GS1-4-5

母音のフォルマント特性に基づいた話者判別

Speaker distinction based on formant properties of Japanese five vowels

○ 酒井香里（沼津高専専攻科）　　鄭萬溶（沼津高専）

Kaori SAKAI, Numazu National College of Technology
Man-Yong JEONG, Numazu National College of Technology

Abstract: Recently, voice recognition is applied in smartphone or car navigation system. However, there are many problems about recognition of speaker and understanding speaker's feelings in the voice recognition technique based on STFT, FFT, cepstrum analysis. Therefore, in this study, Wavelet analysis which is able to analysis voice signal in the time-frequency domain, is adopted in order to make sure the recognition rate of the speaker. On the other hand, vowel formant identified by the vocal tracts of speaker can be evaluated by using LPC because the vocal tracts of speaker are different form each other. In this study, the vowel formants, which may show the transfer characteristics of speaker's vocal tract, will be obtained by applying the linear prediction method to Wavelet's approximation and detail components for each level. In particular, the ingredient and [4 3] can be a significant index to the speaker specification through the examinations.

Key Words: Wavelet Packets analysis, Linear Predictive Method, Speaker specification

1. 緒　論

近年、医療ロボットや介護ロボットなどの開発に伴い、人間にとって最も自然で円滑なコミュニケーション手段である、音声を利用した Human Machine Interface への関心が高まってきている。また、機械の高性能化・多機能化が進む反面、操作が難しいという問題も生じている。このような場面で音声はもっとも強力なコミュニケーションツールとなる。一般に音声には、意味成分・個人成分・感情成分の 3 つの特徴成分が含まれており、現在のスマートフォンやカーナビゲーションなどに数多く見られるように、意味成分の抽出技術の発展は目覚ましく実用化も進んでいる。しかし、個人成分と感情成分の抽出については未だ研究段階であり、多くの課題が残されている。音声に含まれるこれらの成分を特定できるようになれば、さまざまな場面で応用ができ、音声による Human Machine Interface 技術の適用範囲を拡張でき、さらにはセキュリティシステムなどにも応用できる。

一方、Wavelet 解析は時間－周波数解析はもちろんのこと、基底関数の相似性から対象データに含まれている相似的特徴を抽出することができるため、FBI の指紋認証データベース、JPEG や MPEG4 など幅広い分野で利用されている。また、その性質から、非定常波を含む音声認識の解析分野でも有用なツールとして検討されている。

話者判別に関するこれまでの研究では、特性の話者から発せられた複数の音声信号を照合して同一話者であるかどうかを判別する手法と、母音の基本周波数と高次フォルマント、または子音の周波数特性に注目した手法が提案されている。

本研究は、音声に含まれる話者の特徴成分を特定し、話者判別の技術を確立するため、Wavelet 解析と、音声信号からその共振特性を推定できる、線形予測法を用いて、音声信号から個人成分を抽出し、話者を特定する指標を発見することを主な目的とする。

2. 解析手法

本研究では、音声信号から個人成分を抽出するため、Wavelet Packets 解析と線形予測法を用いることにする。本章では、この二つの手法を使用する理由やそれらの手法の特徴について述べる。

2-1 Wavelet Packets 解析

Wavelet 解析には連続 Wavelet 変換と離散 Wavelet 変換があり、離散 Wavelet 変換では、Wavelet 関数を用いて信号は低周波の近似成分と高周波の詳細成分に分けられる。そして、その近似成分は、それ自身をさらに次のレベルの近似と詳細成分に分ける。この操作を繰り返すことを多段階分解と呼び、この多段階分解を使用することによって音声信号をそれぞれの周波数帯域の時系列波形に分解することができる。Wavelet Packets 解析は離散 Wavelet 変換の一つの概念であり、近似成分と同じように詳細成分も分解することで、Fig. 1 に示すように、信号をより細分化することができる。

