Vol.2018-MUS-119 No.32
Vol.2018-SLP-122 No.32
2018/6/16

マイク間到来時間差のばらつきを用いたなりすまし音声検出の検討

矢口 凌也[†1]　塩田 さやか[†1]　小野 順貴[†1]　貴家 仁志[†1]

概要：近年，話者照合システムに録音された登録話者の声を再生するなりすまし攻撃が問題視されている．これまでに入力音声が人間による実発話かスピーカーによる再生音声かを判別する手法として，2 チャネルマイク入力信号間の到来時間差を用いて生体検知を行う手法が提案され，高い識別性能が得られることが報告されている．この手法では発声時の音素毎の音源が微小に変化することを利用するため，口とマイク間の距離や背景雑音によっては検出が難しいという問題点があった．そこで本報告では，逆に無発話区間におけるマイク内到来時間差に着目し，スピーカー再生時の特徴を検出することを検討する．人間が発声する場合，到来時間差検出を行うと無発話区間では一定位置からの音源定位が困難となる．一方，スピーカー再生時は元の音声が無発話区間であっても微小な電子音が再生される場合が多く音源が一定に定位すると考えられる．なりすまし音声を用いた検証実験において，提案法はマイクおよびスピーカーの種類が異なる場合にも高い検出精度が得られたことを報告する．

キーワード：話者照合，声の生体検知，なりすまし音声検出，到来時間差

Spoofing Detection Method using Time Difference Of Arrival To Microphones

Ryoya Yaguchi[†1]　Sayaka Shiota[†1]　Nobutaka Ono[†1]　Hitoshi Kiya[†1]

Abstract: Recently, replay attacks against to speaker verification systems are regarded as one of serious spoofing attacks. In order to detect replay attacks, VoiceLive system, which is a voice liveness detection approach, has been reported. This approach captures time-difference-of-arrival (TDoA) changes in a sequence of phoneme sounds to two microphones, and the experimental results showed the high accuracy on the smartphones. However, since this approach focuses on the sensitive TDoA changes, the performance strongly depends on the recording conditions. Considering this point, this paper proposes a novel spoofing detection approach. The proposed method detects TDoA changes during non-voice periods and distinguishes replay attacks from genuine speakers. The case of replay attacks, there may be stationary electric noise from loudspeakers in non-voice periods. Thus it leads to detect uniform TDoA while it is difficult to cause in the case of genuine speakers. In the spoofing detection experiments, the proposed method obtained the high accuracy even though some different types of microphones and loudspeakers are used.

Keywords: Speaker Verification, Voice Liveness Detection, Anti Spoofing, Time Difference Of Arrival

1. はじめに

近年，スマートフォンや，ネットバンキング等の本人認証において生体認証技術の導入が進んでいる．話者照合は声を用いた生体認証技術であり，技術の発展による照合性能の向上および導入コストの低さから普及しつつある．一方で，再生音声や合成音声の品質向上からなりすまし攻撃として用いられることの危険性が指摘されており，なりすまし音声検出は話者照合における重大な課題の一つとなってきている [1]．特に登録話者の声を録音し再生するなりすまし攻撃は，専門的知識を持たない詐称者でも簡単に詐称が行えてしまうため対策が急務と言える．これまでになりすまし音声検出のコンペティションである ASVspoof が 2015 年 [2] および 2017 年 [3] に開催され，音声合成や録音再生によるなりすまし攻撃への対策手法が数多く提案された．また，なりすまし音声検出とは別に，入力された音声が人間による実発話かスピーカーによる再生音声かを判別する枠組みである声の生体検知も提案された [4]．声の生体検知では，人間が発話する際に必然的に用いる特徴を検出することに着目する．これまでに声の生体検知を実現するための手法としてポップノイズ検出法が提案されている．また，声の生体検知の別の実現手法として，2 本のマイクを用いたマイク間到来時間差を検出する手法が提案され

[†1] 現在，首都大学東京 システムデザイン研究科　情報科学域
Presently with Tokyo Metropolitan University, Faculty School of Systems Design, Department of Computer Science

