社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－ＳＬＰ－69（27）
2007／12／20

英語学習者発話の自動評定における正規化の検討

辻 晃佑† 山下 洋一‡

†立命館大学理工学研究科 〒525-8577 滋賀県草津市野路東 1-1-1
‡立命館大学メディア情報学科 〒525-8577 滋賀県草津市野路東 1-1-1
E-mail: †rs028029@se.ritsumei.ac.jp, ‡yama@media.ritsumei.ac.jp

あらまし 本稿では，英語学習者の発音自動評定におけるスコアの正規化手法について述べる.学習者の発音を自動評定するために，音声認識により得られる認識尤度を用いた手法が広く用いられており，尤度を事後確率化することによって発音評定の精度が向上することが知られている.全音素モデルによる認識尤度によって事後確率化のための特徴ベクトル生起確率を近似する手法において，発話全体の平均のかわりに音素単位での尤度を利用する手法を提案する.また，母語話者が発声する場合にも，音素ごとに尤度のばらつきがあることに注目し，母語話者発声における音素ごとの認識平均尤度を用いて学習者に対する評定スコアを正規化する手法を提案する.さらに，母語話者発声における音素の認識平均尤度にはコンテキスト依存性があることから，音素ごとの認識平均尤度をコンテキストに分けて利用する手法についても述べる.人間教師の評定結果との相関において，提案手法によって相関係数が約 0.1 向上することを評価実験によって示している.

キーワード 事後確率，母語話者発話，英語， CALL，発音

A Study on Normalization for Automatic Scoring of Pronunciation Proficiency of English Learners

Kosuke Tsuji† Yoichi Yamashita‡

† Graduate school of Science and Engineering, Ritsumeikan University

1-1-1 Nojihigashi, Kusatsu-shi, Shiga, 525-8577 Japan
‡ College of Information Science and Engineering, Ritsumeikan University

1-1-1 Nojihigashi, Kusatsu-shi, Shiga, 525-8577 Japan
E-mail: †rs028029@se.ritsumei.ac.jp, ‡yama@media.ritsumei.ac.jp

Abstract This paper describes normalization methods for automatic scoring of pronunciation proficiency of English learners. Acoustic likelihoods calculated by speech recognition techniques are widely used for automatic scoring of pronunciation proficiency of language learners. It is well known that the normalization of acoustic likelihoods with posterior probabilities improves the accuracy of automatic scoring of pronunciation proficiency. The acoustic likelihood with an all-phone model which allows all transitions between phonemes approximate the probability of observed speech signals to calculate the posterior probability. This paper proposes acoustic likelihoods for each phoneme in stead of the average acoustic likelihood of the entire utterance using the all-phone model. Furthermore, an average of the acoustic likelihood is dependent on the phoneme even if a native speaker utters. This paper proposes the introduction of the average acoustic likelihood by native speakers to normalize pronunciation proficiency scores. The context dependency of the average acoustic likelihood by native speakers is also discussed. Evaluation experiments show that the proposed methods improve the correlation coefficient between estimated proficiency scores and human teachers' scores by about 0.1.

Keyword Posterior probability, Native speaker utterance , English, CALL, Pronunciation

1. はじめに

日本人は英語を話すことが苦手である.その原因の一つに,日本の英語教育が読み書き中心であることが挙げられる.近年コンピュータを用いた語学教育（CALL）が注目されており,スピーキングやリスニングに重点をおいたものも多く研究されている[1].スピーキングの上達には教師を伴った発音練習が効果的であるが,時間や場所等の制約から教師による指導を受けられないことも多い.そこで人間教師と同等の評価を計算機が行えるようになれば,学習者の都合に合わせていつでもどこでも効果的な発音練習ができるようになる.そこで音声認識によって得られる認識尤度を利用することにより,学習者の発音評定を行う手法が広く研究されている[2][3][4][5].一般に,正しい発音の音素列に対する尤度を算出することにより評定スコアを求めることができる.この尤度に対して発話時間長や事後確率による正規化を行うことにより発音評定の精度が向上することが報告されている[2][5].同様に,母語話者における音素の平均尤度を用いた正規化が有効であると先行研究により報告されている[3].本報告では,事後確率による認識尤度の正規化の見直しを行い,特徴ベクトルの生起確率を音素ごとに算出することによって手法の改善を試みる.さらに母語話者における音素の平均尤度を用いた正規化を提案し,その音素コンテキスト依存性を考慮した平均尤度の利用についても検討する.

