ピッチ情報を援用した圧縮類似度による方言の自動分類

Automatic classification of Japanese dialects using compression similarity distance assisted by pitch information

谷　研究室　　齋藤 岳丸、山本 峻、鴫原 克幸

Takemaru Saito , Satoshi Yamamoto , Katsuyuki Shigihara

概要

圧縮によるデータ間の類似度に関する距離が定義され、DNA の類似度や言語の類似度、音楽の類似度判定に有用だという実験結果が得られている。昨年、荒堀、福田氏がが圧縮類似度を用いた方言の類似度分析に対して、音声入力データのの自動テキスト化や木の評価方法、圧縮方法の違いによる研究を行ったが、本研究では音声入力データのテキスト化にピッチデータを含むことについての追加研究を行う。

1　はじめに

現在様々な事柄がデータ化されている。もちろん、音楽や文章も「データ」として扱われるようになってきている。近年 Ming Li らが圧縮類似度に基づくデータ間の類似度に関する距離を表す similarity metric を発案（現在 DNA の類似度や言語の類似度、音楽の類似度判定に有用だということが分かっている）した。

昨年、荒堀氏、福田氏が圧縮類似度を用いた方言の類似度分析に対して、音声入力データをドラゴンスピーチという音声認識ソフトで音声データをテキスト化したものを用いて実験を行った。本研究では音声入力データを音声録聞見というソフトウェアによりピッチ情報を抽出し、その情報をテキストに付加したデータを用い、去年のデータとの比較を行う追加研究とする。

第２章では Kolmogorov 記述量について、第３章では圧縮類似度について、第４章では音声のピッチ情報の説明を、第５章ではクラスタリングの説明、第６章では使用したクラスタリングの概要について述べ、第７章で実験の概要と結果について説明をし、第８章で本研究での考察をを述べる。

2　Kolmogorov 記述量の定義

2.1　Kolmogorov 記述量

ある与えられた有限の対象 (例えば文字列) に含まれる情報量を, その対象を生成するプログラムのサイズ（すなわちビット数）と定義する。ただし, ここで言う ’生成するプログラム ’とは全てのメモリが空の初期状態からスタートし, その対象を出力して停止するプログラムのことである。この場合, どんなプログラミング言語を選んでも, それが妥当であればその情報量の差, すなわち各プログラムのサイズの差は定数分の差しかないことが証明されている。このように定義した情報の量のことを kolmogorov 記述量という。

S をプログラム言語、$|p|$ をプログラムサイズとすると、対象 x の Kolmogorov 記述量 $K_s(x)$ は、以下のように定義される。

$$K_s(x) = min\{|p| : S(p) = x\}$$

すなわち $K_s(x)$ は、”プログラム言語 S において x を生成する最小のプログラムの長さ ”と考えていいだろう。

次に補助情報 y が与えられているときの対象 x についての記述量（相対記述量）について考える。文字列 x を、計算可能関数（インタプリタ関数）f、それに文字列 p と y により $f(p, x) = x$ と記述することを考える。インタプリタ関数 f のもとで、補助関数 y に対する x の記述量 K_f を次のように定義する。

$$K_f(x|y) = min\{|p| : P \in \{0, 1\} * f(p, y) = x\}$$

補助情報 y はあらかじめ与えられている情報で, この情報が生成したい対象 x の情報を含んでいれば対象を出力するプログラムは容易になる。

また、x と y を区別できる形で出力させる最短のプログラムの長さ $K(xy)$ とあらわす。$O(logK(xy))$ の誤差範囲では、

$$K(xy) = K(x) + K(y|x)$$

となることが証明されている。

2.2　情報量

相対記述量 $K(x|y)$ が絶対記述量 $K(x)$ よりかなり小さい場合、「y が x についての情報をかなり含んでいる」と解釈することが出来る。したがって、定数分の差異を無視すれば、

$$I(x : y) = K(y) - K(y|x)$$

は ”y に含まれる x の情報量 ”とすることが出来る。また、$K(x|x)$ となる x を選べば、

$$I(x:x) = K(x)$$

となる。
ここで情報の対象性ということを考える。先述より $O(logK(xy))$ の誤差範囲では、

$$K(xy) = K(x) + K(y|x)$$

である。したがって

$$I(x:y) = I(y:x)$$

が成立する。

3 圧縮類似度

数学において距離空間とは，任意の 2 点間で距離が定められた空間のことをいう。

定義

ある集合 X 上の距離とは，実数値関数 $d : X \times X \to R$ で任意の $x, y, z \in X$ に対して次のような性質を満たす。

$$d(x,y) \geq 0$$
$$d(x,y) = 0 \Leftrightarrow x = y$$
$$d(x,y) = d(y,x)$$
$$d(x,y) \leq d(x,z) + d(z,y) : \text{三角不等式}$$

