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

令和　元 年　 ６ 月　 ７ 日現在

機関番号：３３９１０
研究種目：挑戦的萌芽研究
研究期間：2016～2018
課題番号：１６Ｋ１２４６１

研究課題名（和文）音声合成法と話者適応技術および編集合成に基づく詐称音声の相対位相情報による検出法

研究課題名（英文）A detection method using relative phase information for spoofed speech based on speech synthesis, speaker adaptation and edited speech

研究代表者
中川　聖一（NAKAGAWA, Seiichi）

中部大学・工学部・教授

研究者番号：２０１１５８９３

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：　話者照合技術の問題点として、本人の音声を詐称するなりすまし音声がある。なりすまし音声としては、(1)声真似、(2)本人の一部の音声を用いた音声合成、(3)本人の音声への声質変換、(4)本人の音声を録音し再生する録音再生、がある。
　本研究では、代表者が発明した相対位相特徴を改善し、上述の(2)(3)(4)によるなりすまし音声を高精度に検出する技術を開発した。改善点は、使用する位相の周波数帯域を高域まで拡大したこと、周波数軸のスケールを最適に伸縮したことであり、これによって、単独特徴パラメータとしては世界最良パラメータとなった。また従来の特徴パラメータと併用することにより更に検出精度が向上した。

研究成果の学術的意義や社会的意義
　生体認証技術の一つとして話者照合技術がある。本研究では、各話者が約40秒の声を登録しておけば、4秒程度発声した声で、270人の話者から99.7%の精度で正しく発声した話者を識別する技術を開発した。この技術で声による「鍵」などの多くの応用が実現できる。
　一方、声真似や本人の一部の声を用いた音声合成技術や声質変換技術、録音再生技術による、なりすまし音声と本人の音声との区別ができなくなる問題が実用化への妨げとなる。本研究では、このなりすまし音声を高精度に検出する技術を開発した。この技術によって、話者照合技術のセキュリティ分野への応用も可能となった。

研究成果の概要（英文）：　A serious problem for speaker verification is spoofed speech, which is classified into (1) mimic speech (impersonation), (2) speech synthesis using target speaker's speech, (3) voice conversion to target speaker's speech, and (4) record-replay speech of target speaker's speech.
　In this study, we improved relative phase information for spoofed speech detection, which was invented by the proposer. The improvement points are the extension of frequency range to higher frequency to extract relative phase and optimal nonlinear scale of frequency axis. We obtained the best feature parameter, that is, improved relative phase, as single feature in the world. Furthermore, we obtained the higher detection rate by combining this relative phase feature and conventional feature parameters.

研究分野：知覚情報処理

キーワード：話者照合　相対位相情報　振幅スペクトラム　位相スペクトラム　詐称音声　再生音　spoofed speech challenge

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

人が発声する音声は、声帯振動に伴う音源特性（声帯振動波形）と音響（声道）器官の形状に伴う声道伝達特性の畳み込みからなる。プロの声帯模写は、音源特性の一部であるアクセント・イントネーションを模倣するのが中心で、声帯の音源波形や声道特性を模倣することは難しい。そのため、声道特性に直結する振幅スペクトラムによって、模倣音声を検出することは比較的容易である。一方、近年の音声合成技術の進展は著しく、テキスト－音声変換技術で任意のテキストから自然な音声の合成が可能となって来た。また声質変換技術の進展も目覚ましい。これらの技術にターゲット話者（本人）の音声データを用いた話者適応技術（何らかの手段で､詐称者がターゲット話者の音声録音の入手が可能と仮定)を組み合わせることによって、声道特性のレベルでターゲット話者にまねた詐称音声を作ることが可能になっている。また、決まり文句を発声する場合は、ターゲット話者の音声を盗聴・録音しそれを再生すれば、本人になりすますことも可能である。このように、声の“なりすまし”犯罪の発生などにより話者照合技術が個人照合技術として使用できなくなる恐れが出てきている。

このような現状に対して、世界共通の spoofed speech のデータベースが構築され、これを用いて各手法の評価が行なわれるようになった。

２．研究の目的

昨今の音声合成技術の進展は著しく、テキストから自然な音声の合成が可能となってきた。これにより、音声合成技術と話者適応技術によって、任意の話者にまねた詐称音声を作ることが可能になり、声の“なりすまし”犯罪の発生などにより話者照合技術が個人照合技術として使用できなくなる恐れが出てきている。本研究では、我々が提案してきた話者照合に有効な相対位相情報を、編集合成音や録音再生音、残響環境下での詐称音声と本人音声との識別にも頑健に働くようにする技術を開発する。この技術は、雑音や残響環境下における話者認識技術の向上にそのまま使用可能となる。