2. 事後確率による認識尤度の正規化

2.1. 音声認識尤度に基づいた発音評定

特徴ベクトル時系列によって表現された観測音声波形をy,音素系列で表現された文（あるいは単語）をwとすると,音声認識の問題は一般に,

$$P(w|y)=\frac{P(y|w)P(w)}{P(y)} \qquad (1)$$

を最大にするwを決定する問題として定式化される.音声認識では，式(1)の値を最大にするwを求めることが目的となるため，右辺においてwに依存しない$P(y)$を無視し，分子の値だけを算出する.発話内容が既知である文や単語の発音評定を音声認識尤度に基づいて行う場合には，wが与えられた条件で式(1)の値を算出する必要がある.

音声認識では不要として扱われる$P(y)$の値を算出するために，任意の音素連結を許容する音素遷移モデル（いわゆる全音素モデル）を言語モデルとして音響尤度を算出し，その値で$P(y)$を近似する手法が用いられる.

2.2. 手法 1:文章単位での事後確率化

式(1)右辺の値を算出し，発音の評定値とする.ここで，発話内容が既知であることから，その発話内容に対応した正しい音素系列をwとし，$P(y|w)$を算出する.$P(w)$に関しては，どのwに対しても$P(w)$ = 1 とした.$P(y)$は，図 1 に示す全音素モデルでの音響尤度で近似する.用いた音素は英語音素 aa, ae, ah, ao, aw, ax, ay, i, iy, uh, uw, eh, er, ey, ow, oy の母音 16 個と b, ch, d, dh, f, g, h, jh, k, l, m, n, ng, p, r, s, sh, t, th, v, w, y, z, zh の子音 24 個にショートポーズ sp と文頭および文末無音 sil を加えた計 42 個である.

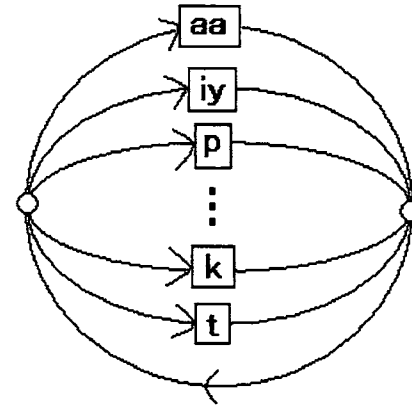

図 1 全音素モデル

正しい音素列による音声認識結果においてi番目の音素p_iのフレーム数をN_i,認識尤度をL_iとする.同じく全音素言語モデルでの音声認識結果においてj番目の音素pa_jのフレーム数をNa_j,認識尤度をLa_jとして,手法 1 での評定値$S1$を式(2)で定義する.

$$S1=\frac{1}{I}\sum_{1}^{I}\left(\frac{L_i}{N_i}-AveLa\right). \qquad (2a)$$

ただし,

$$AveLa = \frac{\sum_{1}^{J} La_j}{\sum_{1}^{J} Na_j} \qquad (2b)$$

であり,I は入力発話における音素数である.

ここで $AveLa$ は,発話に対する全音素モデルでの認識結果において,1 フレームあたりの認識尤度を表し,1 発話中で一定の値をとる.同じく,L_i/N_i は正しい音素列による認識音素 p_i の 1 フレームあたりの認識尤度を表す.音響尤度の算出の際に対数をとっているため,L_i/N_i から $AveLa$ を引くことで事後確率化による正規化をフレームレベルで行っている.