Vol.2018-MUS-119 No.32
Vol.2018-SLP-122 No.32
2018/6/16

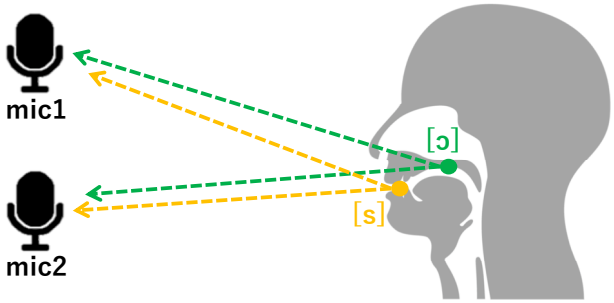

図 1　発音位置の差による到来時間の差異

た [5]．これは図 1 に示すように，人間が発声する際，音素毎に音源位置が若干異なるという現象に着目し，2 チャネルマイク入力信号間の到来時間差により，人間の発音位置が音素毎に異なることを利用して生体検知を行う手法であり，高い識別性能が得られることが報告されている．しかし，この手法で検出される発音位置の違いは極めて微小であるため，口とマイク間の距離や背景雑音によっては検出が難しく，また精度が発話内容に依存しやすいという問題点があった．そこで本報告では，逆に無発話区間に着目した．発話内容に依存しない手法として，スピーカー再生時の特徴を検出することを検討する．到来時間差検出を用いると無発話区間では音が人から発せられないため一定の音源定位が困難である．一方，スピーカーは元の音声が無発話区間であっても微小な電子音が再生される場合が多い．そこで無発話区間における到来時間差の変動に着目したなりすまし検出を行う．検証実験において，提案法はマイクおよびスピーカーの種類が異なる場合にも高い検出精度を得られることを報告する．

2. 関連研究

2.1 ASVspoof2015, 2017 [2, 3]

近年，話者照合において重大な問題となってきているなりすまし攻撃には大きく分けて声真似・音声合成・声質変換・録音再生の 4 種類がある．これらのなりすまし攻撃を検出する方法についても広く議論する必要が出てきたことから，ASVspoof というなりすまし攻撃検出に関するコンペティションが開催された．第一回目の ASVspoof2015 では音声合成，声質変換に焦点を当てていたが，より容易でかつ防ぐのが困難な録音再生について考える必要があるという指摘から ASVspoof2017 という第二回目のコンペティションが開催された．これまでの ASVspoof で発表された手法は主に様々な音響的特徴量を用いた統計モデルによる判別手法だった [6–8]．しかしながら，音響的特徴量を模倣する手法の提案や合成方法が未知の場合に関する脆弱性も問題となっている．

2.2 話者照合のための声の生体検知 [4, 9–11]

ほとんどのなりすまし攻撃に必要となる過程がスピーカーによる音の再生である．そのため入力音声がスピーカー再生なのか実発話なのかを検出できればなりすまし攻撃へのより根本的な解決策になると言える．このスピーカー再生か実発話かを検出する声の生体検知という枠組みが提案された．これを話者照合への前段として使用することで，なりすまし攻撃による話者照合システムの精度低下を防ぐことを目的としている．これまでに声の生体検知を実現する手法として話者の呼気を検出するポップノイズ検出法が提案されており，話者照合と組み合わせることでほぼなりすまし音声を棄却できることを報告している．しかし，呼気を検出する必要があることから収録マイクの性能や話者とマイクとの距離，周囲の環境に大きく依存してしまうといった課題がある．

2.3 到来時間差を用いた音源定位 [5]