これをもとに，情報距離について考える。

Ming Li,Xin Chen らの研究では，情報に関する距離を標準化していて, 任意の文字列 x, y について，以下のように決めている。

$$d(x,y) = \frac{max\{K(x|y), K(y|x)\}}{max\{K(x), K(y)\}}$$

また，$K(y) \geq K(x)$ としたとき，

$$d(x,y) \quad = \quad \frac{K(y|x)}{K(y)}$$

これに対して先に述べた情報量の公式を用いると，

$$d(x,y) \quad = \quad \frac{K(y) - I(x:y)}{K(y)}$$

となる。また 2 節で示した通り，$O(\log K(xy))$ の範囲では $K(xy) = K(x) + K(y|x)$ が成り立つので，

$$d(x,y) = \frac{K(xy) - K(x)}{K(y)}$$

と表すことができる。

kolmogorov 記述量は計算でその値を算出することが不可能であるため, 圧縮プログラムを用いて値を近似的に求めることになる。$bz2(x)$ を文字列 x を bzip2 で圧縮したときのファイルサイズとすると情報距離 $d(x,y)$ は以下のように近似される。

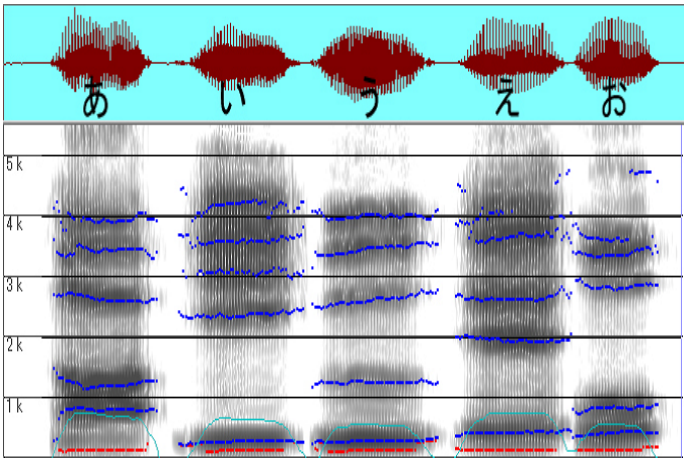
$$d(x,y) \fallingdotseq \frac{b2(xy) - bz(x)}{bz2(y)}$$

また圧縮方法 C が以下を満たすとき $O(\log n)$ の誤差の範囲で万能であると証明されている。
1. 任意の文字列 x に対して $C(xx)$=$C(x)$ であり、空文字列λに対して、$C(\lambda)$=0 である。
2. 任意の文字列 x, y に対して $C(xy) \geqq C(x)$。
3. 任意の文字列 x, y に対して $C(xy) = C(yx)$。
4. 任意の文字列 x, y, z に対して $C(xy)$+C(z)≧C(xz)+C(yz)。

4 音声のピッチ情報

人間の発生する音声は、声帯が振動することによって発生する音を元に構成され、様々な周波数成分を含む。

特に母音が発声される時に声帯が振動し、この周波数をピッチ周波数という。



一番下の線（赤）がピッチ情報

この周波数に方言による特徴が出ると考えられる。

今回は音声録聞見というソフトウェアを使用し、音声のピッチ情報を抽出し、これをデータの一部として利用する。

このデータは音声 1 秒間を 200 個に区切り、その箇所その箇所のピッチ周波数の値を得られる。

	A	B	C	D	E
238	26184	1.19	-9.63	0	
239	26295	1.19	-7.6	0	
240	26405	1.2	-7.96	0	
241	26515	1.2	-13.56	0	
242	26625	1.21	-9.04	0	
243	26736	1.21	-11.37	0	
244	26846	1.22	-8.1	0	
245	26956	1.22	-15.74	0	
246	27066	1.23	-9.46	0	
247	27177	1.23	2.21	178	
248	27287	1.24	11.17	7	
249	27397	1.24	14.74	128	
250	27507	1.25	11.81	136	
251	27618	1.25	17.09	136	