2.3. 手法 2:音素単位での事後確率化

手法 1 では,$P(y|w)P(w)$ が音素ごとに求められるのに対して $P(y)$ は発話全体から求められるため,認識音素範囲ごとの生起確率が正しく求められていないと思われる.そのため,本研究では事後確率化を発話全体で行うのではなく部分的に行う方法を検討した.すなわち,文章全体に対する $P(y)$ を求めるのではなく,音素ごとに $P(y)$ を求め音素単位の事後確率化を行った.p_i の音素時間範囲に対応した全音素遷移可能な条件下での音声認識結果の認識尤度を R_i として,手法 2 での評定値 $S2$ を式(3)で定義する.

$$S2 = \frac{1}{I}\sum_{1}^{I}\left(\frac{L_i}{N_i} - \frac{R_i}{N_i}\right). \qquad (3)$$

ただし L_i に対応した時間範囲において,全音素モデルでの音声認識結果は同じ音素時間幅,音素数になるとは限らないため,R_i はその時間範囲に含まれる全ての音素の認識尤度を,それぞれのフレーム長の割合で掛け,それらを足し合わせることで求めている.音素単位での $P(y)$ の求め方について図 2 に示す.ここで t_i および ta_j はそれぞれ p_i および pa_j の開始時刻を表す.

この図では,正しい音素列による認識音素 p_i の音素区間と全音素モデルでの認識結果を照らし合わせると,複数の音素の区間にまたがっており音素境界も一致していない.

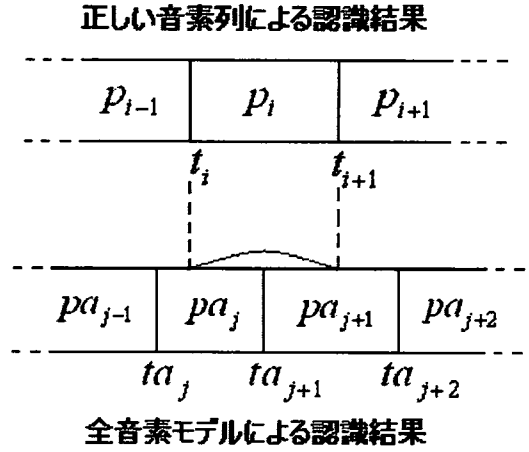

図 2 各音素範囲における $P(y)$ の求め方

この例における $P(y)$ は

$$\frac{R_i}{N_i} = \frac{\left(\frac{ta_{j+1}-t_i}{ta_{j+1}-ta_j}La_j + \frac{t_{i+1}-ta_{j+1}}{ta_{j+2}-ta_{j+1}}La_{j+1}\right)}{t_{i+1}-t_i} \qquad (4)$$

のように求めることができる.p_i の区間が pa_j の区間内にすべて含まれる場合など,他のパターンも在るが,式(4)はそのパターンに対応して変更される.

以降は事後確率化に手法 2 を用いる.

3. 母語話者の音素平均尤度を利用した正規化

3.1. 手法 3:コンテキスト独立母語話者平均音素尤度による正規化

母語話者が発話して尤度が低い音素や高い音素があるとするならば,日本人学習者への評価に加味する必要があると考えられる.そこで,TIMIT データベースの母語話者音声を用いて音素の尤度のばらつきを調べた.音素ごとの認識尤度の平均と分散の一部を表 1 に示す.ただし,学習者発話は手法 2 と同じく事後確率化しているため,母語話者発話も同様に事後確率化を行っている.

