米語話者の日本人英語聴取誤りに対する要因分析

峯 松　信 明

東京大学大学院情報理工学系研究科

（発表者代理：毛利 太郎）

本研究の背景と目的 (#1)

- ***日本の英語教育の現状とそれを取り巻く環境***
  - ✔ 日本語・英語間の音声学的・言語学的差異＝大
    8 年間の英語学習で身に付く音声による語彙伝達能力 = SN比 -3.3 dB に相当（寂しい現実）
  - ✔ 「Native-like な英語」から「intelligible な英語」への学習目標の変化

- ***Intelligibilityって何？***
  - ✔ なまっていてもよい。伝われば，，。
    - ◆ 聞き手の適応能力（人間の高度な処理能力）を当てにした目標設定
  - ✔ 学習者のタスク減，しかし，教師のタスクは果てしなく**「増」**
    - ◆ 許容可・不可の判断には「母語話者の音声知覚プロセス」を持つことが最低限必要
    - ◆ そもそも日本人の英語教師にはその資格はない？　英語教育界が墓穴を掘っているだけ？
    - ◆ 発音教育の基盤を「音声学」から「音声学**＋認知科学**」へと移行させることが必要

- ***日本人教師にとってその獲得が困難な知覚特性が要求される教育戦略***
  - ✔ 学習者は，当然，そんな知覚特性なんて知る由も無い。
  - ✔ 計算機による模擬の可能性は？　Virtual Native Ears の実現は可能か？

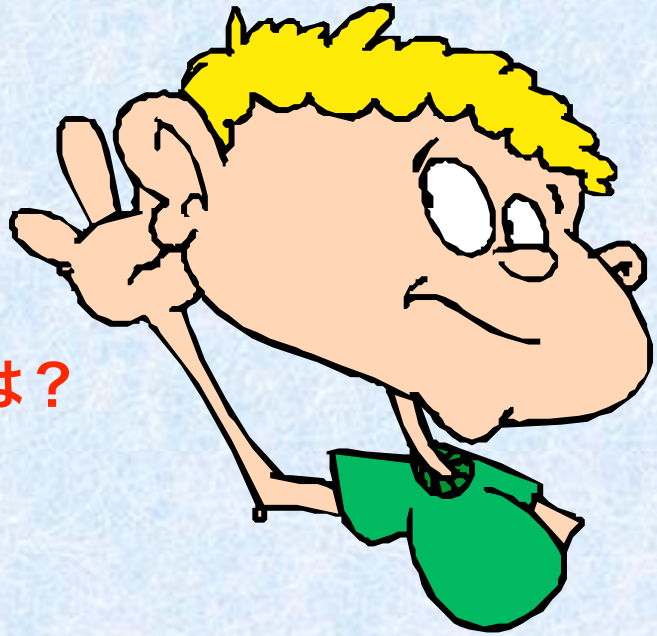
本研究の背景と目的 (#2)

- 「やっていいこと」「やっちゃいけないこと」
  - ✔ どのような発音誤りが母語話者の「聞き取り」を妨げるのか？
    - ◆ 音韻生成のまずさ？　リズム的な歪み？　使う語彙の問題？　或いはその組み合わせ？
  - ✔ Segmental or prosodic ?
    - ◆ 発音学習の本質は分節的側面？　韻律的側面？　　音声の本質から外れた議論か？
  - ✔ **音声 ＝ 分節 ＋ 韻律 ＋ 言語**
    - ◆ 三側面の融合として音声を捉える。
    - ◆ 単語 w の発声を分節属性 ***Si***, 韻律属性 ***Pi***, 言語属性 ***Li*** の組み合わせ及びその値で定義
    - ◆ **Q :** その発声の書き取り率は属性値を使って予測可能か？

連続音声中の単語 w

(S1, S2,,,P1, P2,,,L1, L2,,,) =
(0.3, 1.2,,,4.1, -1.2,,, 2.4, 0.1,,,)

聞き取り率は？



日本人読み上げ英語音声 DB

- ***種々の文セット，単語セット***
  - ✔ 分節的側面と韻律的側面
  - ✔ 文の読み上げと単語の読み上げ
  - ☞ 音素バランス / 韻律バリエーションセット X 文 / 単語セット
  - ☞ 一人約 120 文，220 単語の発声，男女 100 人ずつ（準ランダム抽出）

