアルゴリズムの場合と似た効果があると推測される。

上記2つのアルゴリズムの発想はまったく異なるが、実際の処理は似ている。相関係数の計算は、いわば正規化された乗算である。つまり、音声信号に時間窓をかけるのと、音声信号と基準信号の相関係数を求める計算は、似ていると言える。ここでは、これらの似て非なる方法を比較するため、両者に共通な時間窓/基準信号の形を用い、パラメタの設定(窓長/基準信号長、閾値)も共通としている。

2.3 実験結果(音素ラベルとの照合)

各アルゴリズムによって抽出された局所ピークの位置をコーパス中の音素ラベルと照合し、どの程度正確に音節に対応しているかを評価した。ここでは、先行研究(Ramus et al., 1999; Grabe & Low, 2002)に従い、CV(C)がリズムの構成要素であると考える(便宜的に、この CV(C)を音節と呼ぶ)。C と V はそれぞれ 1 個以上の子音または母音からなる。言い換えれば、連続する母音は 1 つの母音要素(以下、V 要素)、連続する子音は 1 つの子音要素(以下、C 要素)とみなす。本実験では、もし音素ラベル中の各 V 要素の位置に 1 つだけピークがありC 要素の位置にピークがなければ、音節検出は成功したと判断する。もし、V 要素の位置にピークがなければ、検出アルゴリズムはその要素を見逃したことになる。もし、V 要素の位置に 2 つ以上のピークがあったりC 要素の位置にピークがあったりすれば、過剰検出ということになる。ピーク抽出の結果は、どの程度正確に V 要素が検出されるのか(いくつの V 要素が検出されるのか)、および、どの程度正確にピークが V

要素を表しているのか(いくつのピークが正しいのか)、という2つの点から評価した。

2.3.1 RMS ピーク

全体的に、予想されたとおり、窓長と閾値が大きくなるに従い、見逃されたV要素が増え過剰検出されたV要素とC要素が減少した。これは、抽出されたピーク数が全体として減少したことによる。音節の検出は、正しく検出されたV要素の割合と正しく未検出であったC要素の割合の両方が高い時に、正確であると判断される。

図1(左)は、正しく検出されたV要素(1つだけピークがあったV要素)、正しく未検出であったC要素(ピークのなかったC要素)、正しく未検出であったポーズ要素(ピークのなかったポーズ要素)の割合を示している。グラフ中のV要素とC要素に着目すると、窓長64, 128 ms、閾値0, 0.2でもっとも音節が正確に検出されていることが分かる。ポーズ要素を考慮すると、閾値 0、すなわち閾値なしの結果は良くないので、最低限の閾値は必要であることが分かる。

図2(左)は、図1(左)に対応する要素数(正しく検出されたV要素数、正しく未検出であったC要素数、正しく未検出であったポーズ要素数)である。ポーズ要素は、そもそも総数が少ない。

図3(左)は、抽出されたピークが何をさしているかを示したものである。図1, 2が、いかに要素が検出されたか(いくつの要素が正確に検出されたか)を示すのに対し、図3は、抽出されたピークが何を表すか(いくつのピークが信頼できるか)を示す。V1 は音節検出として信頼できるピークの数で、他のものは誤った検出数である。もし、V要素に1つだけピークがあれば、それはV1として数えている。もし、V要素に2つ以上のピークがあれば、その中の1つをV1として数えそれ以外はV2+としている。CとPauは、C要素とポーズ要素中のピークである。ただし、本グラフ中では、1つの要素中に6個以上のピークが出現した場合は6個目以降は数えていない。

図3(左)は、窓長64, 128 msでV1の数が他の数よりも非常に大きいことを示している。また、閾値を0から0.2にすると、V1が少ししか減らないのに対し他のカテゴリが大幅に減ることも示している。これらのパラメタ設定は、図1, 2(左)でも良い結果であった。

