ショートノート

Wavelet解析とSIFT特徴量を用いた音韻の位置合わせ†

松下 真也*1・高野 敏明*2・友次 克子*2

「少数言語の耳を頼りにした書き起こし」という手作業が、言語学者らによって行われている。この手作業は、膨大な時間と複雑な音声表記を必要とする。そこで本研究ではこの作業をサポートするようなシステムの構築を目指す。これまでの実験では、MFCCを使用して「異なる言語間における音のマッチング」を調査したが、精度は低く、精度向上のために前処理が必要と考えられた。そこで本論では、Wavelet解析とSIFT特徴量を用いた前処理に取り組んだ。

キーワード：音声処理、音韻情報

1. はじめに

現在、少数言語（使用者が少ない言語）の保護活動が言語学者らによって行われ、その1つに「少数言語の耳を頼りにした書き起こし」という記録作業がある。この記録作業では少数言語の網羅的な発音の記録が求められることから、国際音声記号（IPA: International Phonetic Alphabet）を用いて書き起こしを行う。ここで、IPAとは音声学から「あらゆる言語の発音を説明することが可能な記号」[1]と定義されるものであり、記録者はIPAを用いるための訓練が要求される。しかし、訓練の度合い（熟練度）は定性的なものであり、記録者によって記録結果が異なる可能性がある。したがって、本研究では記録結果の差異を小さくするために、作業をサポートするようなシステムの構築を目指す。

また、少数言語を扱うにあたって気をつけなければならないことがある。それは、少数言語（以降未知語）は使用者が少ない言語であるだけでなく、音韻情報が未知（IPAに基づく辞書が存在せず、母音・子音の数や種類が未知）であり、加えてどの位置（時間）に音韻情報が含まれるかが未知なことである。これまでは音韻情報の推定（ある少数言語にどんな母音・子音が存在するのか）に焦点をあて、MFCC（Mel Frequency Cepstrum Coefficient）を用いた実験を行ってきたが、結果は精度の高いものでは無かった。この結果より、前処理として「音韻情報の位置合わせ」が必要と考えられたことから、本論文ではWavelet解析とSIFT特徴量を用いた音韻情報の位置合わせを試みた。

2. MFCCを用いた実験と必要な前処理

英語の beautiful の音声を日本人が聞いた際、「びゅーていふ

る」と書き起こしが可能である。「似た音の紐付け」は音声学、言語学の知見（IPA：国際音声記号）によって説明される。そこで本研究では、未知語における音韻情報の確認に「似た音の紐付け」が利用可能と考え、さしあたっては、この知見を数値上で確認することを目的に実験を行った。実験では、異なる言語間として英語と日本語を設定し、両言語の音声データに特徴量抽出（MFCC）を行い、Cos類似度によるマッチングを行った[2,3]。実験条件として、仮定には「異なる言語間における音でも一部相関がある」を設定し、未知語においては、どの時間帯に音韻情報が含まれるか未知であるという状況に合わせ、音声データの録音時間は全て1秒で統一し、表1に従い単一話者によって録音した。実験結果は「英語音素/dʒ/（例：age）と日本語音節「じ」の類似度が高い」を一例にして仮定の成立を確認することができた。しかし、仮定の成立を示す結果は全体の一部であり、精度は高いものでは無かった。

これまでの実験では図1に示すように、異なる言語間において似ている音韻情報が1秒内に含まれていても、その音韻情報の録音位置が異なる場合があり得た。これではMFCCで正しい結果を得られていない可能性があることから、音韻情報のズレを対象に「音韻情報の位置合わせ」を必要の前処理として考えた。図1を参考に、音韻として、特徴的な波形を含む位

表1 録音条件

録音周波数	録音時間	チャンネル	録音環境
44100 Hz	1 秒	1	Praat

図1 音韻情報のズレ

† Phoneme Positioning Using Wavelet Analysis and SIFT Features
Shinya MATSUSHITA, Toshiaki TAKANO,
and Katsuko TOMOTSUGU

*1 三重大学大学院工学研究科
Graduate School of Engineering, Mie University

*2 静岡理科大学大学院理工学研究科
Graduate School of Science and Engineering, Shizuoka Institute of
Science and Technology

置（時間）がより一致するよう位置合わせを行うことで、類似する音韻関係を抽出しやすくなることが期待される。また、こうした前処理には音声の時間情報が必要であることから、どの位置（時間）にどれだけの波形が含まれているかを解析可能な Wavelet 解析に本論では注目し、Wavelet 解析と SIFT 特徴量を用いて音韻の位置合わせを試みる。

