A-002

音素素性による重み付けを用いた発音記号列のアラインメントと精度比較

Alignment for Phonetic Symbol Strings and Accuracy Comparison Using Phoneme Feature Weighting

宮城裕太† 岡崎威生‡
Yuta Miyagi Takeo Okazaki

1 はじめに

言葉は口伝いに広がるため，同じ単語でも発音が異なるという現象が起こる．発音と発音の間の差異を明らかにするためには，発音記号で表された単語の発音どうしを比較し，アラインメントする必要がある．適切なアラインメントを行うには，適切な記号対の重み付けが必要となる．

本研究では，音素素性による重み付けを使用した発音記号列のアラインメントを提案した．そして，この手法によるアラインメントの精度検証を，従来法との比較実験によって行なった．

2 発音記号列のアラインメント

発音記号列は子音と母音，2種の発音記号から構成される．アラインメントとは，配列どうしを比較した際，配列間の要素を何らかの規則に従って対応付ける処理である．図1は，単語「今」についての，2地域間のアラインメント例である．“C”，“V”はそれぞれ子音と母音，“–”は音の消失を表す．

伊仙町伊仙	nj	aː				
宮古島市池間	N	n	a	m	a	

↓

	C	V	C	V	C	V
伊仙町伊仙	–	–	nj	aː	–	–
宮古島市池間	N	–	n	a	m	a

図 1: 発音記号列のアラインメント例（単語「今」）．

以下に，発音記号列のアラインメントを定式化する．まず，子音記号の集合を$C = \{c_1, c_2, \ldots, c_k\}$，母音記号の集合を$V = \{v_1, v_2, \ldots, v_l\}$としたとき，これらの和集合$\sum = \{C \cup V\}$を定義する．このとき，$X, Y \subset \sum$であるような長さ$m$，$n$の2本の発音記号列$X = x_1 x_2 \cdots x_m$，$Y = y_1 y_2 \cdots y_n$のアラインメントは，$X', Y' \subset (\sum \cup \{-\})$であり，配列長が$|X'| = |Y'|$となる発音記号列$X'$，$Y'$として定義される．ここで，“–”は$x'_i$，$x'_{i+1}$，もしくは$y'_i$，$y'_{i+1}$の間の空白，すなわちギャップを示し，$x'_i \neq -$かつ$y'_i \neq -$である．また，各記号対は$(x'_i, y'_i) \in C$，または$(x'_i, y'_i) \in V$を満たす．すなわち，子音は子音と，母音は母音とのみ対応付けられる．ここで，コスト関数$w : (\sum \cup \{-\}) \times (\sum \cup \{-\}) \to \mathbb{R}$を導入する．これは，任意の記号対$(x, y)$に対して，アラインメントされた際の重みを与える関数である．アラインメントされた各記号対の重みの合計値$w(X', Y') = \sum_{i=1}^{|X'|} w(x'_i, y'_i)$をアラインメントスコアと呼ぶ．アラインメントスコアを最適化することで，可能なアラインメントの中から最適アラインメントを決定することができる．

†琉球大学大学院理工学研究科情報工学専攻，Graduate School of Engineering and Science, University of the Ryukyus
‡琉球大学工学部工学科知能情報コース，Computer Science and Intelligent Systems, University of the Ryukyus

3 アラインメントアルゴリズムと発音記号対の重み付け

Needleman-Wunsch法[1]は，動的計画法によって2本の配列間の最適アラインメントを計算するアルゴリズムである．長さm，nの2本の配列X，Yの最適アラインメントは，行と列にそれぞれ配列X，Yの要素を配した$m \times n$の距離行列Dを計算することによって得られる(式(1))．3つの式はそれぞれ配列Xに対するギャップの挿入，配列Yに対するギャップの挿入，記号対の一致･不一致に対応する．$w(x_i, -)$，$w(-, y_j)$はギャップペナルティーと呼ばれ，無駄なギャップの発生を防ぐため，ペナルティーとしての重みが与えられる．

$$D_{i,j} = \min \begin{cases} D_{i-1,j} + w(x_i, -), \\ D_{i,j-1} + w(-, y_j), \\ D_{i-1,j-1} + w(x_i, y_j). \end{cases} \quad (1)$$

