

峯松 信明 (東大・情報理工)

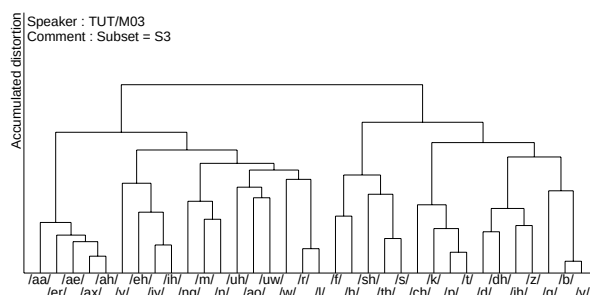
音声コミュニケーションでは、その生成、収録、伝送、再生、聴取の何れの過程においても乗算性或いは線形変換性の歪みが混入する。この不可避的な歪みを表現する次元を保有しない音声の物理表象が提案されている [1]。この表象は音韻論における言語音群の構造記述を、確率論、情報論、及び相対論に基づいて物理の上で実装することで得られる（音響的普遍構造）。この音声表象に基づいて語学学習者を記述すると、そこには性別、年齢、体格、収録機器特性などが一切消え、主に、外国語発音における母国語依存性のみが観測される [1]。本研究では、学習者が呈する音響的普遍構造と、対象言語が呈する言語的語彙構造との整合性に着眼し、発音の分節的明瞭度を定義する。そして、学習者音声を教師音声と音響的に比較するのではなく、学習者音声を対象言語そのものと比較することによる発音評定が可能であることを示す。

ある言語に含まれる各音素を、ケプストラムベクトルによって構成される多次元ガウス分布であると仮定し、音素間距離をバタチャリヤ距離で算出する。

バタチャリヤ距離は二つの確率密度に対して両事象の独立性を仮定して同時確率密度を求め、その平方根を全積分して同時確率へ変換し、その対数で定義される。即ち、同時確率に対する自己情報量を距離と考える。 n 音素に対して、 ${}_nC_2$ 個存在する全音素間距離を考える。これは n 角形の全対角線を考えることに相当し、 n 点で張られる構造を考えることに等しい。

$$c'_t = Ac_t + b$$
$$\mu' = E(c'_*) = A\mu + b$$

$$\Sigma' = E(c'_+ - \mu')(c'_+ - \mu')^T = A \Sigma A^T$$

$$\begin{aligned} & BD(\mu'_u, \Sigma'_u, \mu'_v, \Sigma'_v) \\ &= BD(A\mu_u + b, A\Sigma_u A^T, A\mu_v + b, A\Sigma_v A^T) \\ &= BD(\mu_u, \Sigma_u, \mu_v, \Sigma_v) \end{aligned}$$


これは、情報論的に算出された状態間距離は一次変換に対して不変であり、構造として表象された音声事象は一次変換に対して不変であることを意味する。

ユークリッド空間における構造（点の集合）に対する一次変換は、構造に対して種々の変形を施す。A は構造のスケーリング、せん断、回転及びそれらの組み合わせを、b は構造のシフトを意味する。ここで点が平均ベクトルであり、そこに分布が付随している状況を考える。上記した種々の変形の中で、変形前後で二点（分布）間距離が不変であるのは回転とシフトのみである。つまり、A（例えば声道長の違いがもたらす音響的歪み）は構造の回転となり、b（例えば収録機器の違いがもたらす音響的歪み）は構造のシフトとして観測され、構造は不変である。さて、語学学習者の音声データから本表象を抽出すると、主に母国語発音への依存性に起因する構造歪みが呈されることとなり、これは学習者の今を記述する「文字」が生まれたことに相当する。図 1 に一例を示す。日本人英語特有の癖が明確に表現されている。

3.1 ネイティブ神話からの脱却

日本における英語教育では、かつて「ネイティブ神話」と呼ばれた発音教育指針があった。近年国際語としての英語が認知され、「通じる・明瞭な発音」が求められるようになった。しかし筆者の知る限り「ネイティブ神話」に基づかない発音 CALL システムは存在しない。それは全てのシステムが、母語話者発音の音響音声学的特性と学習者発音のそれとを比較することを大前提としているからである。その結果、音声処理系特有の不安定性を回避できず、また「いつまでネイティブ神話を唱えるのか」との批判にも曝される。

本研究では、この大前提に基づかない方法論を提案する。そこには、不安定性も、ネイティブ神話も無い。前節で述べた学習者を記述する「文字」は、同一話者・同一収録環境下での異なる音素間の距離だけを観測することで得られる。即ち音声認識の世界で忌み

*Automatic scoring of pronunciation based upon the acoustic universal structure in speech

嫌われる mismatch が起こり得ない。故に不安定になり得ない。子供、高齢者を対象とすることで処理精度は落ち得ない。この「文字」の上でシステム構築をすれば、音響的には、安定なシステムしか作り得ない。

3.2 コホートモデルに基づく語彙密度の算出

「ネイティブ神話」から如何に脱却するのか？「発音の明瞭度」を如何に定義するのか？日本人英語の場合音素混同が頻発する。音声知覚過程をコホートモデルとして考えた場合、コホートサイズ（活性化語彙、CS）が増加する。このCSを発音構造と語彙構造を用いて算出し、発音の「分節的明瞭度」と定義する。