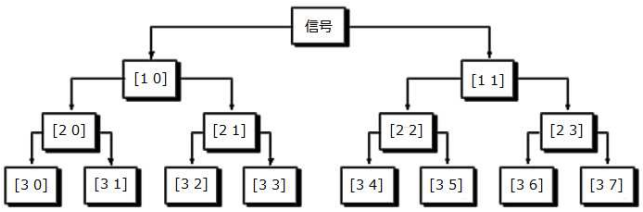

Fig. 1 Wavelet Packets analysis tree by the 3rd-level

また、聴覚器官の音を感知する部位である基底膜は、手前から高周波成分、奥に行くに連れて低周波成分と周波数によって振動する場所が異なる。ここで周波数ごとに分け電気信号として脳に伝達し周波数解析を行っていると考えられている。これは、基底膜上のある一点に着目すると、バンドパスフィルタと見なすことができる。今回使用する、Wavelet Packets 解析では、元の音声データを Wavelet Packets 変換し、高周波成分と低周波成分を分離していき、多分解解析を行うことで、個人の特徴が出る周波数成分を特定する。これは聴覚器官の周波数解析と類似しているため、人間がどの周波数成分に着目して個人を特定しているかの解

析システムとしての使用が考えられる。

2-2 線形予測法

人間の声道には個人差があると言われており、その個人差を測る 1 つの指標となるのが声道の共振特性であるフォルマント構造の推定である。一般に、音声のフォルマント構造を求めるのには線形予測法が用いられている。

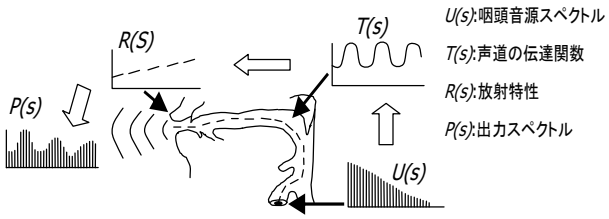

Fig. 2 Sound source filter theory

$$P(s)=U(s)T(s)R(s) \qquad (1.1)$$

線形予測法とは音源フィルタ理論に基づく音声生成モデルを用いてこのモデルを構成する特徴パラメータを抽出するため、スペクトル解析を行ない、得られたスペクトルから声道の共振特性を意味するスペクトル包絡を計算するというものである。一般には、これにより音声信号のスペクトル包絡を算出し、その共振ピーク（フォルマント）などの情報から母音の種類を判別する。本研究では、これらのフォルマントの高次に着目し、話者の特徴成分として捉えることにする。

3. 聴取試験による話者判別

実際に人間がどの周波数成分に着目して個人を特定しているかを調べるために、Wavelet Packets 変換によって日本語の 5 母音を低周波成分、高周波成分と枝状に分離していき、多分解（多段階）解析を行う。もとの音声信号から徐々に高周波成分を取り除いていき、低周波成分である残りの信号を再生し、話者の聞き分けを行い、どの成分が話者判別に有効な成分であるかを調査する。以後この低周波成分を近似成分、高周波成分を詳細成分と呼ぶことにする。

3-1 実験方法

実験では、男女各 3 人、計 6 人から USB マイク Podcaster を使用して日本語の 5 母音を採取した。このとき、サンプリング周波数 44100Hz とし、Wavelet Packets 変換においてマザーウェーブレットは db10 を使用した。また、Wavelet 解析の分解レベルを 6 として分解木を展開した。Wavelet Packets 変換によって音声信号を多段階の近似と詳細成分に枝状に分解した後、各成分を再生して、どの段階まで個人の判別ができるかを聞き分け調査で調べ、その結果について考察を行った。

3-2 実験結果の考察

まず各成分ごとの聞き分け調査を行い、その結果を Table1 に示す。なお、Wavelet Packets 変換のレベル 1 の近似成分を[1 0]、詳細成分を[1 1]とし、[1 0]のレベル 2 の近似成分を[2 0]、詳細成分を[2 1]とする。また、レベル 1 の詳細成分である[1 1]のレベル 2 の近似成分を[2 2]、詳細成分を[2 3]と表記する。