アラインメントに使用される記号対の重み付けには，様々な手法が提案されており，言語学の分野においても，発音記号対の重み付けに関する研究が行われてきた．Wieling[2]は，Levenshtein距離に改良を加えた，以下の3つ重み付けによって，ブルガリア方言の発音記号列を使用したアラインメントを行なった．

1つ目は，Levenshtein距離の計算において子音と母音の記号対の発生を抑制した手法である．Levenshtein距離は式(2)で計算される．ここで，$1_{(x_i \neq y_j)}$は，$x_i \neq y_j$のときに1，それ以外のとき，すなわち一致する発音記号対のときに0となる．

$$D_{i,j} = \min \begin{cases} D_{i-1,j} + 1, \\ D_{i,j-1} + 1, \\ D_{i-1,j-1} + 1_{(x_i \neq y_j)}. \end{cases} \quad (2)$$

2つ目は，Levenshtein距離の計算に発音記号の入れ替えを行うスワップ処理を加え，$x_i = y_{i+1}$，$x_{i+1} = y_i$の場合に限定して適用する手法である．式(2)にスワップ処理を加えると，式(3)のようになる．ここで，$2\left(1_{(x_i \neq y_{j+1})} + 1_{(x_{i+1} \neq y_j)}\right)$は，$x_i \neq y_{i+1}$，$x_{i+1} \neq y_i$の場合にスワップ処理が発生することを防ぐためのペナルティーである．$x_i = y_{i+1}$，$x_{i+1} = y_i$を満たしていれば0となる．Wielingはスワップ処理の重みを0.999とした．これは，ブルガリア方言ではスワップ処理に該当するアラインメント

が比較的発生しやすいことに起因する．

$$D_{i,j} = \min \begin{cases} D_{i-1,j-1} + 1, \\ D_{i-1,j-1} + 1, \\ D_{i-1,j-1} + 1_{(x_i \neq y_j)}, \\ D_{i-2,j-2} + 0.999 \\ \qquad + 2\left(1_{(x_i \neq y_{i+1})} + 1_{(x_{i+1} \neq y_i)}\right). \end{cases} \tag{3}$$

3つ目は，PMI(pointwise mutual information)を，編集操作の重みとして適用した手法である．PMI(式(4))は，ある単語についての全ての2地域間のペアワイズアラインメントにおいて，発音記号xとyのアラインメントのされやすさを示す指標となる．分子の$p(x,y)$は，その中で発音記号xとyがアラインメントされる確率，分母の$p(x)$，$p(y)$はそれぞれ，ペアワイズアラインメント中に発音記号x，yが発生する確率である．

$$\mathrm{PMI}(x,y) = log_2\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{4}$$

まず，子音と母音の発生を防いだLevenshtein距離によるアラインメントを行い，その結果に対してPMIを計算する．計算されたPMI は，0からPMIの値を減算し，PMIの最大値を加算することで編集操作の重みに変換される．そして，このPMIによる重みを使用した再アラインメントを，結果が変化しなくなるまで繰り返す．Wieling[3]は更に，一致する記号対，すなわち$x = y$となる記号対PMIは計算しないこと，式(4)の分子が0の場合に，PMIが∞に発散することを防ぐ改善を行った．

4　音素素性による発音記号対の重み付け

Wielingによる発音記号対の重み付けは，発音の特徴を考慮していない．そこで，音素素性による発音記号対の重み付けを提案した．音素素性とは，舌の位置，唇の形など，その音の発声上の特徴である．子音，母音の発音記号はそれぞれ5種の音素素性を持っている．そこで，各発音記号対の重みを，音素素性の一致数によって与えた．音素素性の一致数は，表1のようにして求める．表中の“|”は音素素性の一致を示す．3つの音素素性が一致しているため，記号対(p,ɸ)の重みは3である．

表 1: 音素素性の一致数(子音記号“p”と“ɸ”).