2.3.2 相関係数のピーク

全体的に、窓長と閾値が大きくなるに従い、見逃されたV要素が増え過剰に検出されたV要素とC要素が減少した。これは、RMSピークと同様の傾向である。

しかし、予想に反して、窓長128 ms(閾値0, 0.2)が、もっとも良い結果を示した(図1〜3右参照)。これは、256 ms(3.9 Hz)でもっとも良く音節を捕らえられるという予想に反するものである。

相関係数による方法の最良のパラメタ設定(窓長128 ms、閾値0, 0.2)で得られた結果は、RMSによる方法の最良のパラメタ設定(窓長64, 128 ms、閾値0, 0.2)の結果とほとんど同じであった。

3. 変調スペクトル

図4は、2節で用いられた音声データの変調スペクトルを示す。朗読音声と模擬自発発話音声のスペクトルは、ほとんど同じ形をしている。両者とも先行研究から予想されるように4〜5 Hz前後にピークがあるが、模擬自発発話のピークの方がやや高い周波数にある。模擬自発発話の方が少し速く発話されているのかもしれない。

図5は、モーラ長の分布をモーラの種類別に示したものである。CV, CJVが他の種類よりも長い。

図6の「モーラ」(mora)は、図5の全種類を合計したものである。「音韻的音節」(phonological syllable)は、モーラ単独、あるいはそれにVのみからなるモーラおよび特殊モーラ(N, Q)が後続したものである。ただし、ここでの音節の定義は、音韻論的・形態論的な制約は考慮していないので、

音韻論で言われるものとは異なる。「音声的音節」(phonetic syllable)は、音韻的音節単独、あるいはそれに無声化モーラが先行、ポーズの前で後続したものである。(音声的音節が2節で言う音節に相当する。)言い換えると、音韻的音節、音声的音節は、モーラ1個から構成されるか、あるいは、それに1個以上の非C(J)Vモーラあるいは無声化モーラを加えたものから成る。朗読のグラフ中の200～250 ms付近の瘤(図6左)は、このような追加モーラが長さを保っていることを示唆しており、この部分が変調スペクトルの 4～5Hz のピークに寄与しているかもしれない。模擬自発音声(図 6右)では長さの変化が緩やかになっている。

図7は、連続する2モーラおよび2音節の長さを示したものである。すべての分布が、200～250 ms、あるいはそれに近いピークを持っている。この時間長は4～5 Hzに対応する。

図5では200～250 msのピークが見られず、図6では、2モーラ以上から構成されていると推定される音韻的音節、音声的音節が 200～250 ms 前後に分布している。さらに、図 7 ではピークが200～250 msにある。以上のことは、単独のモーラや音節ではなく、2モーラまたは2音節からなる単位、すなわち、脚が、変調スペクトルのピークに対応することを示唆している。日本語はモーラリズムであると言われるが、モーラだけでなく、英語同様、脚もリズムに関与している可能性がある。

4. 結論

音節はリズムにとって重要な概念であるが、非分節的アプローチと分節的アプローチは融合されていなかった。本研究は、強さ曲線が適切な分析パラメタの設定で音節位置を示し(2 節)、強さ曲線から得られた変調スペクトルのピークが2モーラまたは2音節からなる脚に対応することを示唆した(3 節)。非分節的アプローチと分節的アプローチの分析結果を対応させることにより、音節・脚が、英語同様、日本語のリズムにおいても重要な概念であると考えられることを指摘した。

参考文献

北澤茂良(編). (2004). *日本語MULTEXT*[CD-ROM]. 静岡大学, 浜松市.

Aoki, T., Komatsu, M., Arai, T., & Murahara, Y. (2002). Temporal envelope modulation using syllable search method for robust language identification. *Proc. Forum Acusticum Sevilla 2002* [CD-Rom].

Arai, T., & Greenberg, S. (1997). The temporal properties of spoken Japanese are similar to those of English. *Proc. Eurospeech '97,* 1011-1014.

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). Berlin: Mouton de Gruyter.

Greenberg, S., & Arai, T. (2004). What are the essential cues for understanding spoken language? *IEICE Trans. Inf. & Syst.*, vol. E87-D, no. 5, pp. 1059-1070.