- ***音声収録の手順***
  1. ［収録前］収録に先立って読み上げ文 / 単語に対する発音練習を要請
  2. ［収録中］発声者自身が正しいと考える発音ができるまで繰り返しの発声を要請
  3. ［収録後］収録サイトの技術スタッフによるチェック
  - ☞ 発声者にとっては「正しい」英語であるが，発音誤りは山のように存在

- ***読み上げ ＝ 文法的誤りは無し？***
  - ✔ 厳密には言語的誤りは存在しない
  - ✔ 聴取実験で使用した文セット＝音素バランスセット
    - ◆ 出現頻度の低い（親密度の低い），語，句，言い回しを含む
    - ◆ 誤りではないが，低親密度表現として利用可能

提示音声の選択

- ***JE-DB（男声）からバランスのとれた文音声選択が必要***
  - ✔ 音素バランス文 460 文・発音習熟度の低い 90 人に着眼（上位者を無視, 90/95）
    - ◆ Better45 と poorer45 の 2 グループへ区分
  - ✔ 人間の短期記憶容量（7 chunks）を考慮し単語数により文を区分
    - ◆ 5 単語以下文，6・7 単語文，8 単語以上文
  - ✔ 単語の予測のし難さ（PP）に基づいて文を区分
    - ◆ Less / rather / more predictable → 最終的に文セットを 9 (3 x 3) つのサブセットに分割
  - ✔ 話者と文の対応
    - ◆ Better45 = より難しい 8 セットから 1 文ずつを発声
    - ◆ Poorer45 = より易しい 8 セットから 1 文ずつを発声
    - ◆ Better45 = 45 x 8 = 360，Poor45 = 45 x 8 = 360，合計 720 発声
  - ✔ 2 つの文音声セット
    - ◆ Set-a = [Better45 + Poorer45] / 2 = 360 発声
    - ◆ Set-b = 720 - Set-a = 360 発声
    - ◆ Set-a, Set-b は同一の 360 文，同一の 90 人。但し，文と話者の対応は異なる。
  - ✔ **文長，予測のし難さ，話者，習熟度の観点からバランスのとれたセット**
    - ◆ DB の存在により初めて可能となった提示音声セット

書き取り実験に向けての準備

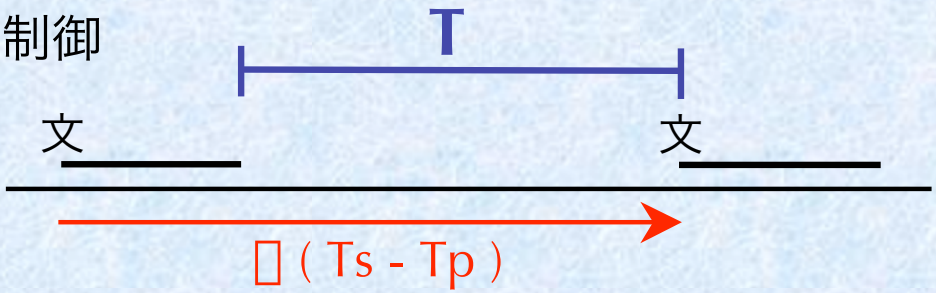
- ***聞き取れた単語だけを書き取って欲しい，邪推はして欲しくない，，，***
  - ✔ 被験者のタイピング能力に基づく音声提示間隔の制限
  - ✔ 文音声提示間隔（文尾～次文頭）T の制御

    T = α (Ts - Tp) - Ts

    Ts : 文音声長

    Tp : その文音声中のポーズ長

    α : タイピング能力係数（3.0～4.0，母語話者音声の聴取に基づき決定）

    文提示と同時に書き取りは開始される。よって，書き取り時間そのものは α (Ts - Tp)
  - ✔ 提示刺激音声の構成
    - ◆ 刺激時間間隔 T で連続提示，但し，3 文毎にタイピングエラーの修正を許可
    - ◆ エラー修正における時間制限は無し
  - ✔ 被験者
    - ◆ 日本在住暦 1 年未満の米国人 6 人（但し，日本語学習歴があるなどの親日家）
    - ◆ 峯松以外の日本人と話したことが無いカナダ人 1 人（ICSLP 会場でゲット，貴重なデータ）
  - ✔ タスク
    - ◆ Transcribe what you heard without deep guessing.
    - ◆ 各文書き取り後，その内容について「不確かなものがあるかどうか」 X or O で表記