３．研究の方法

（１）当初の研究の方法

本研究は、①編集合成音の音声素片の接合が自然であるかどうかを位相の時間変化を測度として検定する方法、②異なるマイクロフォンや残響の伝達特性に違いによる位相特性の変化の抽出法、を新しく開発することが目的である。そのために、①に対しては、音声波形の１フレームごとや１サンプルごとに我々が提案している相対位相特徴を求め、この時間的変化が周波数帯域ごとに滑らかであるかどうかを検証することによって行う。②に対しては、インパルス応答より求めた伝達特性の位相特性により正規化する方法を試み、実験的にこの手法の問題点を明らかする。次に、伝達特性の違いによる位相特性の正規化を最尤推定法により解折的に行う方法とニューラルネットによる非線形変換で行う方法を開発する。

（２）本研究の方法

当初の研究の方法を見直し､まず､代表者が発明した相対位相特徴の抽出法の改善に取り組む。従来の相対位相は、比較的低域の周波数帯域から抽出してきた。一方、その後の研究で中域の周波数帯域にも話者情報が含まれていることが分かった。そこで、本研究では、中高域の周波数帯域から相対位相を抽出する方法を検討する。また、なりすまし音声の検出に有効として提案されてきた特徴パラメータとの比較、併用に関して検討する。

なお、評価データベースは、世界的共通データベースである Spoofed Challenge 2015 と Spoofed Challenge　2017 のデータを使用する。

４．研究成果

（１）我々は、話者が直接発声した音声と音声合成技術等によって合成された（詐称）音声を識別する技術を開発してきた。我々が提案した相対位相情報を用いた識別手法を IEEE Journal of Selected Topics in Signal Processing　に論文投稿し掲載された。 次の目標であった録音再生音によるなりすまし音声の検出については Eurasip Journal on Audio, Speech and Music Processing に論文投稿し採録された。また、雑音重畳により話者識別精度が劣化するため、雑音重畳音声の振幅スペクトラムと位相スペクトラムからクリーン音声の振幅スペクトラムと位相スペクトラムをディープネットワークを用いて復元する手法を提案した。これにより話者識別精度が改善できることを示した。また、話者認識の基本技術として、音声の残差波形の相対位相情報も補間的特徴量として有効なことを示した

（２）我々が提案した相対位相情報を、ある人の声になりすます詐称音声の識別に利用する研究を世界共通データベースを用いて進めてきた。今までに、相対位相情報が、音声合成技術や音声変換技術により作成された詐称音声と原音声とを識別することに有用であることを示してきた。 表 1 に、その結果を示す[Wang, 2017]。表１より、本提案法が優れていることを示し

ているが、同時に、S10 の編集合成音声に対しては、元の音声を繋いで合成しているので、相対位相特徴では直接音声との識別は難しいこともわかった。ここで、MFCC はメルケプストラム係数、MGDCC は群遅延ケプストラム係数、RPS は相対位相シフトを示す。

表１　声質変換音声、音声合成音声、編集合成音声による「なりすまし音声」の検出結果
RP, PRSRP, MRP は提案法、　値は、等誤り率[%]を示す。
s1, s2, s5-s9: 声質変換音声、s3,s4: 音声合成音声、　s10:編集合成音声

Features	Known attacks						Unknown attacks					
	s1	s2	s3	s4	s5	Ave.	s6	s7	s8	s9	s10	Ave.
MFCC	0.054	5.621	0.023	0.041	2.651	1.678	3.212	0.475	2.217	0.705	9.373	3.196
MGDCC	0.227	1.157	0.002	**0.000**	1.694	0.616	1.230	0.252	0.236	0.169	15.513	3.480
RPS (result in [17])	0.266	0.169	0.022	0.036	0.144	0.127	0.515	0.008	**0.091**	0.011	43.468	8.819
Cosine phase (result in [31])	—	—	—	—	—	—	—	—	—	—	—	—
RP	0.043	0.158	0.008	0.010	0.456	0.135	0.808	0.147	0.361	0.064	40.338	8.344
PPSRP [26]	**0.000**	**0.025**	**0.000**	**0.000**	**0.025**	**0.010**	0.285	**0.005**	1.179	**0.000**	37.728	7.839
MRP	0.038	0.086	**0.000**	0.004	0.175	0.060	**0.274**	0.061	0.453	0.032	28.757	5.915