ピッチの抽出

5 クラスタリング

今回は NJ 法、UPGMA、quartet-method の三つのクラスタリング手法を採用した。

これらはどれだけ似ているかを示す尺度として距離を用いるという性質があることと、例年との比較を行うために使用した。

5.1 NJ 法

近隣結合法と呼ばれ、全ての枝の長さの総和が最小になるように系統樹を作成していく無根系統樹。

効率が良く、他の方法では扱えないような大量のデータを扱うことができるが、出力された木が最適とは限らない。

5.2 UPGMA

平均距離法と呼ばれる単純な系統樹制作法で、最も近縁なクラスタ間の距離の平均を求めながら系統樹を作成する。　葉から決まり、最後の根の位置が決まる有根系統樹。

5.3 Quartet Method

木が距離表にどの程度沿っているか評価する基準があり、ランダムに木を変更し評価値を更新していくヒューリスティック（試行錯誤）アルゴリズムである。

NJ 法、UPGMA 法と比較すると大幅に処理時間を要する。

6 前処理

6.1 前処理の目的

先行研究での前処理の目的

先行研究では文字情報での比較を行っていた。文字情報は一般的には SJIS,EUC,UTF-8 等の文字コードとして表現されるが、圧縮類似度を算出する場合には無駄な情報を含んでいるのは好ましくない為、より簡潔にデータを表すべきである。

そこで、文字を数値に置き換えて単純化している。

本研究での前処理の目的

文字を数値に置き換えて単純化しているのは先行研究と同じだが、本研究では文字情報にピッチ情報を組み込む必要がある。

先行研究よりも複雑かつ方法が無限に考えられるため、本研究の要点ともいえる。

今回は前処理を 2 種類検討し、実験を行った。

6.2 文字情報の単純化

文字情報の単純化の具体的な手法について述べる。

下表は「あ」「か」「さ」という文字の各文字コードでの表現と、単純化変換を行った後の表現である。

SJIS,EUC は 1 文字を 2byte(UTF-8 は 3byte) で表現するが、先頭の 1byte(UTF-8 は 2byte) は共通の値が入っている。

圧縮類似度を求める場合はこのような情報は不要なものなため、ひらがなに対して文字コード順に番号を割り当て、1 文字 1byte の情報に変換を行う。

単純化例

文字	SJIS	EUC	UTF-8	単純化
あ	0x82A0	0xA4A2	0xE38182	0x02
か	0x82A9	0xA4AB	0xE3818B	0x0B
さ	0x82B3	0xA4B5	0xE38195	0x15

6.3 前処理１

ピッチ情報 (Hz) の値は話者によって差があるが、概ね 300Hz 前後までの値で構成されているため、2byte で表すことにする。

文字情報は先の単純化に従い変換を行うが、ピッチ情報とのデータの長さを揃えるために 2byte で表すことにする。

この情報を交互に並べ、実験データとする。

この実験データは、文字情報にピッチ情報をそのまま付加することを目的としている。

前処理１例

文字	文字 (単純化後)	ピッチ (Hz)
お	10：0x000A	116：0x0074
ば	48：0x0030	155：0x009B
あ	2：0x0002	145：0x0091
さ	21：0x0015	98：0x0062
ん	83：0x0053	130：0x0082
が	12：0x000C	140：0x008C

→完成データ (16 進数表現)

000A 0074 0030 009B 0002 0091 0015 0062 0053 0082 000C 008C

6.4 前処理２

特定の話者のピッチ情報 (Hz) の平均値をとり、その値と 128(1byte 符号なし整数表現 0〜255 の中央値) の差をオフセット値とする。

ある時点のピッチ情報 (Hz) からオフセット値を引き、その値をピッチ情報として使用し、1byte で表す。

文字情報は先の単純化に従い変換を行い、1byte で表す。

この情報を交互に並べ、実験データとする。

この実験データは、文字情報にピッチ情報を付加するのは前処理１と同じだが、ピッチ情報の話者の男女差等の個人差を平均化することを目的としている。

前処理２例

ピッチ情報平均値：147

オフセット値：19

文字	文字 (単純化後)	ピッチ (Hz)
お	10：0x0A	116 - 19 = 97：0x61
ば	48：0x30	155 - 19 = 136：0x88
あ	2：0x02	145 - 19 = 126：0x7E
さ	21：0x15	98 - 19 = 79：0x4F
ん	83：0x53	130 - 19 = 111：0x6F
が	12：0x0C	140 - 19 = 121：0x79