書き取り実験手順

- *米国語話者音声に対する書き取り実験*
  - ✔ タイピング能力係数 α の決定
- *JE 用 360 文とは異なる文を使った，母語話者発声に対する書き取り実験*
  - ✔ 日本人英語書き取り性能に対する比較対照
- *同様の枠組みで日本人英語 720 文音声の書き取り*
  - ✔ Set-a, Set-b の順に提示
- *書き取り結果に対する後処理*
  - ✔ 単語尾（時制，単複数）の違いは無視した集計
  - ✔ 米国人 6 人（親日家）とカナダ人 1 人は別個に集計
  - ✔ 360 文中の約 2600 単語に対して，0/6 ～ 6/6 の書き取り率が定義（米国人結果）

予測対象

発音誤り分析 (#1)

- *音素生成における誤りの分析*
  - ✔ 日本語の音を余り知らない米語話者の聴取を模擬
    - ◆ 日本人英語をどのような米語音素列として書き取るのか？
    - ◆ 米語音響モデルのみを使った音素誤り分析
    - ◆ TIMIT-DB から，方言なまりの強い話者，単音間の linking が極端に強い話者を除いて音響モデル（multi-mixture monophone）を作成
  - ✔ Forced alignment 結果を用いた，音素誤り予測ネットワーク認識文法の作成
    - ◆ 日本人英語の特徴を反映
    - ◆ 音素コンテキストに依存した，置換，脱落，挿入，同一化（マージ）規則の導入
    - ◆ 日本語音素・米語音素間の音的類似性に基づく規則の導入
    - ◆ 母音に対応するスペルのローマ字読みに基づく規則の導入
    - ◆ などなど

thinks

```
T        $T_rep
I        $I_rep
G   →    ( $G_rep [ $vow_ins ] )
k        ( ( $k_rep [ $vow_ins ] ) | k_mrg )
s        ( $s_rep [ $vow_ins ] )
```

発音誤り分析 (#2)

- *文強勢生成における誤りの分析*
  - ネットワーク文法を通して得られる音素系列をシラブル化
    - tsylb v2.1 を使ったシラビフィケーション
  - 各シラブルを強勢・弱勢 HMM を用いて強弱判定（2段階）
    - 文強勢に注意して発声された文音声 700 音声を使った HMM 学習
    - シラブルの構造に基づき，シラブルを 12 分類 x 強・弱
    - スペクトル，パワー，ピッチ，有声度，継続長を考慮したモデリング
    - シラブル構造別に異なる HMM topology を使用
    - / schwa / として認識された母音は無条件で弱勢へ
    - Closed 評価実験における性能 96 %

```
Y y i T D i G k s D i T z u h a z I
  ➡  [ Y ][ y i T ][ D i G k s ][ D i T ][ z u ][ h a ][ z I ]
```

- *イントネーションは？　　今回は無視*
  - 着眼するのが語彙同定の可否（パラ言語情報の伝達ではない）
  - 殆どの文が肯定文。両言語間差異は，音素セット，リズムの方が遥かに大

発音誤り分析 (#3)

- *言語的属性値の推定*
  - ✔ 読み上げ文中の各単語，単語対に対して，1-gram，2-gram 値を抽出
    - ◆ 1-gram 値＝単語頻度＝その単語の親密度の粗い推定値（単語知覚への影響大）
  - ✔ WSJ コーパスより作成された言語モデルを使用