次の目標は、録音された音声をスピーカーを通して再生された音声と直接マイクに向かって発声した音声との識別に取り掛かった。つまり、ユーザの生体情報によるセキュリティ手段の一つである話者照合技術に関して、悪意をもってユーザになり済ます録音再生音による詐称音声を検出するために、ユーザが直接マイクに向かって発声した音声と人の声を録音再生した音声との弁別を行う手法を検討した。相対位相情報は、伝達特性の影響を受けることが知られている。録音機器を介在するか否かの違いは伝達特性の違いとして現れるため、相対位相が有効な特徴量であることが予想された。 評価実験の結果、我々の提案する相対位相が、このような詐称音声に有効であることを示した。詳細に分析した結果、録音再生機器の性能が良ければ、録音再生機器を通した音声と原音声との識別が困難なこと、スマートフォンなどの録音再生機器の性能が悪い場合は、識別が容易なことも分かった。これは識別率が、 伝達特性の変動量と大きな相関があるためである。単独特徴量としては、詐称音声の識別には相対位相が最も優れていることを示した。表 2 にその結果を示す[Liu, 2019]。ここで、IMFCC は、MFCC とは逆に高域の周波数分解能を高めた振幅スペクトラムのケプストラム係数、CQCC はコンスタントQケプストラム係数である。位相情報の改良点として、従来は低周波数帯域～中周波数帯域の相対位相情報に限定してきたが、これを高周波数帯域まで拡張したこと、そのため、 相対位相特徴をメル周波数軸に変換し、次元削減を行ったことである（MelRP）。これによっても、識別性能が向上した。さらに、周波数帯域による話者情報の偏在性に着目し、頑健に弁別しやすくなるように周波数軸スケールの伸縮を行った。すなわち、なりすまし音声の検出に有効な帯域（4000Hz～5000Hz）には、周波数分解能を上げるフィルタバンク形状を自動学習する方法の検討を行った(ARP)。同様の手法を振幅スペクトラムについても行い(AFCC)、代表的な特徴パラメータとの比較を行った。表２より、提案法の相対位相が最も優れた特徴パラメータであることがわかる。

表 2　単独特徴パラメータによる録音生成音声の検出結果
RP, MelRP, AFCC, ARP が提案法
Development: 開発データ
Evaluation:　 評価データ
値は、等誤り率[%]を示す

Feature	Development	Evaluation
CQCC	10.35	28.48
MFCC	13.74	34.39
IMFCC	4.83	28.59
MGDCC	25.92	38.10
RP	19.86	25.68
MelRP	10.36	16.03
AFCC	**4.01**	27.80
ARP	9.11	**12.65**

（３）実際には、従来の MFCC や CQCC などと併用すると、さらに識別率は向上した。表 3 に、複数特徴パラメータの組み合わせによる直接音声と録音再生音声との識別結果を示す。提案法である相対位相を用いた特徴パラメータの併用の効果がわかる。

表３　特徴パラメータの組み合わせによる録音再生音声の検出結果
値は、等誤り率[%]を示す

Feature	Development	Evaluation
CQCC+MFCC	10.75	29.33
CQCC+AFCC	3.57	28.02
CQCC+RP	9.06	20.98
CQCC+MelRP	5.02	13.88
CQCC+ARP	2.26	12.58
AFCC+ARP	2.23	11.95
ARP+AFCC+CQCC	2.20	11.43

（４）話者照合の基本技術の開発を検討し、音声の原波形と線形予測分析による残差波形に対して、メル周波数ケプストラム係数(MFCC)と群遅延ケプストラム係数、相対位相情報を抽出し、これらのパラメータの併用による話者照合の改善を行った。相対位相に対しては、使用する周波数帯域を低域から中域までの 60Hz ～2800Hz まで拡大した。その結果、原波形から抽出した特徴パラメータが残差波形から抽出した特徴パラメータよりも良かったが、残差波形にも多くの話者情報 が保存されることが分かった。男性 135 名、女性 135 名、合計 270 名の話者認識において、話者情報の登録に 5 発話の約 20 秒間の音声を用いた場合と 10 発話の約 40 秒 間の音声を用いた場合で、1 発話ごとによる話者認識で、それぞれ約 99.2%と 99.7%の認識率を得ることができた。いずれの場合も相対位相情報が振幅スペクト ラム情報の補完情報として有用であることを実証した。