→完成データ (16 進数表現)

0A61 3088 027E 154F 536F 0C79

7 実験

実験データを２つと、それをクラスタリングするための前処理を２種類用意する。

前処理を施した実験データは、それぞれ前年度の研究結果から bzip2 にて圧縮をし、距離表を出す。

その４種類のデータを３種類のクラスタリング手法で視覚化し、結果についての考察を述べる。

7.1 実験データ

実験データは以下の手順で作成される。

音声を文字データに変換する（手作業）

↓

音声からピッチ情報を取得する

↓

取得したピッチ情報に対応する文字データを割当てる

↓

このデータに各種前処理を行い、実験データとする。

データ１

方言ももたろう（富士通ビーエスシー）の音声データから音声録聞見（東京大学 大学院 医学系研究科）でピッチを抽出し、一昨年の研究で使用したテキストデータを手動で割り当てた csv ファイルを用いる。

データ２

方言ももたろうの音声データは、録音時のノイズが入っているものも多く、正しいピッチ情報が取得できていない可能性があるので、Audacity（オープンソース）を用い、ある程度ノイズを取った後、音声録聞見でピッチを抽出し、テキストデータを割り当てる。

以上二つのデータに前処理を施したものを実験データとする。

7.2 結果

視覚化

視覚化したときに見やすいよう、以下のように地域によって色分けをした。

北海道	黒
東北	緑
関東	青
中部	黄色
近畿	ピンク
四国	紫
中国	オレンジ
九州	水色
沖縄	白

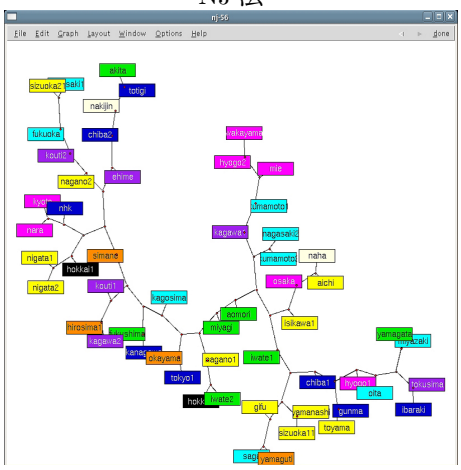
この色分けで、ノイズカットを行った場合と行わない場合の、前処理１，２についてのクラスタリング結果を以下に示す。