- *分析結果の例*

```
------- +1.000 +1.000 - silB [       0- 3600000]<-63.33> == silB [       0- 3600000]<-63.33>  silB    match -
iris    -1.645 -1.645 S    Y [ 3600000- 5800000]<-60.60> ==    Y [ 3600000- 5700000]<-60.33> Y_cor    match S
iris    -1.645 -1.645 -    r [ 5800000- 6100000]<-90.74> ==    y [ 5700000- 6200000]<-73.09> y_rep    match -
iris    -1.645 -1.645 W    I [ 6100000- 7200000]<-69.31> ==    i [ 6200000- 7200000]<-58.44> i_rep    match S
iris    -1.645 -1.645 -    s [ 7200000- 8000000]<-68.13> ==    T [ 7200000- 9300000]<-63.58> T_rep    match -
------- +1.000 +1.000 - null [ 8000000- 8000000]< +0.00> == null [ 9300000- 9300000]< +0.00>  null    match -
thinks  -4.292 -3.731 -    T [ 8000000- 9600000]<-64.39> ==    D [ 9300000- 9600000]<-72.58> D_rep    match -
thinks  -4.292 -3.731 S    I [ 9600000-10000000]<-71.58> ==    i [ 9600000-10600000]<-58.34> i_rep    match S
thinks  -4.292 -3.731 -    G [10000000-11300000]<-68.55> ==    G [10600000-11300000]<-76.30> G_cor    match -
thinks  -4.292 -3.731 -    k [11300000-12400000]<-79.76> ==    k [11300000-12400000]<-79.76> k_cor    match -
thinks  -4.292 -3.731 -    s [12400000-14300000]<-63.36> ==    s [12400000-14300000]<-63.36> s_cor    match -
------- +1.000 +1.000 -   sp [14300000-23700000]<-56.24> ==   sp [14300000-23700000]<-56.24>    sp    match -
this    -2.634 -1.884 -    D [23700000-24300000]<-86.28> ==    D [23700000-24300000]<-86.28> D_cor    match -
this    -2.634 -1.884 S    I [24300000-24900000]<-73.14> ==    i [24300000-24900000]<-68.32> i_rep    match S
this    -2.634 -1.884 -    s [24900000-25600000]<-67.22> ==    T [24900000-26400000]<-65.91> T_rep    match -
------- +1.000 +1.000 -   sp [25600000-25800000]<-68.10> == null [26400000-26400000]< +0.00>  null    match -
zoo     -5.510 -5.404 -    z [25800000-26400000]<-71.18> == null [26400000-26400000]< +0.00> z_mrg mismatch -
zoo     -5.510 -5.404 S    u [26400000-27300000]<-71.79> ==    u [26400000-27300000]<-71.79> u_cor    match W
------- +1.000 +1.000 - null [27300000-27300000]< +0.00> == null [27300000-27300000]< +0.00>  null    match -
has     -2.467 -1.747 -    h [27300000-28500000]<-66.45> ==    h [27300000-28200000]<-66.74> h_cor    match -
has     -2.467 -1.747 S    @ [28500000-29700000]<-64.88> ==    a [28200000-29700000]<-59.31> a_rep    match S
has     -2.467 -1.747 -    z [29700000-30800000]<-78.49> ==    z [29700000-30500000]<-71.87> z_cor    match -
has     -2.467 -1.747                                      I [30500000-30800000]<-93.01> I_ins    match W
------- +1.000 +1.000 -   sp [30800000-31100000]<-91.46> ==   sp [30800000-31100000]<-91.46>    sp    match -
```

CART による書き取り率の予測 (#1)

- *実験で使用した説明変数*
  - ✔ 分節的特徴に関する説明変数

segmental factors	level
#phonemes	P
#vowels	P
#consonants	P
#vowel replacements	P
distance vector of vowel rep.	P
#vowel insertions	P
#vowel deletions	P
#cons. rep.	P
distance vector of cons. rep.	P
#cons. insertions	P
#cons. deletions	P
#mismatches	P
word-level likelihood	W
phoneme-level likelihood	P
averaged likelihood	F