別の声の生体検知手法として，マイク間到来時間差を用いた手法が提案されている．これは，スマートフォン等の機器にマイクが複数個搭載されることが増えてきたことに着目し，複数チャネルで音声を収録することが可能であることを利用した手法となっている．人間は発声する際に口内の奥の方や舌先など様々な位置で音を生成している．到来時間差を用いる手法では図 1 に示すように，複数のマイクを用い音素毎に音源定位の位置が細かく変わることを検出している．具体的には，時刻 t における 2 チャネル信号 $\mathrm{mic}_1(t)$，$\mathrm{mic}_2(t)$ において，式 (1) で示される相互相関関数 $\mathrm{CC}(d)$ の最大値を求め，そこから到来時間差を算出している．

$$\mathrm{CC}(d) = \frac{\sum_i[(\mathrm{mic}_1(i) - \overline{\mathrm{mic}_1(i)}) * (\mathrm{mic}_2(i+d) - \overline{\mathrm{mic}_2(i+d)})]}{\sqrt{\sum_i(\mathrm{mic}_1(i) - \overline{\mathrm{mic}_1(i)})^2}\sqrt{\sum_i(\mathrm{mic}_2(i+d) - \overline{\mathrm{mic}_2(i+d)})^2}} \quad (1)$$

$$\Delta t = \arg\max_d \mathrm{CC}(d), \quad (2)$$

ここで，d は発生している遅延点数であり，Δt は到来時間差を示す．しかし，単純に相互相関関数を用いて到来時間差を検出する手法は背景の雑音や残響に強く影響を受け，到来時間差を算出する精度が低下してしまう．そこで，一般化相互相関関数 [12] を用い，相互相関関数の計算時に重み付け関数を使用することで，位相情報のみを用いる手法を組み込み，無相関な雑音等に対する頑健性を向上させている．論文では高いサンプリング周波数を用い，マイクと口の関係が正しくセッティングされている状況で収録された音声に関して非常に高い検出精度を得ることが報告されている．

3. 提案法

3.1 到来時間差検出法の問題点

2.3 節で述べた到来時間差を用いる手法は微細な音源差を検出する必要があるために収録条件が限定されている．

Vol.2018-MUS-119 No.32
Vol.2018-SLP-122 No.32
2018/6/16

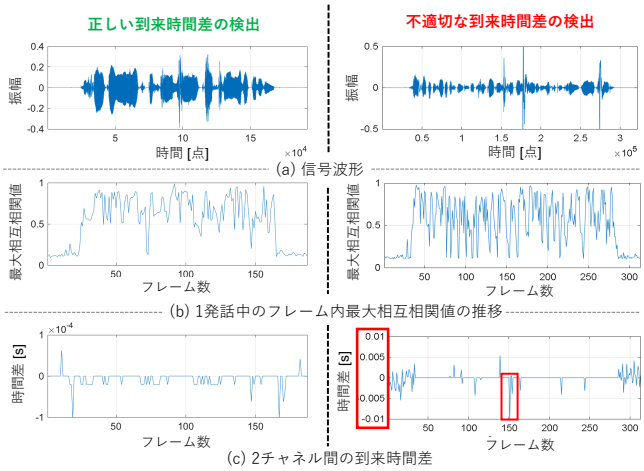

図 2 到来時間差の検出例

適切な環境で収録された音声を用いない場合に正しい挙動を示すことが難しい．図 2 に正しく検出できる例とできない例を示す．図 2 の左側では正しく到来時間差が検出できており，図 2(c) のように発話区間で常に微小な音源位置の変化が確認できている．しかし，図 2 右側の例の赤枠部においては音源位置の変化が大きすぎてしまい定位の範囲が口内にとどまらず正しく検出できていないことがわかる．文献 [5] では更に音素毎の変化を正しく検出する必要があるため収録条件に注意を払わないと実現が難しいことがわかる．

3.2 無発話区間におけるスピーカー特性

人間の発話をマイクで収録する場合の観測モデルを，時間周波数領域で表すと，以下のように書くことができる．

$$M_1(t,\omega) = H_1(\omega)S(t,\omega) + N_1(t,\omega) \quad (3)$$
$$M_2(t,\omega) = H_2(\omega)S(t,\omega) + N_2(t,\omega) \quad (4)$$

ここで，M_1, M_2 がマイク 1, 2 での観測信号，S が音声信号，H_1, H_2 が発話位置からマイク 1, 2 までの伝達特性，N_1, N_2 が背景雑音を表す．無発話区間，すなわち音声信号 $S(t,\omega)$ が 0 の区間では、

$$M_1(t,\omega) = N_1(t,\omega) \quad (5)$$
$$M_2(t,\omega) = N_2(t,\omega) \quad (6)$$