結果

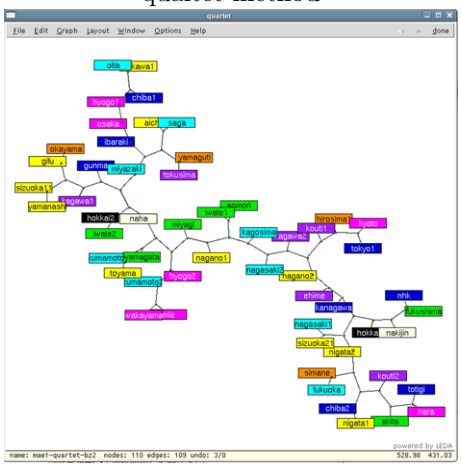
前処理 1

■ノイズカットをしない場合

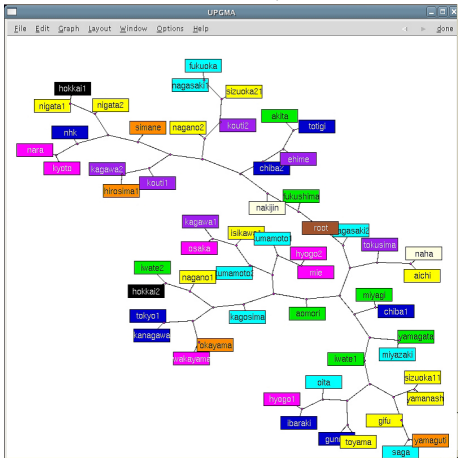
NJ 法



qualtet-method

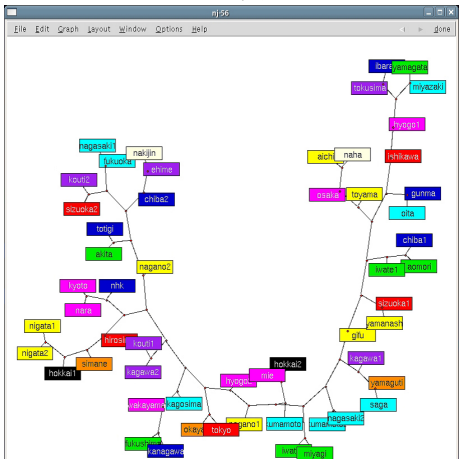


UPGMA 法

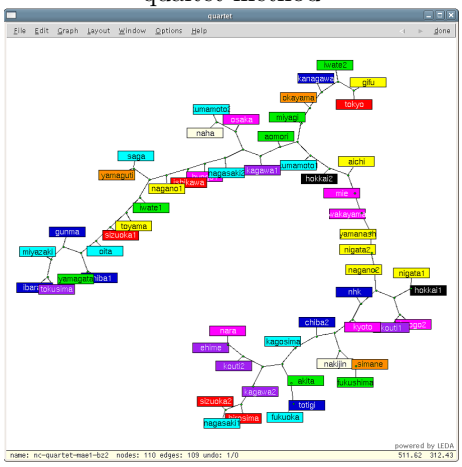


■ノイズカットをした場合

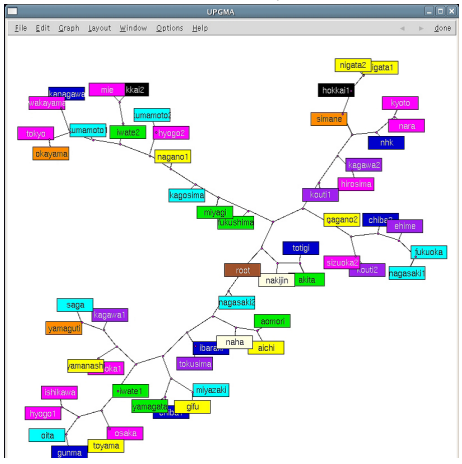
NJ 法



qualtet-method



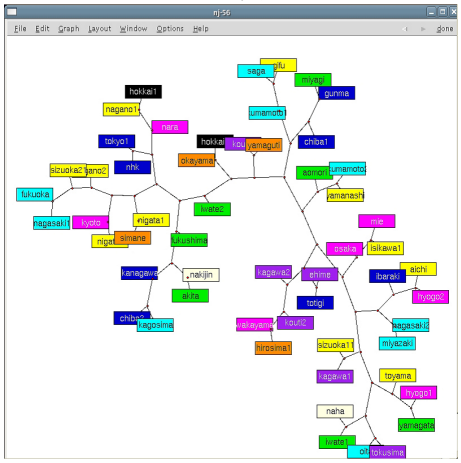
UPGMA 法



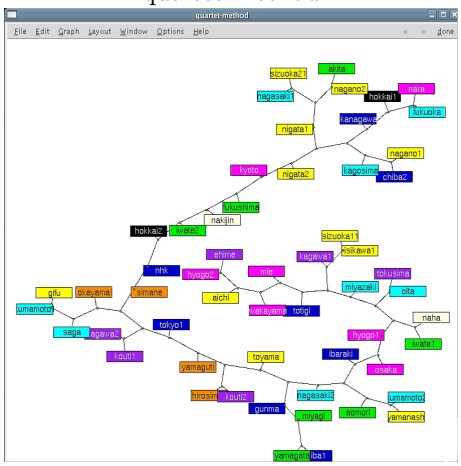
前処理２

■ノイズカットをしない場合

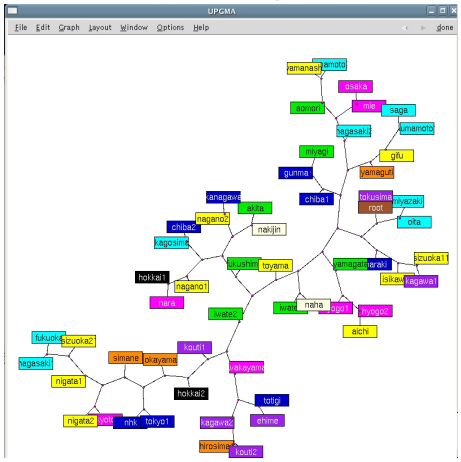
NJ 法



qualtet-method

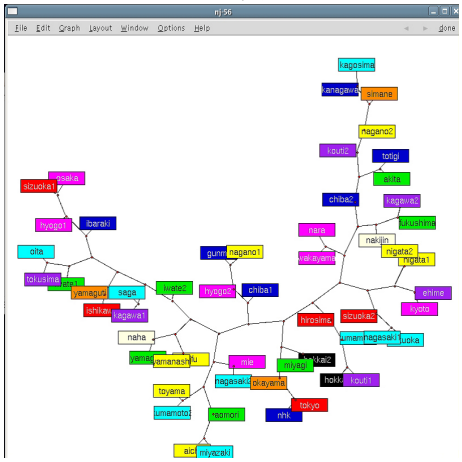


UPGMA 法

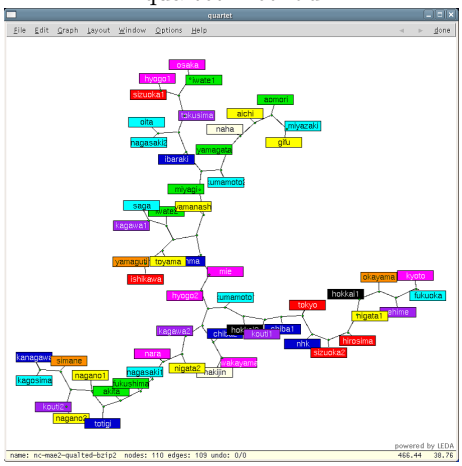


■ノイズカットをした場合

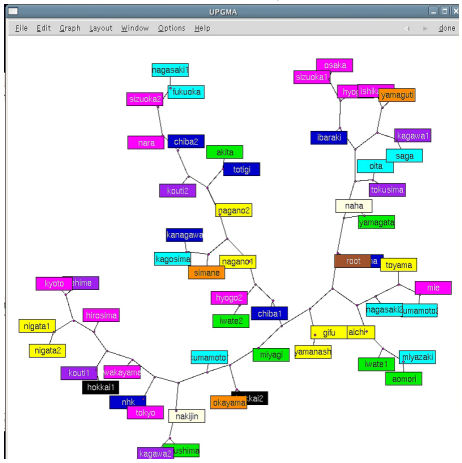
NJ 法



qualtet-method



UPGMA 法



8　考察