Table1 に示すように、[1 0]成分では、意味と話者の両方が特定でき、音声が元データよりクリアになった。[2 0]成分では、音声に少しの違和感はあるものの、話者と意味の両方を特定できた。[3 0]成分では、意味は判別できるが、話者を完全に特定できなくなった。以上の結果により、人間が話者を特定する際に必要としている成分が、レベル 2 の[2 0]とレベル 3 の[3 0]成分から取り除かれていることが確認できた。したがって、その成分は[1 0]と[2 0]成分の詳細成分である[2 1]と[3 1]成分に含まれていることがわかった。また意味に関しては、話者判別が不可能になる[3 0]成分以降の近似成分でも認識可能だったため、過去の研究が明らかにしているように、低周波成分が意味を表していることが確認できた。

Table 1 Listening test for meaning and speaker identification

five vowels	meaning	speaker
[1 0]	○	○
[1 1]	×	×
[2 0]	○	○
[3 0]	○	×
[4 0]	○	×
[4 1]	△	×
[5 0]	△	×
[6 0]	×	×

4. Wavelet Packets 変換を用いた母音解析

ここでは、Wavelet Packets 変換によって日本語の 5 母音を近似・詳細成分と分解していき、線形予測法を用いてスペクトル包絡を得て、どの周波数帯域の形状に個人差が現れるかを調査し、母音のスペクトル包絡の違いから個人判別の可能性を探る。

4.1 実験方法

3 章で扱った、日本語 5 母音の音声データを Wavelet Packets 変換し、それぞれの周波数帯域に分解し、線形予測法を用いてスペクトル包絡を得る。それらの形状を比較・考察する。ここでは、比較的に高い周波数帯域の共振特性も現れるようにするため、線形予測法の次数を 16 と通常より高く設定した。

4.2 実験と解析結果の考察

それぞれ各 5 回録音した 5 母音をレベル 1 の[1 0]成分からレベル 3 の[3 7]成分までの包絡線を描くと、レベル 3 の[3 1]と[3 3]成分でその波形に違いが表れた。男性 3 人から採取した 5 母音を Wavelet Packets 変換し、レベル 3 の信号を線形予測法に適用して得た、スペクトル包絡線を Fig. 3 に示す。

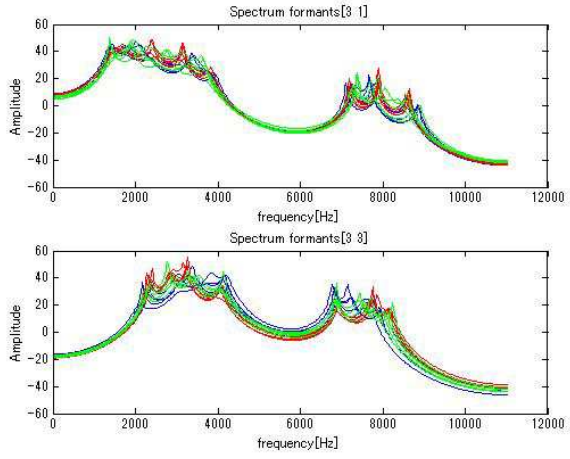

Fig. 3 Formant structures extracted from DWT of level 3 for vowel /a/ of three male adults

Fig. 3 は上から順に男性 3 人の母音/a/のレベル 3 の[3 1]成分と[3 3]成分を示し、同一話者は同じ色で表示している。図に示されているように、同一話者同士ではほぼ同じ曲線となり、別話者同士ではその違いがはっきり表れている。本論文には示さないが、この違いは男女計 6 人のそれぞれの比較で明確に示されていた。この波形の違いは声道の形状の違いによる共振特性が関係している。これらの結果から、[3 1]成分と[3 3]成分に何らかの個人成分が含まれていることが確認できた。

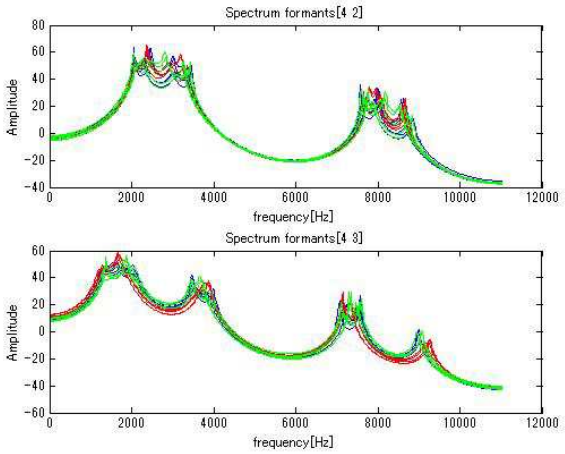

Fig. 4 Formant structures extracted from DWT of level 4 for vowel /a/ of three male adults