	位置	方法	口蓋・円唇	喉頭化	声帯振動
p	両唇	破裂/閉鎖音	両方なし	区別なし	無声音
	|		|	|	|
ɸ	両唇	摩擦音	両方なし	区別なし	無声音

本研究で扱う琉球諸語の発音記号列は，81種の子音，35種の母音記号から構成されるため，各発音記号対の重みは，式(5)に示す行列によって定義される．また，子音と母音の記号対についてはアラインメントを行わないため，音素素性による重み付けは子音記号対，母音記号対についてのみ行った．f，gをそれぞれ，子音記号対，母音記号対の音素素性の一致数を示す関数とすると，子音記号対，母音記号対の重みs_1，s_2は$s_1 = f(c_i, c_j)$，$s_2 = g(v_i, v_j)$と定義される．

$$\begin{array}{c} \\ c_1 \\ c_2 \\ \vdots \\ c_{81} \\ v_1 \\ v_2 \\ \vdots \\ v_{37} \end{array} \begin{array}{c} \begin{array}{cccccccc} c_1 & c_2 & \cdots & c_{81} & v_1 & v_2 & \cdots & v_{37} \end{array} \\ \begin{pmatrix} s_1 & s_1 & \cdots & s_1 & & & & \\ & \ddots & \ddots & \vdots & & & & \\ & & \ddots & s_1 & & & & \\ & & & s_1 & & & & \\ & & & & s_2 & s_2 & \cdots & s_2 \\ & & & & & \ddots & \ddots & \vdots \\ & & & & & & \ddots & s_2 \\ & & & & & & & s_2 \end{pmatrix} \end{array} \tag{5}$$

s_1：子音記号対のスコア
s_2：母音記号対のスコア

各発音記号が持つ音素素性は5個であるため，$0 \leq s_1, s_2 \leq 5$である．

この音素素性による重み付けは，各発音記号対の類似度を表す．つまり，アラインメントスコアが大きいほど，配列間の類似度が高いことを示す．よって，式(6)に従いアラインメントスコアを最大化することで，最適アラインメントが得られる．ここで，$p(<0)$はギャップペナルティー，Sは音素素性による重みを定義した行列(式(5))である．

$$D_{i,j} = \max \begin{cases} D_{i-1,j} + p, \\ D_{i,j-1} + p, \\ D_{i-1,j-1} + S_{i,j}. \end{cases} \tag{6}$$

次に，最適なギャップペナルティーを決定するために予備実験を行なった．実験に使用したのは，琉球諸語110単語の発音記号列である．単語ごとに最大で95地域，最小で2地域についての発音記号列がある．アラインメントは，単語ごとに2地域間の発音記号列の組み合わせ全てに対してNeedleman-Wunsch法を適用して行なった．次に，アラインメント結果の精度を検証するため，狩俣[4]によるアラインメントを使用した．狩俣は，琉球諸語の発音について調査，記録を行い，地域間の発音の差を明らかにするため，単語ごとに発音記号列のアラインメントを行った．狩俣によるアラインメントと，音素素性による重み付けを使用したアラインメントを比較し，単語ごとに一致したアラインメント数を，その単語の総アラインメント数で除し，一致率を計算した(式(7)).

$$\text{一致率} = \frac{\text{一致したアラインメント数}}{\text{アラインメント総数}} \tag{7}$$

アラインメントは，−1から−10のギャップペナルティーについて行い，式(7)によって単語ごとの一致率を計算した．そして，110単語全体での平均一致率を計算し，平均一致率と最低一致率との差が小さいことを基準に最適なギャップペナルティーを決定した．ギャップペナルティーが−3から−10のとき，結果は同一で，平均一致率と最低一致率との差が最も小さくなった．そこで，−3をギャップペナルティーの最適値とした．

以上の結果を踏まえ，本研究では以下に示すアラインメント手法を提案する．

1. 与えられた単語の発音記号列のうち，2地域の発音記号列を選択する．

2. 入力となる2地域の発音記号列間のアラインメントを，$p = -3$とした式(6)に従い，Needleman-Wuhnsch法によって計算する．

3. 以上の過程を，2地域の組み合わせ全てについて行い，アラインメントを計算する．

5 アラインメントの精度検証

音素素性による発音記号対の重み付けを使用した発音記号のアラインメントと，Wielingが提案した，Levenshtein距離に改良を加えた3つの重み付けを使用したアラインメント精度の比較を行った．琉球諸語110単語の発音記号列を使用し，前節と同様の方法でNeedleman-Wunsch法によるアラインメントを行い，平均一致率，および一致率が100%に達した単語数を求めた(表2)．