CART = Classification And Regression Tree

韻律的＆言語的特徴に関する説明変数

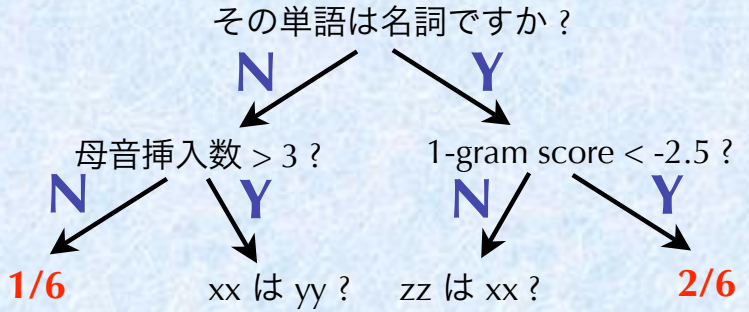
prosodic factors	level
#syllables	Sy
stressed syl. %correct	Sy
stressed syl. accuracy	Sy
#stressed syllables correctly produced	Sy
#rep. of stress with unstress	Sy
#rep. of unstress with stress	Sy
#inserted stressed syllables	Sy
#inserted unstressed syllables	Sy
word duration	W
averaged syllable duration	Sy
pause length before the word	W
pause length after the word	W
averaged stress-to-stress interval	S
variance of stress-to-stress intervals	S
linguistic factors	**level**
part of speech	W
position in the sentence	W
1-gram score	W
2-gram score	W

- ✔ 発音辞書において，与えられた単語の近隣に存在する単語数（Neighborhood サイズ，その単語が本来持つ他単語との音響的な混同し易さ）は未導入

CART による書き取り率の予測 (#2)

- *決定木の学習*
  - ✔ 2,100 語を学習，500 語を評価用データとして使用
  - ✔ 学習データの分布には「偏り」あり（ 6 / 6 データが半数以上）
    - ◆ n / 6 (n < 6) データを複数回カウントし，疑似的に偏りをなくした学習も検討
  - ✔ 説明変数の異なる 5 つの実験条件

CASE-1	only with segmental factors
CASE-2	only with prosodic factors
CASE-3	only with linguistic factors
CASE-4	only with acoustic factors
CASE-5	with all the factors



決定木の一例（推定されたものとは異なる）

- *評価尺度*
  - ✔ +1/6, -1/6 の予測エラーは無視して，recall 及び precision を算出
    - ◆ 2/6 データに対する recall
      - ☞ [その中で正しく 2(±1)/6 と判定された数] / [2/6 が正解である評価データ数]
    - ◆ 2/6 データに対する precision
      - ☞ [その中で 2(±1)/6 が正解であるデータ数] / [2/6 と判定された評価データ数]

実験結果と考察 (#1)

❒ *米国語話者の書き取り能力*

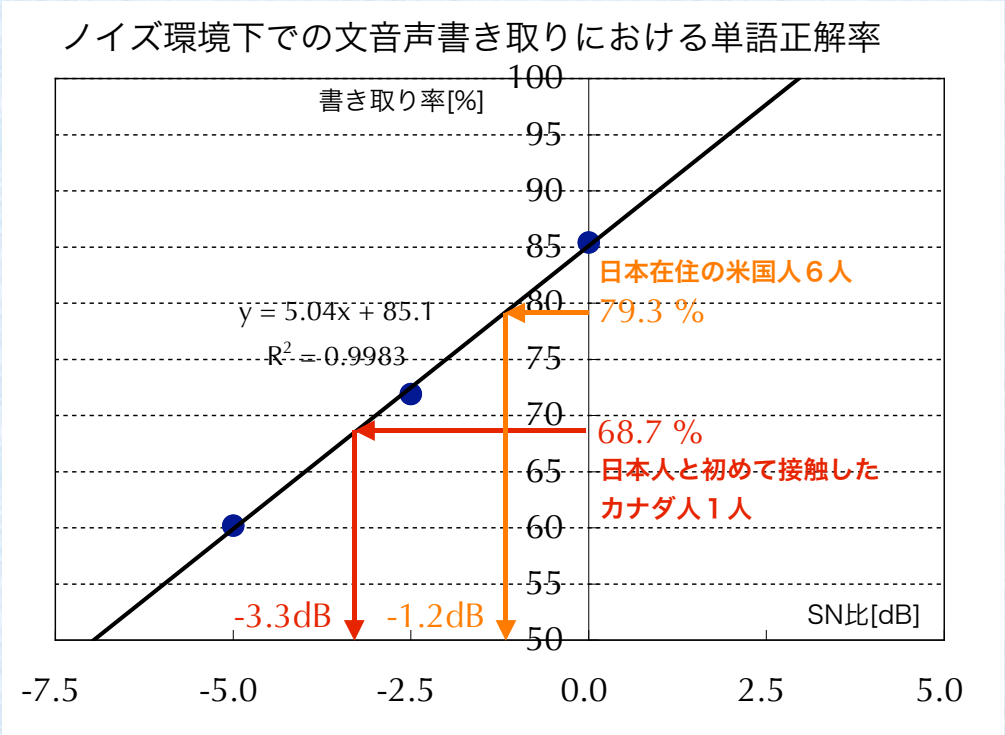
prof. level	#spk.	#uttr.	%correct	rate of X
~2	2	16	64.1%	83.3%
~2.5	27	216	75.4%	56.7%
~3	38	304	82.3%	44.7%
~3.5	21	168	83.4%	33.7%
~4	2	16	91.3%	20.8%