この結果を踏まえ、より狭い帯域で個人成分を抽出するために[3 1]と[3 3]成分をさらにレベル 6 まで分解した結果を Fig. 4 に示す。[3 3]成分を分解した[4 6]成分と[4 7]成分では波形の形状の差は見られなかった。そして、[3 1]成分を分解した[4 2]と[4 3]成分では波形の形状に違いが見られた。しかし、レベル 5・6 では見られなかった。したがって、[3 3]成分に関しては、分解レベル 3、[3 1]成分に関しては、分解レベル 4 が適当であると分かった。これらの成分は、3 章の聴取試験で注目した、[2 1]成分のレベル 3 の詳細成分である[3 3]成分と[3 1]成分のレベル 4 の詳細・近似成分である[4 2]と[4 3]成分であるため、より狭い帯域で個人成分を抽出することができたと考えられる。

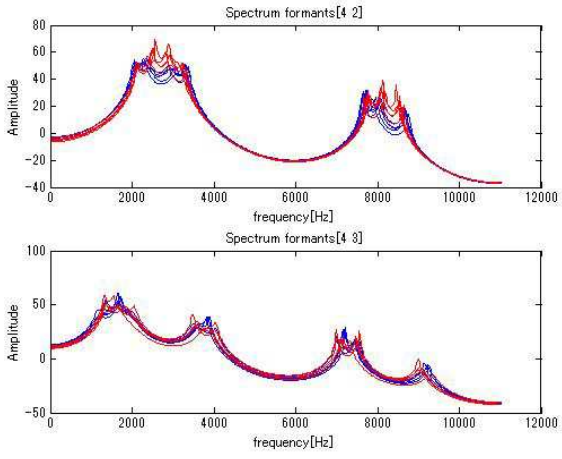

Fig. 5 Formant structures extracted from DWT of level 4 for vowel /i/ of male adults A and B

5. フォルマントに着目した評価

前節の別話者間での波形の違いをより明確に視覚化するためにフォルマントに着目して評価する。

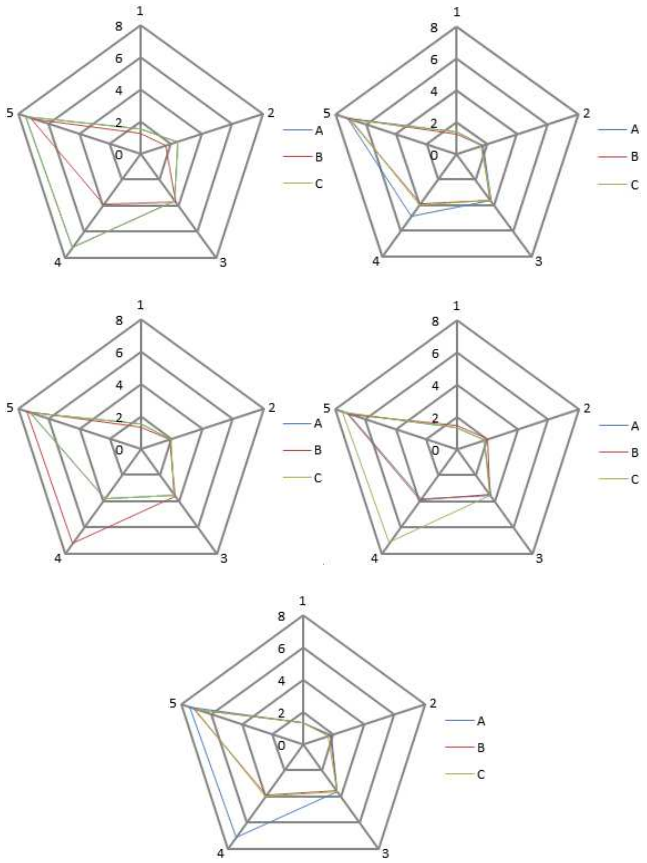

Fig. 6 Formants structures of five vowels of the ingredient [4 3] for speaker A, B and C