表 1 母語話者音声における各音素の認識尤度
(音素コンテキスト独立)

音素	データ数	認識尤度 平均	分散
t	10733	-2.6	13
i	18347	-1.6	7.5
n	11874	-1.2	7.1
…	…	…	…
aa	4197	-0.3	2.1
ch	1081	-0.3	1.0
s	10114	-0.1	0.9

その結果を見ると,スコアが高くなり易い音素や低くなり易い音素,ばらつきの大きい音素や小さい音素があることが分かる.このことから,母語話者の音素平均尤度のばらつきを考慮して学習者発話の認識尤度の正規化を行う[1].音素 p_i における母語話者の認識尤度の平均を $AVE(p)$,分散を $VAR(p)$,標準偏差を $SD(p)$ として,手法 3 での評定値 $S3$ を式(5)で定義する.

$$S3 = \frac{1}{I}\sum_{1}^{I}\frac{\frac{L_i}{N_i} - \frac{R_i}{N_i} - AVE(p_i)}{SD(p_i)} \quad (5a)$$

ただし,

$$SD(p_i) = \sqrt{VAR(p_i)} \quad (5b)$$

である.

3.2. 手法 4:コンテキスト依存の母語話者平均音素尤度による正規化

音素は前後の音素により影響を受けることが一般的に知られている.つまり同じ音素であっても前後の音素の影響を受けることで異なった性質を持つと考えられる.3.1 で述べた母語話者による音素ごとの平均尤度における直前の音素とのコンテキスト依存性を分析した.認識尤度の平均と分散の一部を表 2 に示す.

この結果を見ると,同じ音素であっても直前の音素が異なれば,スコアの高くなり易さや低くなり易さ,ばらつきの大きさに差があることが分かる.このことから,母語話者音素を用いた尤度の正規化手法に音素のコンテキスト依存性を考慮した手法を検討する.

表 2 母語話者音声における各音素の認識尤度
(音素コンテキスト依存)

直前の音素	音素	データ数	認識尤度 平均	分散
…	…	…	…	…
ay	i	140	-3.89	13.88
l	i	910	-2.55	10.98
ch	i	159	-2.3	14.8
…	…	…	…	…
f	i	365	-1.33	4.67
z	i	709	-1.28	4.48
eh	i	3	0.75	1.64
…	…	…	…	…

式は手法 3 と同様に式(4)を用い評定値を算出する.$AVE(p_i)$, $SD(p_i)$ にそれぞれコンテキスト依存の音素ごとの平均,分散を用い,その評定値を $S4$ とする.

ただし,音素を細分化しすぎると各音素のばらつきを十分に表現するためのデータ数が確保できなくなる.そのため,今回は依存性を考慮する部分を制限した.具体的には音素を大きく母音グループと子音グループに分類し,母音グループの場合には直前の音素から,子音グループの場合には直後の音素から影響を受けると想定して音素分類を行った.それを図 3 に示す.

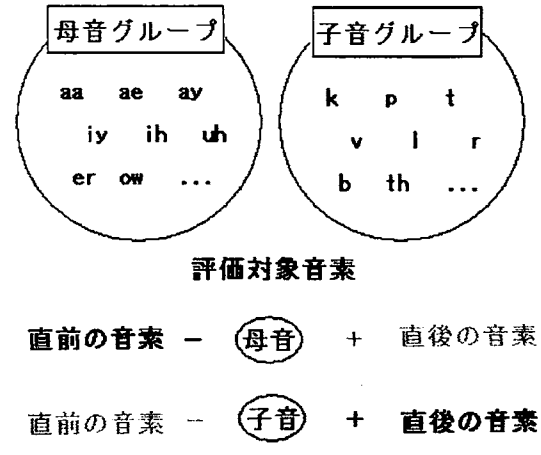

図 3 部分的なコンテキスト依存性の利用

また,十分に分散を表すのに必要な最小データ数の閾値を経験的に 20 個と定め,データ数がこれに満たない場合ではコンテキスト独立での音素の平均と分散で代用した.

4. 音声データ

英語の音響モデルは,TIMIT データベースの音声試料(462 名,各 10 文)を学習データとしてHTK を用いて作成されたものを使用した.また,認識エンジンには HTK を用いた.評価データには日本人による英語読み上げ音声データベース[6]から 895 文を使用した.このデータベースでは,文章の各学習者発話に対してイントネーション,リズム,発音を評価した人間英語教師による評価値が付与されている[7].本研究では発音に注目した評価値を用い,それぞれの手法による評定値との相関を調べた.ただし教師は 5 名,教師評定値は 5 段階(1:下手,5:上手)であり,教師ごとに評定値を正規化し 5 人の平均値を教師評定値とする.