（米国人 4 人，set-a, bの結果）

- ✔ 平均書き取り率
  - ◆ Set-a　[米国人 6 人] = 79.3 %,　[カナダ人 1 人] = 68.7 %
  - ◆ Set-b　[米国人 6 人] = 84.3 %,　[カナダ人 1 人] = 67.2 %
  - ◆ Set-a / b は同一文セット，音響的学習以外に言語的な学習効果もあり？
  - ◆ 国際会議発表において PC / プロジェクタが壊れた場合，**あなたの**英語は（その研究テーマにあまり通じていない聴取者には）このくらいしか聞き取ってもらえていない？
- ✔ ~3と~3.5 は単語正解率には大きな差は見られないが，理解率には有意な差
  - ◆ 語彙の伝搬能力ではなく，意図の伝搬能力（コミュニケーション能力）の差

実験結果と考察 (#2)

- ***平均書き取り率 70〜80 %って，どんな音声？***
  - ✔ 日本語音声にノイズを重畳することで，単語正解率 70 〜 80 % の音声を生成
  - ✔ ATR 503 文から，10 単語以下の文を 30 文選択し 18 名の被験者を使って調査



  - ✔ 単語明瞭度 79.3 % = SN比 -1.2 dB，　68.7 % = SN比 -3.3 dB
    - ◆ 8 年間の英語学習において身に付いた音声による語彙伝搬能力
    - ◆ 現在の英語教育が有する発音指導に関するパフォーマンス？
    - ◆ 教育の**「支援」**が必要なのではなく，必要なのは教育の**「変革」**
    - ◆ その変革のために音声技術 / 計算機 / ネットワーク / マルチメディアは何をもたらすのか？

実験結果と考察 (#3)

- *CART による書き取り率の予測 （+1/6, -1/6 のエラーは無視）*

予測正解率 [%]

		0/6	1/6	2/6	3/6	4/6	5/6	6/6
#test data		24	21	27	31	45	92	292
C-1	recall	29	43	53	19	38	94	95
	prec.	58	57	61	50	68	90	79
C-2	recall	44	35	54	36	37	93	94
	prec.	55	16	47	64	82	82	83
C-3	recall	29	43	37	61	27	42	32
	prec.	13	10	14	26	36	84	68
C-4	recall	54	34	42	29	53	85	92
	prec.	50	46	48	43	72	80	86
C-5	recall	50	34	48	64	57	89	96
	prec.	55	42	50	53	66	88	83
B·L	recall	29	43	43	43	43	43	29
	prec.	8	14	15	20	32	81	72

CASE-1	only with segmental factors
CASE-2	only with prosodic factors
CASE-3	only with linguistic factors
CASE-4	only with acoustic factors
CASE-5	with all the factors

実験結果と考察 (#4)