線形予測法で描いた包絡線のピーク値を取る。周波数の小さい順に第 1 、第 2、第 3、第 4、第 5 フォルマントと呼び、5 つのフォルマントで 5 角形を描き別話者間での違いを 5 角形の形に表現し、どのように現れるかを評価する。これは、すべての話者の各成分において、ピーク値の数の平均をとると、5 つのフォルマントが安定的に抽出されることからこの数を採用した。まず Fig. 6 に上から男性 3 人の[/a/,/i/,/u/,/e/,/o/]の 5 母音(それぞれ 3 回平均)を分析した結果から得たフォルマントにより形成された 5 角形を示す。[4 2]成分では 3 回の同一話者の再現性はあったものの、別話者との違いをうまく表現できなかったため、この評価法の指標として向かないことがわかった。一方、[4 3]成分では、Fig. 6 に示されているように、別話者間での差がはっきりと表れた。これは 3 回平均を表したものであり、これを通して同一話者間での再現性も確認できた。[3 3]成分も同様に評価した結果、別話者間での 5 角形の形に違いが現れた。また、5 母音をそれぞれ比較してみると、各母音によって特徴が顕著に現れる話者が異なるため、[3 3]と[4 3]成分の 5 母音の 5 角形の形状の組み合わせによって話者が特定できると判断される。

6. 環境ノイズに対する評価

話者判別技術を実用化するためには、一般的の想定できる環境ノイズを考慮し、環境ノイズが話者判別に及ぼす影

響についても把握しておく必要がある。ここでは、意味判別に関係する低次フォルマント、すなわち、低周波領域の環境ノイズは考えずに話者判別と関係する比較的高い周波数領域に対して、学校の室内環境において、環境ノイズの影響を調べることにする。話者判別において重要となる、[3 3]と[4 3]成分は、比較的高周波成分のため、同じ周波数領域での環境ノイズに対する評価は必須となる。音声と同じ手順で解析し、録音時に常時どのくらい高周波成分の環境ノイズが含まれているかを調査した。

6-1 実験方法

音声を録音する時と同じ環境下で、サンプリング周波数 44100Hz で環境ノイズを録音し、音声と同じ手順で解析し 5 角形の形状を比較する。実験場所は研究室とし、在室の人々には、静粛にするように依頼した。

6-2 実験と解析結果の考察

まず、マイク録音時に必ず入ってしまう環境ノイズの[3 3]と[4 3]成分を Fig. 7 と 8 に示し、フォルマント構造の再現性を評価した。

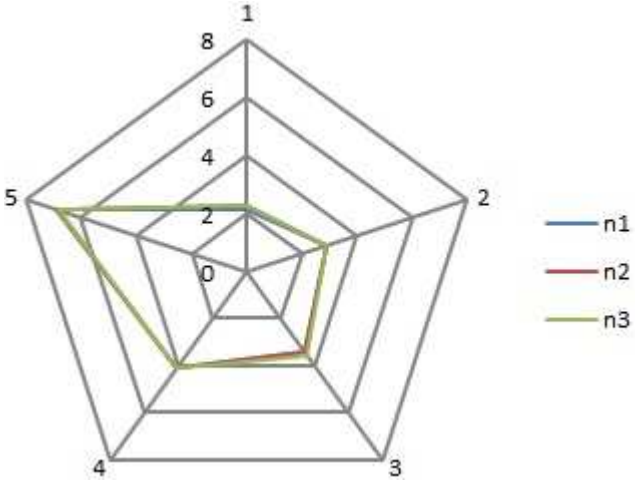

Fig. 7 Environmental noise of the ingredient [3 3]