3. Wavelet 解析と SIFT 特徴量を用いた音韻情報の位置合わせ

Wavelet 解析とは時間的周波数解析の 1 つであり、どの地点にどれだけの波形が含まれるかを解析する [4]。本論では異なる言語の音に Wavelet 解析を行い、スカログラムに出力した。

$$\Psi(t) = \frac{1}{\sqrt{a}} \Psi\left(\frac{x-b}{a}\right) \quad (1)$$

$$w(a, b) = \int_{-\infty}^{\infty} x(t) \frac{1}{\sqrt{a}} \Psi\left(\frac{x-b}{a}\right) dt \quad (2)$$

ここで、 $\Psi(t)$ は Mother Wavelet 関数を意味し、 $x(t)$ は元の波形、 a 、 b はそれぞれ $\Psi(t)$ の拡張量、平行移動量を表す。

SIFT 特徴量とは、画像処理において用いられる特徴量抽出法の 1 つとして画像の類似検索などに利用され、画像の回転や拡大縮小に強い特徴量である [5, 6]。Wavelet 解析で音韻の時間情報を含むスカログラムを出力したとき、SIFT 特徴量による対応点は時間情報を有している。したがってこの対応点を利用することで「音韻情報の位置合わせ」が可能か本論文で検討した。

4. 実験

本論文では 2 つの実験を行った。

実験 1： 人工データにおいて、Wavelet 解析と SIFT 特徴量によるアプローチが有効かの検証

実験 2： 異なる言語間において、Wavelet 解析と SIFT 特徴量によるアプローチが有効かの検証

Wavelet 解析と SIFT 特徴量によるアプローチでは、Wavelet 解析による時間的周波数解析を行い、出力されたスカログラムに SIFT 特徴量を用いて対応点を紐づけた。この対応点の座標は時間情報を含むことから、対応点の差分（特徴点の横軸上距離）の平均を取ることで、波形全体のズレを算出できる。そこで、Wavelet 解析と SIFT 特徴量の利用が、「音韻情報の位置合わせ」に有効な手段かを上記 2 つの実験を通して確認した。なお、実験で用いた音声データは表 1 に従って単一話者によって録音されており、この単一話者は英語を母国語とする英語と日本語の男性バイリンガル話者である。

4.1 実験 1：人工データによるズレの検証

実験手順を図 2 に示す。表 1 の下、録音された各英語音素（46 種類）の前後 0.1 秒を削除し、前に 0.2 秒の無音を追加した音声、後ろに 0.2 秒の無音を追加した音声の 2 パターンを用意した。これにより同じ音声異なる時間帯に録音されている状況が作られ、この状況の下で Wavelet 解析と SIFT 特徴量による位置合わせが可能かを確認した。

Wavelet 解析と SIFT 特徴量を用いた実験結果の一部を図 3

図 2 実験手順

図 3 同言語間（英語音素）の SIFT 特徴量マッチング

図 4 「あ」と/A/の SIFT 特徴量マッチング

（縦：周波数、横：時間）に示す。図 3 では英語音素 /eɪ/（二重母音）を入力データとしており、このデータは図 2 に従って音韻情報の録音位置が異なるものである。各英語音素（46 種類）でも同様の実験を行った結果、得られた特徴点数（対応点）の平均は 25 個であった。波形全体のズレを算出したところ、英語音素（46 種類）において、波形全体のズレと仮定した値の平均は 146 pixel（0.2085 秒）であった。1 秒音声を Wavelet 解析して得られた画像（スカログラム）は 700 pixel であることから、実験条件による 0.2 秒のズレは 140 pixel である。若干の誤差を許容するならば、146 pixel という結果はおおよそ波形全体のズレ（0.2 秒のズレ）を検出できていると考えられる。したがって、図 1 のような音韻情報のズレの補正を目的としたとき、Wavelet 解析と SIFT 特徴量によるアプローチは、同言語間では有効であると考えられる。

4.2 実験 2：異なる言語におけるズレの検証

同言語間で行った実験結果より、Wavelet 解析と SIFT 特徴量を用いた音韻情報の位置合わせは可能であると考えられた。そこで、異なる言語間として英語、日本語間で似た音声（「あ」と/A/）を用意し、Wavelet 解析と SIFT 特徴量による音韻情報の位置合わせを試みた。なお、実験 2 で用いた音声データは実験 1 と同様にして、単一話者によって録音された日本語（音節）と英語（音素）音声である。

実験結果の一部を図4（縦：周波数，横：時間）に示す。図4では日本語音節「あ」と英語音素/ʌ/を入力としており，対応点は1点であった。実験1に比べ，実験2で得られた特徴点数の平均は5個と少ないことから，対応点の差分の平均を用いた位置補正は困難である。また，マッチングした特徴点数が少ない原因として，話者の発話速度が考えられる。話者の発話速度とは，例えば，同じ文言でも，早口で話すときとそうでないときは話す時間に違いが生じる。Wavelet 解析は，時間的周波数解析の一手法であり，本提案では，こうした話す速度の違いを考慮していないことが，SIFT 特徴量によるマッチングに影響を与えた1つと考えられる。今後としては，話者の発話速度も考慮した手法によるマッチングについても検証の必要があると考えられる。

