様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　２０　日現在

機関番号：　１２４０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０１８７

研究課題名（和文）気導音声と骨導音声を同時利用する基本周波数抽出法の確立とその話者認識への応用

研究課題名（英文）Fundamental Frequency Detection Utilizing Air and Bone Conducted Speeches and Its Application to Speaker Recognition

研究代表者

島村　徹也（SHIMAMURA, TETSUYA）

埼玉大学・情報メディア基盤センター・教授

研究者番号：４０２３５６３５

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：本研究では、雑音環境下、特には高騒音環境下においてさえも高い信号雑音比が確保できることから、骨伝導マイクロホン(骨導マイクロホン)の利用に着目する。骨導マイクロホンから収録された音声、骨導音声、を利用し、通常の気導マイクロホンから得られる音声、気導音声、と併用し、新しい基本周波数抽出の方法を導出した。また、音声の基本周波数を利用する話者認識アルゴリズムを導出することにも成功し、雑音環境下での高精度な話者認識への一方向性を示した。

研究成果の概要（英文）：In this work, we focus on the use of bone-conducted microphone, since it provides high signal-to-noise ratio speech. The speech obtained through the bone-conducted microphone is called bone-conducted speech. The bone conducted speech is combined with that obtained through air-conducted microphone, air-conducted speech, and a new fundamental frequency detection algorithm is derived. A specific algorithm for speaker recognition, in which the fundamental frequency of speech is used, is also derived. One direction for speaker recognition in noisy environments is suggested.

研究分野：　ディジタル信号処理

キーワード：　骨伝導

１．研究開始当初の背景

個人認証のために、音声の個人差を用いて誰の声であるかを自動的に判定する技術は特に話者認識と呼ばれ、これまでの研究で無雑音環境下においては、98%を越える、理想状態に近いレベルまで達していることが知られている。しかしながら、実環境としての雑音環境下においては、認識率が 50%以下になるなど大幅に低下してしまうことがよくある。そこで、a) 雑音付加音声に雑音低減を施し、信号対雑音比(SNR)を改善する方法や、b) 雑音にも強靭な話者の特徴パラメータの利用、c) 音声の変動を表現する統計モデルの工夫、などが検討されている。しかしながら、a)は低減する雑音量を増大すると、元々の音声の音質が劣化されてしまい、受け入れられる認識率が達成できない。また、b)と c)は、話者認識アルゴリズムとして、特に両者を統合する形で研究が進められているが、多くの計算量を必要とし高い認識率を得るか、計算量を妥協して低い認識率に止めるかのようなトレードオフの関係になり、確約されるアルゴリズムが存在していない。

２．研究の目的

本研究は、音声を利用する個人認証システムの質的向上および利用環境の拡大を目指し、従来困難とされてきた雑音・騒音環境下での高精度な基本周波数抽出を実現し、それを認証システムに組み入れることで、実環境に頑強な認証システムを構築することを目的とする。

３．研究の方法

これまでに特に、気導マイクと骨導マイクを同時に利用し、気導音声の高周波数成分と骨導音声の低周波数成分を加え合わせることで、良質な音声が得られる知見が得られている。本研究では、この組み合わせのアイデアを音声の基本周波数(基本周期の逆数)の抽出問題に発展する。骨導マイクと気導マイクを準備し、それらを同時に利用し、骨導音声と気導音声を同時に取得する。そして、雑音環境下でのそれぞれの音声信号の特徴を考慮して、それぞれに基本周波数抽出を施し、得られる特徴量を組み合わせることで、より雑音に耐性のある結果を得る。また、得られた基本周波数を話者の特徴パラメータとして利用し、雑音・騒音環境下においても高精度で安定した認識結果を与える話者認識システムを構築する。そして、得られたシステムの有効性を実験的に検討する。

４．研究成果

主な研究成果は５.発表論文の(2)(3)であるが、ここでは特に(3)の概要を述べることにする。

<提案法の構成>

深層ニューラルネットワークを用いた話者認識において、振幅スペクトルに加え、音声の基本周波数から生成した調波バイナリベクトルを入力に加える手法を提案する。基本周波数が雑音環境下でも正確に推定されれば、このバイナリベクトルは雑音に頑強な特徴量であると考えられる。これを用いることにより、ネットワークへの入力が雑音に対して少ない変動となり、かつ個人性を保持したものであるため、雑音環境下での精度の向上に貢献すると期待できる。学習処理は信号の切り出し、基本周波数推定、調波バイナリベクトル生成、対数振幅特性の計算、ネットワーク計算、の 5 行程から成る。認識処理は入力ベクトル生成、ネットワーク計算、出力ベクトルのユニットごとの加算、認識結果の出力、の 4 行程から成る。

<調波バイナリベクトル>

基本周波数は、音声の周期性を表すとともに個人性が表れる特徴量とされる。この基本周波数を深層ニューラルネットワークに効果的に導入する方法を考える。基本周波数推定には，計算が容易でかつ雑音に対しても比較的強いとされる自己相関関数法を考える。具体的には、自己相関関数法を改良した重み

付き自己相関関数法を用いる。重み付き自己相関関数の最大ピーク位置から基本周期が求まり、その逆数から基本周波数が求まる。ネットワークの基本周波数に対する解釈を補助するため、基本周波数の値そのものを入力せず、振幅スペクトルと同様の尺度を持つ入力ベクトルを与える。具体的には、スペクトルの構造に注目し、振幅スペクトルと同次元のベクトルで、基本周波数とその倍音の周波数ビンを 1 とし、それ以外の成分を 0 とするものを用いる。