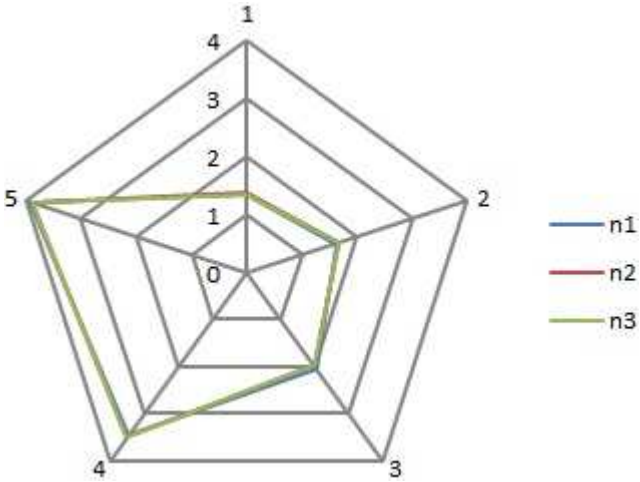

Fig. 8 Environmental noise of the ingredient [4 3]

Fig. 7、8 に示されるように、3 つの無音声ノイズデータを比較すると、ばらつきはなく、ほぼ一定を保っていることがわかる。つまり、[3 3]と[4 3]成分には男女 6 人とも同じだけの環境ノイズが入っており、[3 3]と[4 3]成分ではノイズが話者判別に影響しないことが明らかになった。しかし、本論文には示さないが、[3 3]成分と各母音の五角形を比較してみると、ノイズとそれぞれの話者のフォルマントの値が近く、形状が類似している母音があるため、この評価法では個人成分の差が表れにくくなることが考えられる。3 章で示されるように、[2 1]成分は個人の特定に多少の違和感を与える程度で、[4 3]成分より含まれる個人成分が少ないため、[3 3]成分はノイズとの分離を考える必要がある。一方、[4 3]成分では各母音の五角形とノイズの値には大きな差があり、十分に個人成分を含んでいることがわかる。これらの結果より、一般的な室内環境において[4 3]のフォルマント特性は個人の特徴を判別するための指標にできると考えられる。

7. 結　論

本研究では、Wavelet Packets 解析を用いて母音のフォルマント特性に基づいた話者判別について検討を行い、以下のような結論を得た。

(1) Wavelet Packets 変換によって音声信号を多段階分解すると、[2 0]成分で話者が少しわかりにくくなり、[3 0]成分で話者の特定が完全に不可能になることが確認できた。また、各成分の詳細成分である[2 1]と[3 1]成分に個人の特徴成分が含まれており、それらの成分に着目することにより話者判別が可能であることを証明できた。

(2) Wavelet Packets 変換によって得られた各成分に線形予測法を適用し、包絡線を描くと、[3 1]成分と[3 3]成分に個人差が確認でき、話者判別の指標として使用できることを明らかにした。

(3) Wavelet Packets 変換によって得られた各成分に線形予測法を適用することにより母音のフォルマントの伝達特性を抽出でき、その違いから話者判別が可能であることが確認できた。特に[4 3]成分は話者の特徴成分となる重要な指標と成りうることを明らかにした。

今後の研究としては、単母音だけでなく、多音節からの母音の抽出による比較も行い、基本周波数の変化と[4 3]成分の比較も行いたい。そして、より個人の差が目で見て確認できるように、よりよい評価法の確立を目指していきたい。また、Wavelet Packets 解析は使うフィルタの閾値で分けられる成分が変化するので、話者判別に適した閾値を調査していく必要がある。

参考文献

(1) 杉山　昂太郎, 鄭萬溶,「ウェーブレット解析を用いた話者判別手法に関する検討」,平成 23 年度専攻科学習成果レポート

(2) 富士原　照久, 鄭萬溶,「Wavelet 解析を用いた個人成分抽出に関する検討」,平成 19 年度専卒業研究発表会報告書集(0708)

(3) 田中友基, 鄭萬溶, “音声認識に関する基礎的研究”D&D2005, 朱鷺メッセ(新潟), No.329, 2005.

(4) 石川陽平, 田中友基, 鄭萬溶, “Wavelet 解析を用いた音声認識に関する基礎的研究”, D&D2006, 名古屋大学, No.313, 2006

(5) Mohadese Eshaghi, M.R. Karami Mollaei,” Voice activity detection based on using wavelet packet”, Digital Signal Processing 20 (2010) 1102-1115