5. まとめと今後

似ていると判断される音声データにおいて収録される地点が異なると MFCC に影響を及ぼし，出力結果に正しいものが得られていない可能性があった。そこで，似た音声における特徴的な部分が含まれる地点を合わせることを目的に「音韻情報の位置合わせ」を考えた。実験1では Wavelet 解析と SIFT 特徴量を用いた結果で対応点が多く得られ，それら対応点の差分の平均を用いることで音韻のズレの補正が可能と考えられた。この知見を元に，異なる言語間でも Wavelet 解析と SIFT 特徴量による実験2を行った結果，対応点は減少し，この要因として考えられることが，話者の発話速度の違いであった。この違いは Wavelet 解析の結果に影響を及ぼしたと考えられる。今後は話者の発話速度の違いに対応するような Wavelet 解析を検証する必要がある。

参 考 文 献

- [1] 川原繁人: ビジュアル音声学, 三省堂, 2018.
- [2] 松下真也, 高野敏明, 友次克子: “単一話者による英語音素と日本語音節の対応関係の一考察,” 第47回東海ファジィ研究会予稿集 (蒲研 2019), 8月6-7日, 蒲郡, pp. 12-1-12-3, 2019.
- [3] 松下真也, 高野敏明, 友次克子: “Wavelet 解析と SIFT 特徴量を用いた音韻の位置合わせ,” 第48回東海ファジィ研究会予稿集 (ヒマ研 2020), 2月17-18日, 南知多, pp. P1-01-1-P1-01-2, 2020.
- [4] 金谷健一: これなら分かる応用数学教室—最小二乗法からウェーブレットまで—, 共立出版, 2003.
- [5] R. C. Gonzalez and R. E. Woods: *Digital Image Processing*, 4th Edition, Pearson Education, 2018.
- [6] 画像情報教育振興協会: デジタル画像処理, 画像情報教育振興協会, 2015.

(2020年4月11日 受付)

(2020年4月19日 採録)

[問い合わせ先]

〒514-8507 三重県津市栗真町屋町 1577

三重大学大学院工学研究科

松下 真也

E-mail: 420d052@m.mie-u.ac.jp

— 著 者 紹 介 —

まつした しんや
松下 真也 [非会員]

2018年静岡理科大学情報学部情報デザイン学科卒業。2020年同大学大学院理工学研究科修士(理工学)。現在、三重大学大学院博士後期課程システム工学専攻に在学中。

たかの としあき
高野 敏明 [正会員]

2008年三重大学工学部電気電子工学科卒業。2010年同大学院工学研究科博士前期課程電気電子工学専攻修士(工学)。2013年同大学院博士後期課程システム工学専攻修士(工学, 三重大学)。2013年立命館大学情報理工学部特任助教。2016年静岡理科大学情報学部講師を経て, 2020年同学部准教授として現在に至る。機械学習, 特に強化学習の研究に従事。ファジィ学会, 人工知能学会, 日本ロボット学会, IEEE 各会員。

ともたけ かつこ
友次 克子 [非会員]

1985年同志社大学文学部英文学科卒業。1987年同大学院文学研究科博士前期課程修士。1990年同博士後期課程満期退学。1991年静岡理科大学理工学部講師, 2001年助教授, 2015年情報学部教授。現在に至る。認知言語学, 特にコーパスに基づく構文研究に従事。日本認知言語学会, 日本英語学会, ICLA 各会員。

Phoneme Positioning Using Wavelet Analysis and SIFT Features

by

Shinya MATSUSHITA, Toshiaki TAKANO, and Katsuko TOMOTSUGU**Abstract:**

Linguists transcribe minor languages usually relying on their ears. This manual work requires time-consuming efforts and elaborate phonetic notation. In this study, we aim to construct a support system for fieldwork linguists documenting minor or low resource languages. In the previous preliminary experiments, we investigated sound matching between two languages using MFCC, but the correspondence accuracy was below expectations due in part to a lack of preprocessing data. Therefore, in this paper, we have worked on audio preprocessing using Wavelet analysis and SIFT features.

Keywords: audio processing, voice recognition

Contact Address: **Shinya MATSUSHITA**

*Graduate School of Engineering, Mie University
1577 Kurima-machiya, Tsu, Mie 514-8507, Japan
E-mail: 420d052@m.mie-u.ac.jp*