<学習方法>

ネットワークを学習させるための入力ベクトルは、音声から窓関数によって切り取られた短時間フレーム信号から生成される。フレーム信号に対して離散フーリエ変換して得られた振幅スペクトルと、フレーム信号の基本周波数から生成される調波バイナリベクトルを結合して入力ベクトルとする。 このとき，振幅スペクトルは対数尺度に変換する。母音はパワーが大きく周期性を持つことから話者の識別に有用だと考えられる。よって、信号のパワーとゼロ交差率から判別した母音区間のみを入力の対象フレームとする。そして、得られた入力ベクトルをネットワークに入力させ、誤差逆伝播法によってネットワークの重みを更新する。

<認識方法>

学習時と同様にフレーム単位で入力ベクトルを生成し、ネットワークから出力を得る。この際、ユニットの出力は各フレームにおける各話者の尤度を表している。これを対象となる音声の全フレームに対して求め、それらの尤度をユニットごとに加算し、最大となるユニットから話者を判定する。また学習の際と同様に、信号のパワーとゼロ交差率から判別した母音区間のみを入力の対象フレームとする。

<実験>

データベースには CSTR VCTK Corpus を用い、話者 10 名を今回の実験に利用した。各話者とも 200 音声を用意し、学習には 180 音声、テストには発話区間が比較的長い 20 音声を使用した。 本実験では学習にクリーン音声を、テストに雑音混入音声を用い、未学習の雑音に対するシステムの頑健性を確認する。テスト音声に付加する雑音はホワイトノイズ、ピンクノイズ、バブルノイズ、カーノイズ、ファクトリーノイズの 5 種類とし、それぞれの雑音に対して信号対雑音比(SNR)を 0dB、5dB、10dB、15dB に設定した。ネットワークのサイズは、本来の目的であるクリーン音声に対する認識の精度がより高くなるように選択する。事前実験の結果、中間層を 4 層、各層で 128 ユニットとなるよう設定した。なお，隣接層間のユニットが全結合した順伝播型ネットワークとした。出力層は対象話者の人数と同じ 10 のユニットを持ち、各ユニットが対応する話者の尤度を表す。入力には 256 次元の対数振幅スペクトルを用いる従来法と、これに調波バイナリベクトルを加えた 512 次元を用いる提案法を比較する。評価には適合率と再現率の調和平均として定義される F 値を用いた。F 値は最大で 1 となり、高い値ほどシステムが高い認識性能を有することを示す。

各 SNR における 10 名の平均 F 値を求めた。それらの結果が表 1 から表 4 に示される。それぞれの SNR において平均して 0.07 ポイント、0.05 ポイント、0.03 ポイント、0.02 ポイントの F 値の改善が見られ、雑音環境下での認識精度の向上が確認できた。特にホワイトノイズ、ピンクノイズに対して効果が確認でき、0dB においてはそれぞれ 0.08 ポイント、0.13 ポイント、5dB においてはそれぞれ 0.07 ポイント、0.08 ポイント、10dB においてはそれぞれ 0.08 ポイント、0.05 ポイント、15dB

においてはそれぞれ 0.03 ポイント、0.04 ポイントの改善となった。 これは広帯域雑音であるホワイトノイズ、ピンクノイズがスペクトル形状の大きな変化をもたらす一方、提案法では調波バイナリベクトルが入力ベクトルの変動を抑える役割を果たしたためだと考えられる。

表１ 雑音と F 値（SNR=0dB）

雑音	従来法	提案法
ホワイト	0.32	0.40
ピンク	0.47	0.60
バブル	0.91	0.95
カー	0.96	0.96
ファクトリ	0.65	0.75
平均	0.66	0.73

表２ 雑音と F 値（SNR=5dB）

雑音	従来法	提案法
ホワイト	0.48	0.55
ピンク	0.74	0.82
バブル	0.99	1.00
カー	0.97	0.99
ファクトリ	0.86	0.89
平均	0.80	0.85

表３ 雑音と F 値（SNR=10dB）

雑音	従来法	提案法
ホワイト	0.67	0.75
ピンク	0.91	0.96
バブル	1.00	1.00
カー	0.99	0.99
ファクトリ	0.94	0.96
平均	0.90	0.93

表４ 雑音と F 値（SNR=15dB）

雑音	従来法	提案法
ホワイト	0.89	0.92
ピンク	0.95	0.99
バブル	1.00	1.00
カー	0.99	0.99
ファクトリ	0.99	0.99
平均	0.96	0.98

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１ 件）

(1) M.A. Rahman, Y. Sugiura, T. Shimamura and H.Makinae, “LP-Based Quality Improvement of Noisy Bone-Conducted Speech,” IEEJ Trans. Electronics, Information and Systems, Vol.137, No.1, pp.197-198, 2017.

〔学会発表〕（計２ 件）

(1) S. Zhang, Y. Sugiura, T. Shimamura and H.Makinae, “Fundamental Frequency Estimation Combining Air-Conducted Speech with Bone-Conducted Speech in Noisy Environments,” Proceedings of IEEE International Conference on Electrical, Computer and Communication Engineering, pp.244-247, 2017.

(2) 鈴木良啓, 杉浦陽介, 島村徹也, “雑音に頑強な話者認識のための基本周波数を用いた深層ニューラルネットワーク,” 電子情報通信学会技術研究報告, SP2016-58, pp.53-56, 2016. 12.

〔図書〕（計０ 件）

６．研究組織

(1)研究代表者

島村 徹也（SHIMAMURA, Tetsuya）
埼玉大学・情報メディア基盤センター・教授
研究者番号：40235635

(2)研究分担者

なし