Houtgast, T., & Steeneken, H. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J. Acoust. Soc. Am., 77,* 1069-1077.

Komatsu, M., & Arai, T. (2003). Acoustic realization of prosodic types: Constructing average syllables. *LACUS Forum, 29,* 259-269.

Komatsu, M., & Arai, T. (2006). What is rhythm? Can we capture syllable shapes from intensity contours? *IEICE Tech. Rep.*, vol. 105, no. 685, pp. 121-126.

Komatsu, M., & Miyakoda, H. (in press). Acoustic measurement of rhythm types: A stress language vs. a mora language. *Linguistik International, 15.* Bern: Peter Lang.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition, 73,* pp. 265-292.

* 本発表の一部は、Komatsu & Arai(2006)を改訂したものである。本研究は、一部、北海道医療大学個体差健康科学研究所の助成を受けています。

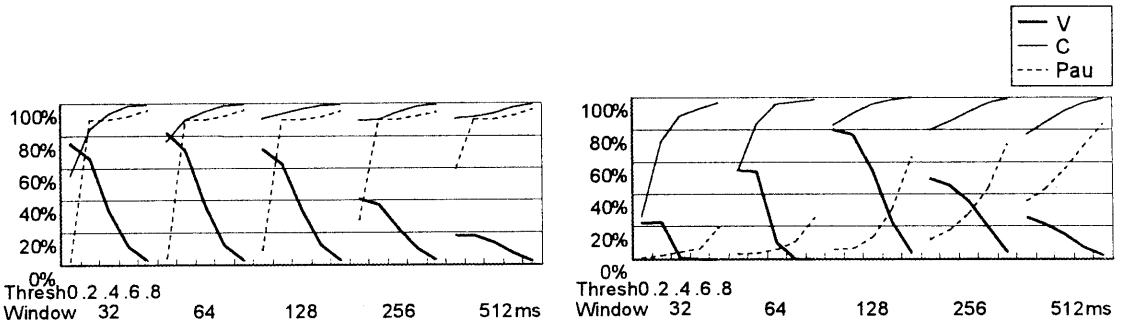

図 1: 正しい検出率／未検出率（左：RMS ピーク、右：相関係数のピーク）

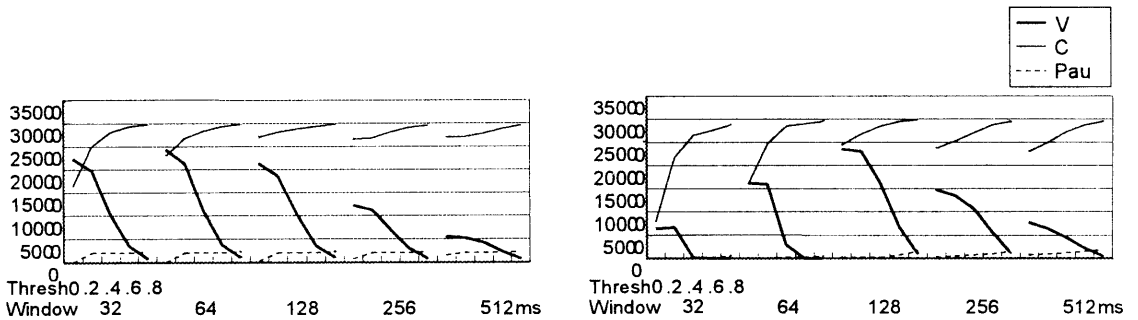

図 2: 正しい検出数／未検出数（左：RMS ピーク、右：相関係数のピーク）

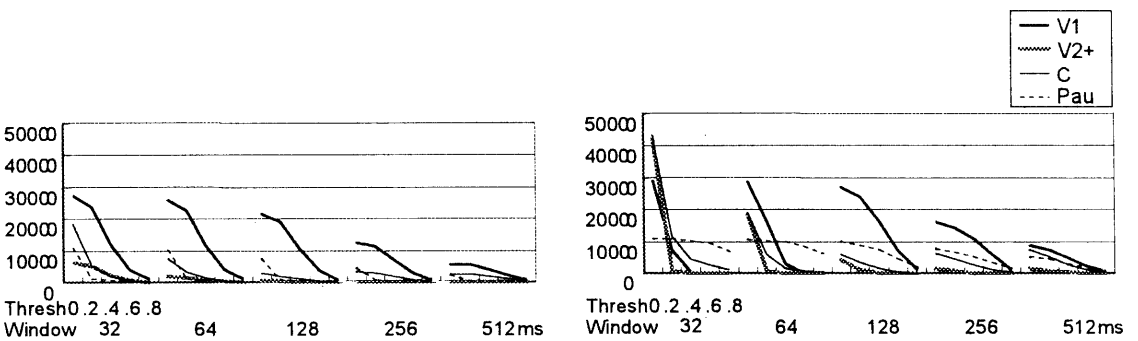

図 3: ピークの指すもの（左：RMS ピーク、右：相関係数のピーク）

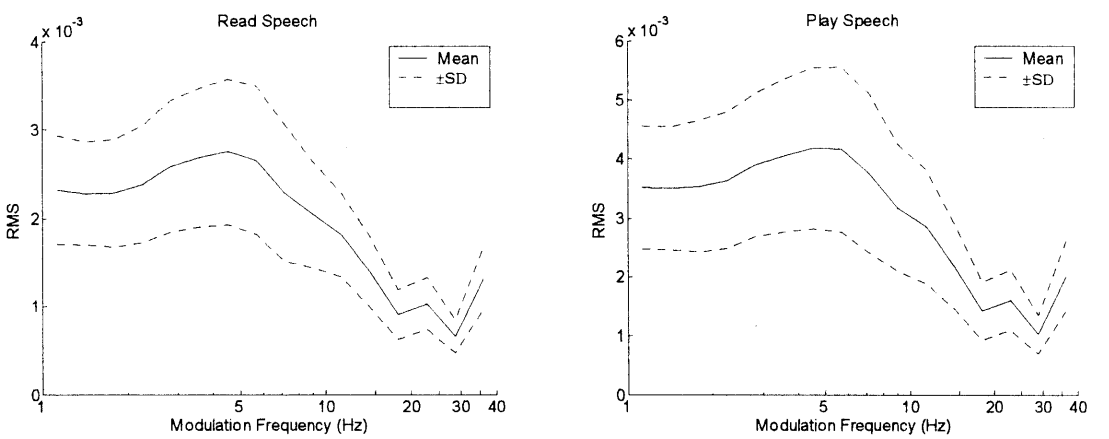

図 4: 変調スペクトル（左：朗読、右：模擬自発発話）

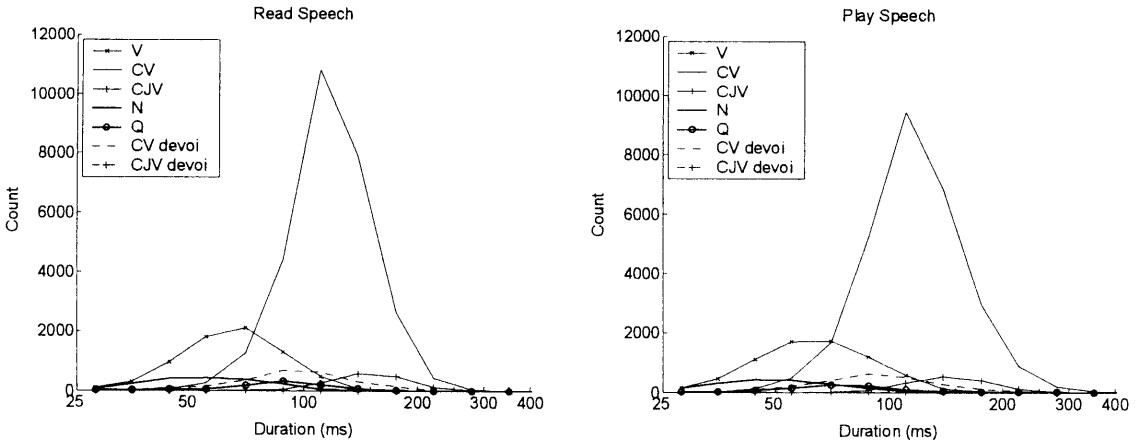

図 5: 種類ごとのモーラ長の分布(左:朗読、右:模擬自発発話)