- ***CART による書き取り率の予測***
  - ✔ Baseline = チャンスレベル
    - ◆ +1/6, -1/6 の予測誤りを無視
    - ◆ Precision 計算時には，学習データが持つ書き取り率の偏りを（事前分布として）考慮
  - ✔ 予測最高性能は CASE-5 （分 / 韻 / 言全てを考慮）で実現
  - ✔ 学習データが持つ「偏り」
    - ◆ 通常の CART 学習では，6 / 6 から 0 / 6 へ向けて精度が落ちる
    - ◆ 「偏り」を人工的に排除した学習
      - ☞ Recall = 100, 100, 99, 98, 95, 87, 62 % for 6 / 6 to 0 /6.
      - ☞ 最終的に欲しい性能（の特性）は応用場面に依存。決定木学習形態の制御で対応可
  - ✔ 予測に有効な説明変数は？
    1. **Variance of stress-to-stress intervals**
    2. **1-gram score**
    3. **Phoneme-level likelihood**
    4. ：
    5. ：

リズム感よく
易しい単語を
個々の音を正しく

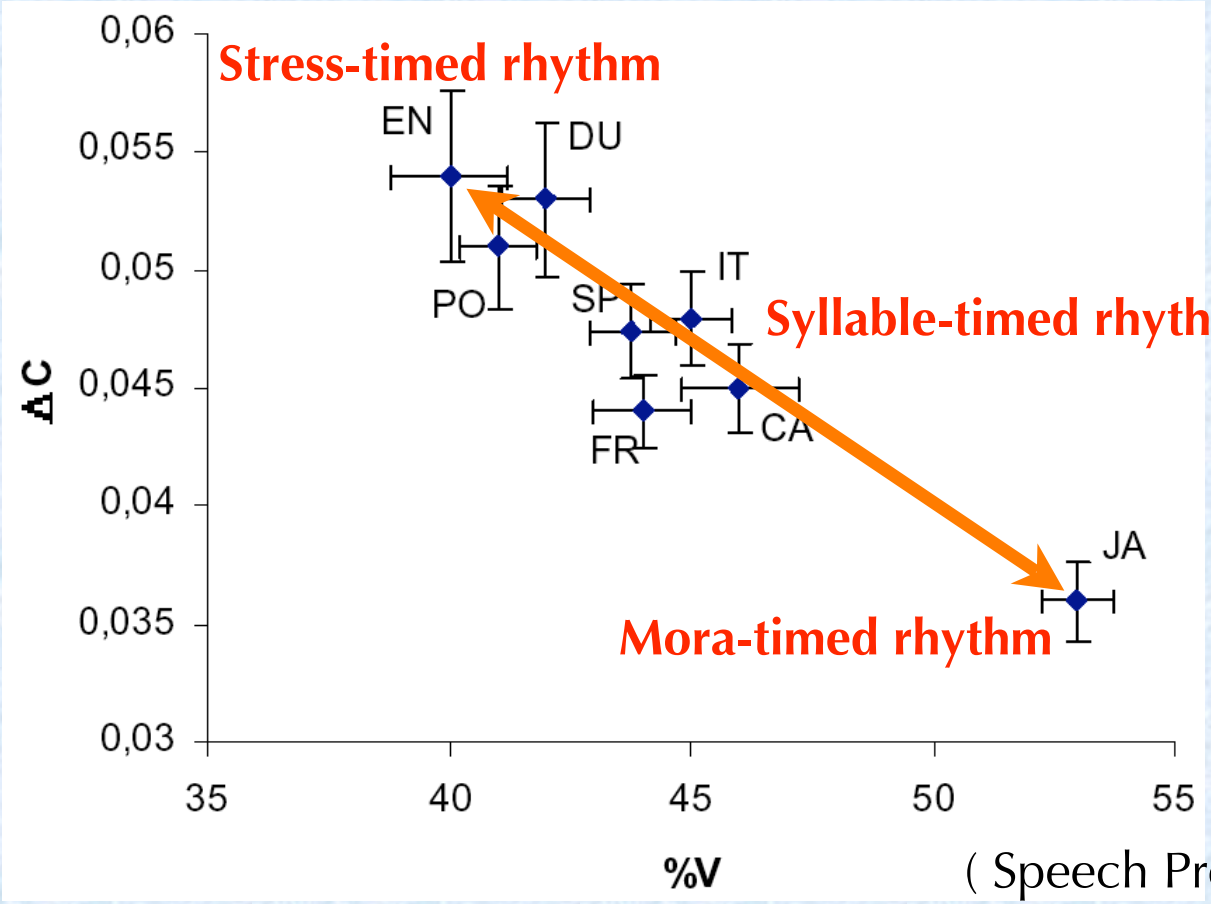
見ているのは「通じるか，通じないか」
でもリズムが狂っていると，通じない‥

何でリズムなの‥‥？ 1つの仮説

- *「ことば」の本質って「話しことば」？「書きことば」？*
  - ✔ 「話しことば」でしょう。だって，誰もが先に獲得するじゃない。
- *「話しことば」の本質って「分節」「リズム」「イントネーション」？*
  - ✔ あんまり縦割りすべきじゃないけど，獲得過程を考えれば「分節」は最後だねぇ。
- *だったら「リズム」が本質になるのかなぁ？*
  - ✔ だって胎内にいる時から，ノイズに包まれた心音を聞いてるじゃない。
  - ✔ 半年は聞いてるって聞くけど。あるのは暗闇とノイズとリズムだけ。
  - ✔ 規則正しいリズム → 日本語のモーラ・フットの知覚的等時性？
    英語の強勢シラブルの知覚的等時性？
- *と考えると，リズムが崩れると，まあ，聞きたかぁないわな，，，*
  - ✔ かもね
- *従来の発音教育 ＝ 音声学だけに根差す方法論 ＝ 聞き手不在の方法論？*
  - ✔ リズムが異なる言語を習得する場合，それ専用の方法論があって然るべき

日本語リズムって英語とそんなに違うの？

- *リズム研究が示す一枚のグラフ*



（Speech Prosody'2002 より）