のように背景雑音のみとなり，これを定位した場合，どこに定位されるかは不確定となる．一方、なりすまし攻撃を行うためにマイク p で録音した音声は以下のように表せる．

$$M_p(t,\omega) = H_p(\omega)S(t,\omega) + N_p(t,\omega) \quad (7)$$

この録音音声をスピーカーで再生した場合，観測信号は

$$M_1(t,\omega) = H_1'(\omega)(M_p(t,\omega) + N_s(t,\omega)) + N_1(t,\omega) \quad (8)$$
$$M_2(t,\omega) = H_2'(\omega)(M_p(t,\omega) + N_s(t,\omega)) + N_2(t,\omega) \quad (9)$$

と表せる．$H'1(\omega)$, $H'2(\omega)$ はスピーカーからマイク 1, 2

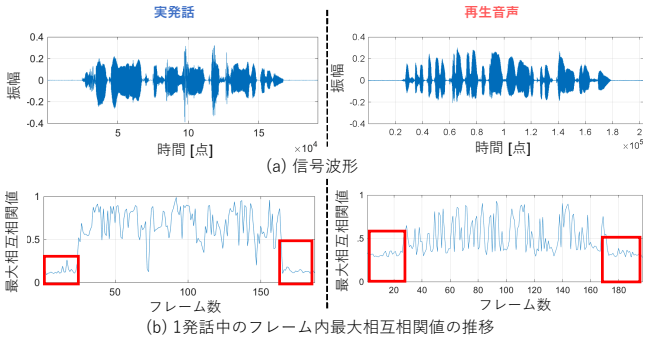

図 3 フレーム間最大相互相関値の推移

までの伝達特性，$N_s(t,\omega)$ は再生系で生じる雑音である．無発話区間，すなわち音声信号 $S(t,\omega)$ が 0 の区間では，

$$M_1(t,\omega) = H_1'(\omega)(N_p(t,\omega) + N_s(t,\omega)) + N_1(t,\omega) \quad (10)$$
$$M_2(t,\omega) = H_2'(\omega)(N_p(t,\omega) + N_s(t,\omega)) + N_2(t,\omega) \quad (11)$$

となる．つまり，$S(t,\omega) = 0$ であっても，録音時に録音された背景雑音および再生時に再生系で発生した雑音はスピーカーから，一定の伝達特性を介してマイクに到来するため，$H_1'(\omega)$, $H_2'(\omega)$ で決まる音源位置に定位されることになる．

図 3 に実発話と再生音声において同一の文章を読み上げた時の信号波形と，各フレームにおけるマイク間の相互相関の最大値の推移を示す．赤枠の無発話区間を見ると，実発話では最大相互相関が小さい値をとるのに対し，再生音声では概ね実発話より大きいことがわかる．これは，上記のように実発話の無発話時には一定の音源がないため，相互相関が全体的に低くなり，その中からピーク値を一点選ぶと，一定な値を取ることが難しくなるためである．一方，再生音声では無発話区間の再生においてもスピーカーから微弱な電子音が発されるため，2 チャネルマイク信号間は実発話の無音区間に比べ高い相関値が得られることがわかる．

3.3 無発話区間における最大相互相関値の平均を用いたなりすまし検出

前節の結果より，入力音声の無発話区間において，2 チャネル信号の最大相互相関の平均はなりすまし音声の方が実発話よりも高いと想定される．そこで，本研究では無発話区間のチャンネル間最大相互相関の平均を算出し，その値を用いたなりすまし検出を行うことを提案する．手順としては以下のように行う．

1. 入力音声の無発話区間のみを抽出．ただし，チャネル間では共通の時間を用いる．
2. 各チャネルの信号をそれぞれフレーム分割し，短時間フーリエ変換を行う．
3. チャネル間の一般化相互相関をフレーム毎に算出．
4. フレーム単位の一般化相互相関の平均値を算出．閾値