表 2: 平均一致率，および一致率が100%に達した単語数の比較(PF : 音素素性，LV : Levenshtein距離，SW : スワップ処理，PMI : PMI-weighting)

	平均一致率(%)	一致率100%の単語数
PF	96.438	70
LV	94.921	70
SW	94.921	70
PMI	95.077	63

平均一致率が最も高かったのはPFによるアラインメント，続いてPMI，LV，SWによるアラインメントという結果となった．一致率が100%に達した単語数では，PF，LV，SWによるアラインメントがともに70単語，続いてPMI によるアラインメントとなった．SW，およびLVのアラインメント結果が一致しているのは，検証に使用した琉球諸語の発音記号列では，スワップ処理が発生しなかったためである．この結果は，琉球諸語ではスワップ処理に該当する発音の変化が発生しないことを示す．従って，Wielingによるスワップ処理の発音記号列への適用は，琉球諸語においては適切ではない．そのため，PF，LV，PMIの結果についてのみ述べる．

図2, 3, 4はそれぞれ, PF, LV, PMIによるアラインメント結果の，単語ごとの一致率の分布である．平均一致率や，一致率が100%に達した単語数から，ほとんどの単語について，PFによるアラインメントが有効であると考えられる．表3, 4はそれぞれ，PFによるアラインメント精度が，LV，PMIを下回った結果の一部である．LV，PMIによるアラインメントが，狩俣によるアラインメントと一致している．表5, 6は, PFによるアラインメント精度が, LV, PMIを上回った結果の一部である．PFによるアラインメントが，狩俣によるアラインメントと一致している．PF対にして，LV，PMIによるアラインメントは，同じ記号対を作成するために，比較的ギャップを挿入しやすい傾向にあった．表6の単語「耳」のアラインメントの場合，記号対(i, i)を作成するために，合計3つのギャップが挿入されている．これは，Levenshtein距離では記号の置換，ギャップの挿入に対して同じ重みが与えられていることが影響していると考えられる．これによって表3，4のように，PFによるアラインメントよりも精度が高くなった単語があった．しかし，前述の理由から，これらのアラインメントが狩俣によるアラインメントと一致したのは，偶然性の高い結果であると考えられる．

6 おわりに

本稿では，発音記号列のアラインメントを行うために，音素素性による発音記号対の重み付けを提案した．そして，Wielingによる重み付けを使用したアラインメント精度との比較実験を，琉球諸語110単語の発音記号列で，Needleman-Wunsch法を使用して行った．いずれの結果においても，音素素性による重み付けを使用したアラインメントの精度が，Wielingによる手法の精度を上回った．

今後は，音素素性と，Wielingによる重み付けとを組み合わせた手法の提案と，他言語の発音記号列を使用した精度検証を試みる．

参考文献

[1] S. B. Needleman and C. D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443 – 453, 1970.

[2] M. Wieling, J. Prokić, and J. Nerbonne, “Evaluating the pairwise string alignment of pronunciations,” in *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, (Stroudsburg, PA, USA), pp. 26–34, Association for Computational Linguistics, 2009.

[3] M. Wieling and J. Nerbonne, “Measuring linguistic variation commensurably,” *Dialectologia*, vol. Special Issue II, pp. 141–162, 01 2011.

[4] 狩俣繁久, “危機言語としての琉球方言の研究状況: 日本復帰 後から今日までの活動についてのおぼえがき,” **国立民族学博物館調査報告**, vol. 39, pp. 257–267, 2003.

表 3: PFの精度がLVの精度を下回ったアラインメント(PFが不一致，LVが一致)

地域名	PF	LV
09-015足(1-5)		
喜界町志戸桶	pjh – – a:	pjh a: – –
国頭村辺土名	ɸ i s a	ɸ i s a
18-036ゑけり(1-1)		
旧笠利町笠利	j i h i r i	j i h i r i
与那国町祖納	b i – – g i	b i g i – –

表 4: PFの精度がPMIの精度を下回ったアラインメント(PFが不一致，PMIが一致)

地域名	PF	PMI
18-036ゑけり(1-1)		
宇検村宇検	j e – – r i	j e r i – –
瀬戸内町嘉鉄	j i ç i r ï	j i ç i r ï
53-103B鍬(1-7)		
喜界町阿伝	k’ – e:	k’ e: –
和泊町和泊	k’ o i	k’ o i