局所に地域ごとのまとまりが見られるが、全体的にはばらばらでいい結果は得られず、文字データだけで分類した先行研究より劣る結果となった。

これは抽出したピッチ情報が録音環境に左右されたデータである可能性が高い事や、　実験データの構成の仕方がよくない事などが考えられる。

また、方言の自動分類にピッチ情報を組み込むのが有効かどうかも判断し難い。

このような結果から、今後の課題として以下の項目が考えられる。

・最適なピッチ情報を取得できるような音声処理。
・音声の長さを一定にする。
・音声の男女差、声質差などの違いを考慮する。
・ピッチ以外の音声情報の抽出・適用。
・適切な前処理・実験データの作成方法を考案する。
・自動的に音声から日本語割り当てを行う。
・人間の声についての知識を深める。
・各地方の歴史と方言のかかわり方。

9　参考文献

(1)　谷 聖一　情報数学特別講義（2008）

(2)　Ming Li and Paul M.B.Vitanyi. 渡辺治翻訳 Kolmogorov Complexity and its Applications. コンピュータ基礎理論ハンドブック (1994)Ming Li and Paul M.B.Vitanyi. 渡辺治翻訳 Kolmogorov Complexity and its Applications. コンピュータ基礎理論ハンドブック (1994)

(3)　今石 元久　音声研究入門（2005）

(4)　高野 浩明　Kolmogorov 記述量に基づく類似度を用いた方言の自動分類　(2007)

(5)　堀中 幸司　Kolmogorov 記述量に基づく類似度を用いた方言の自動分類　(2007)