参考までに,人間教師の評価における相関を表 3 に示す.表 3 では,ある教師による評価とその教師を除いた 4 人の教師による評価の平均値との相関を表しており,最後に教師の平均値を示している.

表 3:教師評価における相関

	相関係数
教師 A	0.660
教師 B	0.650
教師 C	0.630
教師 D	0.661
教師 E	0.619
平均	0.644

5. 実験結果

人間英語教師による評定値と提案手法を含む4つの手法で求めた評定値との相関を表4に示す.事後確率化を文章単位で行った結果と比較して,音素単位で行った結果では大きく相関係数が向上している.またすでに報告されているように[1],母語話者の音素平均尤度を用いて正規化を行うことによって相関係数が高くなっている.それに加え,母語話者の音素平均尤度にコンテキスト依存性を考慮することによってさらに相関係数が高くなっている.しかしながら人間教師による相互の相関は表 3 に示すように約 0.644 であることから,自動評定の手法をさらに改善する必要があると思われる.

表 4:各手法のスコアと教師の評価との相関

	正規化手法				相関係数
	事後確率化		平均尤度	平均尤度	
	文章	音素	コンテキスト独立	コンテキスト依存	
手法 1	○				0.368
手法 2		○			0.447
手法 3		○	○		0.461
手法 4		○		○	0.473

6. おわりに

本稿では人間教師による発音評定値と高い相関を持つ発音自動評定手法の実現を目的として,事後確率による正規化を部分的に行う手法と母語話者の音素平均尤度を用いた発音評定に音素コンテキスト依存を考慮した手法の 2 つの正規化について検討した.その結果として,前者では文章単位で事後確率化を行うより音素単位で行う方が人間教師評価との相関が上昇することが分かり,後者ではコンテキスト依存での母語話者の音素平均尤度による正規化が評価性能の向上に有効であることを示すことができた.

今後の課題としては,コンテキスト依存性の部分的な利用についてさらに効果的な組み合わせを検討することが挙げられる.今後は人間教師による評価の視点を取り入れた評定手法を検討していく予定である.現在は機能語や内容語といった英語の性質,音素の発話速度などの要素を重みとして評定手法に組み入れ,本稿で提案した手法とも組み合わせていく.

文　献

[1] 中川聖一: “科学研究費特定研究(A)『メディア教育利用』－音声言語処理技術を用いた語学 CAI－”, 日本音響学会誌, Vol.56 No.11, pp.767-770,2000.

[2] 大田圭, 中川聖一: “日本人の英語文発話の発音評価法”, 日本音響学会春季講演論文集, pp.247-248, 2006.

[3] 大西優子, 山下洋一: "語学学習者の自動評定における尤度正規化について", 日本音響学会秋季講演論文集, pp.896-897, 2005.

[4] 坂口福太郎, 緒方淳, 有木康雄: "日本語・英語HMMを用いた発音評価と単語発声の誤り検出", 音響学会春季講演論文集, pp.151-152, 2001.

[5] 中村直生, 中川聖一: "日本人の英語音の評価法", 電子情報通信学会技術研究報告 SP 音声, Vol.102 No.107 , pp. 13-18, 2002.

[6] 峯松信明,富山義弘,塘辻正剛,吉本啓,清水克正,中川聖一,牧野正三: "日本人話者による英語文・単語音声データベースの構築", 日本音響学会秋季講論集, pp.199-200 ,2001.

[7] 峯松信明, 富山義弘, 吉本啓, 清水克正, 中川聖一, 塘辻正剛, 牧野正三: "日本人英語音声に対する母語話者英語教師による評価ラベリング", 日本音響学会秋季講演論文集, pp.215-226, 2002.