５．主な発表論文等

〔雑誌論文〕（計 2 件）

① Z. Oo, L. Wang, K. Phapatanaburi, M. Liu, S. Nakagawa, M. Iwahashi, J. Dang, “Replay attack detection with auditory filter-based relative phase features”, Eurasip Journal on Audio, Speech and Music Processing, Accepted, 2019、査読有

② L. Wang, S. Nakagawa, Z. Zhang, Y. Yoshida, Y. Kawakami, “Spoofing speech detection using modified relative phase information”, IEEE Journal of Selected Topics in Signal Processing, Vol.11, No.4, 2017, 660-670, DOI:10.1109/JSTSP.2017.2694139、査読有

〔学会発表〕（計 9 件）

① M. Liu, L. Wang, J. Dang, S. Nakagawa, H. Guan, X. Li, “Replay attack detection using magnitude and phase information with attention-based adaptive filters”, Proc. ICASSP, 2019, 6201-6205, DOI:10.1109/ICASSP.2019.8682739、査読有

② 山本滉己、山本一公、中川聖一、原音声波形と残差波形からの MFCC と相対位相情報による話者認識の比較、電子情報通信学会、総合全国大会、学生ポスターセッション、ISS-P-017, 2019、査読無

③ 中川聖一、山本滉己、山本一公、残差波形の相対位相情報の話者認識の有効性の検討、電子情報通信学会、音声研究会、SP2018-92, 2019、査読無

④ M. Liu, L. Wang, Z. Oo, J. Dang, D. Li, S. Nakagawa”, Replay attacks detection using phase and magnitude features with various frequency resolutions”, Proc. 11th ISCSLP, 2018, DOI:10.1109/ISCSLP.2018.8706628、査読有

⑤ M. Ge, L. Wang, S. Nakagawa, Y. Kawakami, J. Dang, X. Li, D. Chuxing, “Pitch synchronized relative phase with peak error detection for noise robust speaker recognition”, Proc. 11th ISCSLP, 2018, DOI:10.1109/iscslp.2018.8706701、査読有

⑥ D. Li, L. Wang, J. Dang, M. Liu, Z. Oo, S. Nakagawa, H. Guan, X. Li, “Multiple phase information combination for replay attacks detection”, Proc. Interspeech, 2018,

656-659, DOI:10.21437/Interspeech.2018-2001、査読有

⑦ Z. Oo, L. Wang, L. Meng, S. Nakagawa, M. Iwahashi, “Automatic speaker verification for replay attacks using mel-scale phase and magnitude features”, 日本音響学会、春季講演論文集、2-8-4, 2018、査読無

⑧ L. Wang, S. Nakagawa, J. Dang, J. Wei, T. Shen, “Pseudo-pitch-synchronized phase information extraction and its application for robust speaker recognition”, Proc. GCCE, 2017, 253-257, DOI:10.1109/GCCE.2017.8229401、査読有

⑨ Z. Oo, Y. Kawakami, L. Wang, S. Nakagawa, X. Xiao, M. Iwahashi, “DNN-based amplitude and phase feature enhancement for noise robust speaker identification”, Proc. Interspeech, 2016, 2204-2208, DOI:10.21437/Interspeech.2016-717、査読有

〔図書〕（計　1 件）

① 中川聖一（編著)、音声言語処理と自然言語処理（増補)、コロナ社、2018、査読無

６．研究組織

(1)研究分担者

研究分担者氏名：岩橋　政宏

ローマ字氏名：　(IWAHASHI, Masahiro)

所属研究機関名：長岡技術科学大学

部局名：工学研究科

職名：教授

研究者番号（8 桁)：30251854

研究分担者氏名：王　龍標

ローマ字氏名：　(WANG, Longbiao)

所属研究機関名：長岡技術科学大学

部局名：工学研究科

職名：准教授

研究者番号（8 桁)：30510458

「2016 年 11 月 4 日削除」