- ✔ %V = 文音声において母音が占める時間的割合
- ✔ ΔC = 連続する子音によって構成される区間の一文内での標準偏差

まとめと今後の課題

- ***Native-like* から *intelligible* へ**
  - ✔ 聞き手の存在（適応能力）を意識した教育方法
  - ✔ 母語話者の知覚特性が必要となる教育方法（日本人教師には不可能？）
  - ✔ だったら，それを計算機でシミュレートできないのか？
- ***母語話者による日本人英語聴取誤りの分節的・韻律的・言語的分析***
  - ✔ JE-DB からバランス良く選択された 720 音声の聴取＆書き取り
  - ✔ 日本人であること＝ -3.3 dB のホワイトノイズを付加することに相当
  - ✔ 聴取音声に対する分節，韻律，言語的側面からの（発音誤り）分析
  - ✔ CART による誤聴取単語（書き取り率）の予測
  - ✔ 予測に貢献する要因＝1. リズム，2. 親密度，3. 単音の正しさ，，，，
- ***何故リズム？***
  - ✔ 誰しもが最初に獲得する「ことば」の側面 ＝ リズム ？
- ***CART 以外の規則生成の検討 ／ 有効な他説明変数の追及***
- ***日本人の英語教師はどのくらい intelligibility を予測できるのか？***
  - ✔ 新しい教育戦略に対する準備はできているのか？

!!!!!! お 知 ら せ !!!!!!

ご覧戴いたように，本研究では「日本人英語中の各単語が母語話者によってどのくらいの確率で聞き取られるのか」を現在の音声・言語技術を用いて予測する，という研究を行っています。「伝わる英語」を教育目標とする場合「やっていいこと，いけないこと」の定義が必要ですが，母語話者の知覚プロセスをシミュレートする形でそれを実装しよう，というものです。

まだまだ予測精度が低いのはご覧の通りですが，日本人の英語教師がどの程度「伝わる，伝わらない」を正しく評価できるのか，について参考データとして収集したい，と思っております。

具体的な作業としては，JE-DB からバランス良く選択選択した 360 発声に対して，それを聴取して戴き，意図された文における各単語に 0/6 ～ 6/6 のスコアを付けて戴く，という流れを予定しております。

協力して戴ける場合は，是非ともご連絡下さい。

連絡先：**mine@gavo.t.u-tokyo.ac.jp**

（なお峯松は今，長期出張で Fant 先生のお膝元，Sweden, KTH にいます）

峯松は「人間にできること・できないこと，計算機にできること・できないこと，を総合的に考えて教育戦略を練るべきだ」と考えております。