表 5: PFの精度がLVの精度を上回ったアラインメント(PFが一致，LVが不一致)

地域名	PF	LV
01-003首(2-2)		
宮古島市池間	n u b u i	n u b u i
石垣市大浜	n u b – y	n u b y –
07-013声(1-1)		
瀬戸内町管鈍ドン	kʻ u i:	kʻ u i: –
旧具志川村兼城	kw i: i	kw – i: i

表 6: PFの精度がPMIの精度を上回ったアラインメント(PFが一致，PMIが不一致)

地域名	PF	PMI
05-009耳(1-2)		
旧笠利町笠利	m i N –	m – – i N
勝連町津堅	m i: m i	m i: m i –
07-013声(1-1)		
瀬戸内町管鈍ドン	kʻ u i:	kʻ u i: –
旧具志川村兼城	kw i: i	kw – i: i

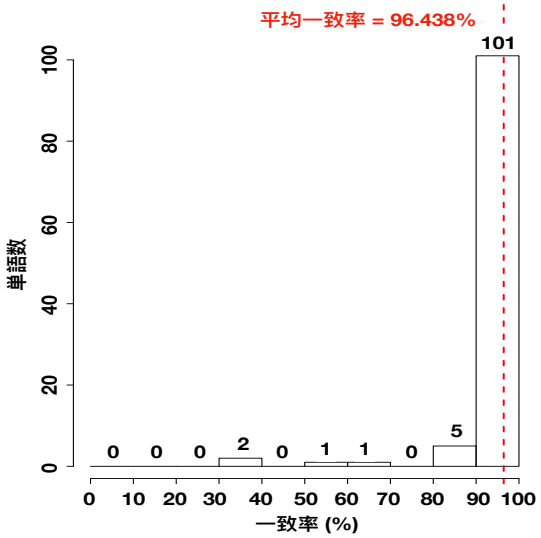

図 2: 単語ごとのアラインメント一致率の分布(PF)

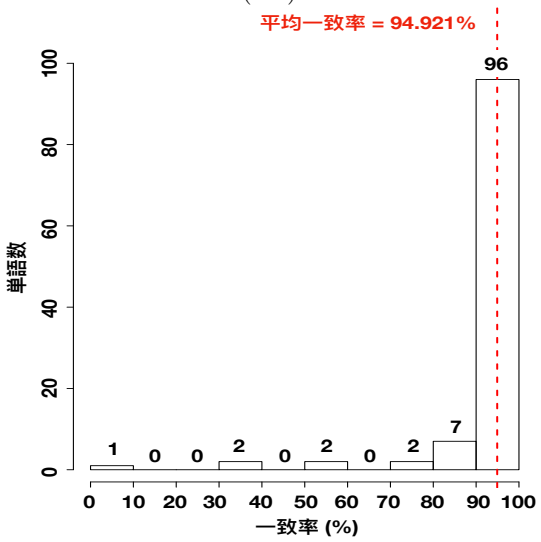

図 3: 単語ごとのアラインメント一致率の分布(LV)

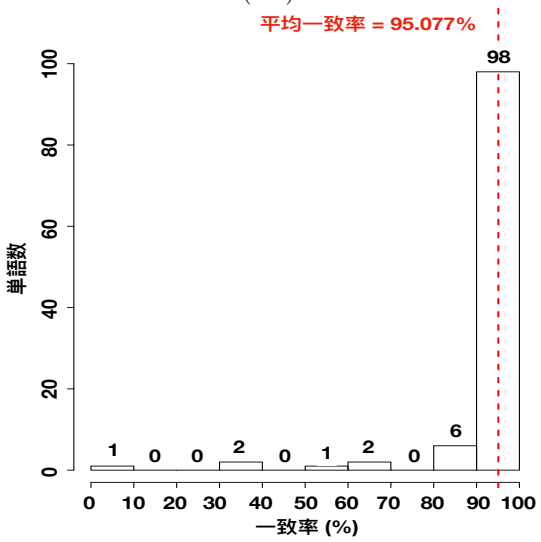

図 4: 単語ごとのアラインメント一致率の分布(PMI)