本研究では、シラブルを単位としたコホートを考え、第一シラブル入力時のCSを推定する。語彙セットとしては、語彙数20KのWSJのunigramを使用した。この辞書の全エントリをtsylb[2]を用いてシラブルに分割した。各異なりシラブルについて $CS_0(s_i, \theta)$ を求める。 $CS_0(s_i, \theta)$ とは、シラブル s_i 或は s_i から音響的に近い（距離 θ 以下）シラブルを語頭に持つ単語数の総和である。次に $CS_1(w_j, \theta)$ を求める。 $CS_1(w_j, \theta)$ とは w_j の先頭シラブルを $s^1(w_j)$ とした時 $CS_0(s^1(w_j), \theta)$ である。つまり w_j の先頭シラブル或はそのシラブルから音響的に近いシラブルを先頭とする単語数である。最終的にCSの全単語に対する期待値、ECS(Expected Cohort Size)を次式で求める。

$$ECS(\theta) = \sum_j p(w_j) CS_1(w_j, \theta)$$

$p(w_j)$ は w_j のunigram値である。こうして、単語の生起頻度まで考慮したCSが、シラブル間距離閾値 θ の関数として定義される。任意の2シラブル間の距離は、音素間距離行列を参照し、シラブル（音素連鎖）間のDPにより計算できる。即ち、本研究で検討する音響的普遍構造と言語的語彙構造の整合性の定量化は、音声を構造として表象する際に得られるパラメータ（音素間距離行列）だけが必要となる。

3.3 実験結果と考察

ERJデータベース[3]の各話者に対して、GMMで音素をモデル化し、音素間距離行列を算出した。一話者当り60文（音素バランス文の一部、話者毎に異なる）を用いた。一部の話者のデータには、二重母音が存在しなかったため、二重母音以外の音素を対象とした。CS推定も二重母音で始まる単語は無視した。

図2にシラブル間距離 θ の関数としてECSを示す。発音構造は文セットに依存する傾向があるので、ここでは、ERJの発音評定ラベルより幅広い発音習熟度の話者が得られた、文セット6の結果を示す。図中、スコアが5点満点の話者は母語話者（4人）である。図右側には、日本人学習者の中で英語教師の採点の上位3人、下位3人のスコアを示している。成績下位者のCSは非常に大きい。成績上位者の一部が母語話者よりもCSが小さく見積もられている。音素間距離は、話速や調音努力などの影響を受ける。例えば、

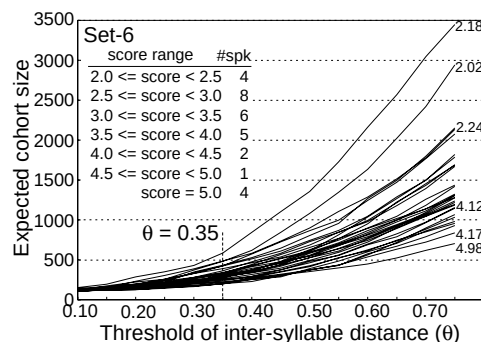


図2. シラブル間距離 θ の関数として表現したECS

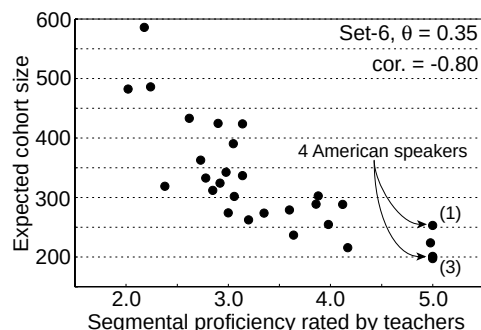


図3. 自動発音評定結果

母語話者であっても話速が早い場合は音素間距離が小さくなりCSは増加する。収録時に発声速度・様式を統制するか、事後的に正規化する処理が必要である。

図2に対して、 $\theta = 0.35$ においてECSを求め、それと英語教師による評定との関係を図3に示す。なお母語話者は5点満点としている。相関係数は -0.80 であり、比較的良好な対応がとれている。注意して戴きたいのは、本手法では母語話者音響モデルとの照合のみならず、母語話者音声データそのものを一切用いていない点である。母語話者の発音構造が得られた場合、構造の一部を学習者と置換し、コホートが効率的に低減する部分置換を求めることで、学習者毎に異なる発音学習計画を立てることも可能である[4]。

4 まとめ

音響的普遍構造として抽出される学習者の発音構造と、対象言語の言語的語彙構造との整合性に着眼して発音の分節的明瞭度を定義し、実験的にその妥当性を示した。今後は幅広い年齢層の音声に対する検証、及び実際の教育の場への導入などを考えている。

参考文献

- [1] 峯松, “音声に内在する音響的普遍構造とそれに基づく語学学習者モデリング”, 電子情報通信学会音声研究会資料, SP2003-179 (2004)
- [2] Tsylb: <http://www.nist.gov/speech/tools/tsylb2-11tar.Z.htm>
- [3] 峯松他, “英語CALL構築を目的とした日本人及び米国人による読み上げ英語音声データベースの構築”, 日本教育工学会論文誌 (2004, 採録決定)
- [4] 峯松, “音響的普遍構造と言語的普遍構造間の整合性に基づく発音明瞭度の評定”, 電子情報通信学会音声研究会資料, SP2003-181 (2004)