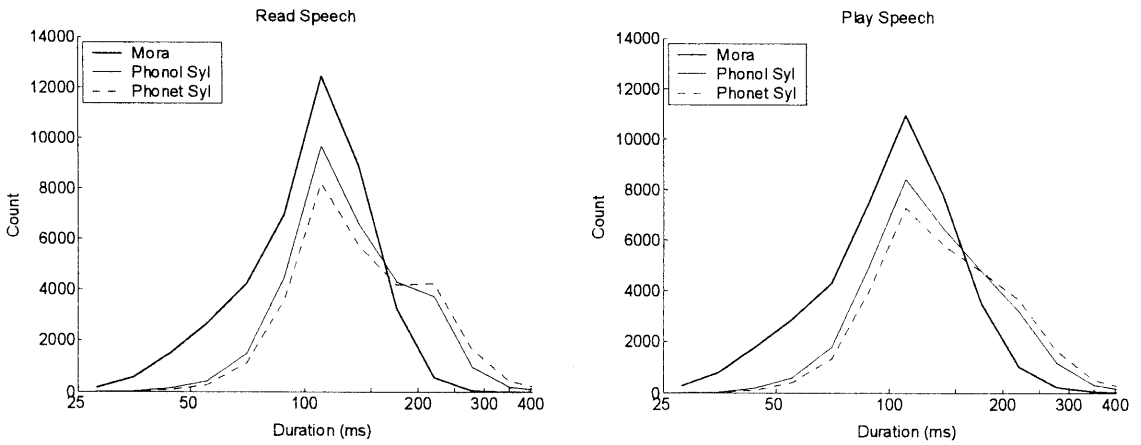

図 6: モーラ、音韻的音節、音声的音節の長さの分布 (左:朗読、右:模擬自発発話)

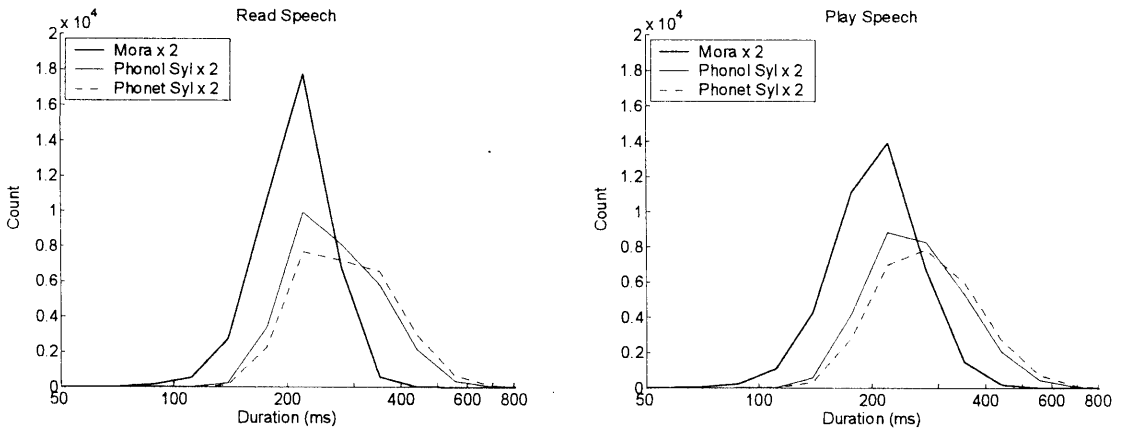

図 7: 連続した 2 つのモーラ、音韻的音節、音声的音節の長さの分布(左:朗読、右:模擬自発発話)