表 1　VLD データベースの収録条件

サンプリング周波数	48 kHz
量子化ビット数	24 bit
話者	女性 15 名
マイク	3 種類 × 2 本
スピーカー	BOSE 111AD
収録環境	防音室

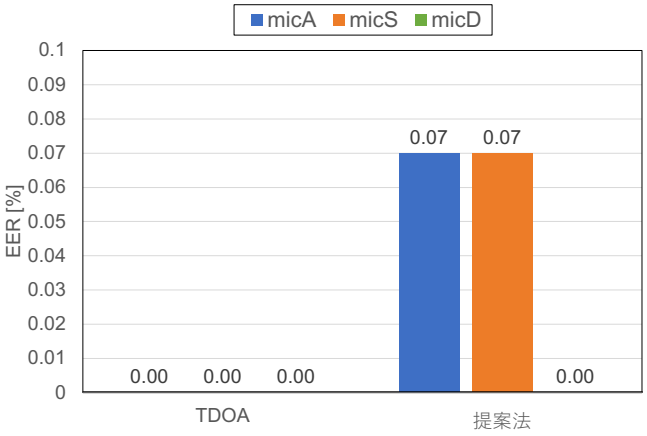

図 4　収録機器別の EER

より高ければ再生音声，低ければ実発話とする．ただし，手順 1 において，発話区間が入ると平均値に大きく影響を与えるため，無発話区間の抽出は厳しく設定する必要がある．

4. 評価実験

提案法の性能評価をするためになりすまし音声の検出実験を行った．本実験では収録機器別および再生機器別の 2 種類の実験を行なった．

4.1 収録機器別実験

4.1.1 実験条件

まず，VLD データベース [4] を用いてスピーカーが一つ，収録マイクが 3 種類ある場合の実験を行った．収録条件を表 1 に示す．用いたマイクは AKG P170（micA），SONY ECM-XYST1M（micS），DPA 4066（micD）の 3 種類でそれぞれ 2 本ずつ使用して到来時間差の検出を行った．文章数は実発話 1500 文，再生音声は各マイク毎に 1500 文である．比較手法は以下の 2 つである．

TDOA

各発話の全フレームを用いて一般化相互相関関数より到来時間差を算出し，それらの最大値および最小値を得る．最大値と最小値の差を到来時間差の変動範囲として定義し，設定した閾値より大きければ実発話として受理し，小さい場合はなりすまし音声として棄却する．なお，文献 [5] の音素情報を用いない簡易版の実験とみなす．

表 2　収録データの詳細

サンプリング周波数	48 kHz
量子化ビット数	16 bit
話者	男性 1 名
マイク	micA
スピーカー	3 種類
収録環境	EV ホール

提案法

3.3 節の手順に従う．なお，無発話区間の抽出については音声波形の振幅が閾値を超える場合を取り除いている．無発話区間をより厳密に取るために閾値は非常に小さい値にしている．

TDOA および提案法どちらにおいても到来時間差検出には，周波数分解能 46.9 Hz，分解窓幅 21 msec，オーバーラップなしという条件を用いた．性能評価に使用する尺度は，実発話誤棄却率と再生音声誤受理率が等しくなる点である等価エラー率（Equal Error Rare; EER）を用いた．ここで，TDOA では到来時間差の変動範囲を，提案法では平均値を受理もしくは棄却するために閾値を変えて EER を算出する．

4.1.2 実験結果

図 4 に収録機器別の声の生体検知実験における EER を示す．まず TDOA についてみると，マイクの種類を問わず誤りがないことがわかる．しかし，実際の到来時間差を見ると本来の目的である音素毎に異なる音源の定位を行なっているとは言えない場合も多い．そのため音素情報も合わせて用いる場合には精度が悪化すると考えられる．次に提案法と比較すると，マイクの種類を問わず高精度で識別できていることがわかる．提案法は手順が簡単でマイクが 2 本あれば実現可能でありながら TDOA にほぼ匹敵する精度を得ることができた．

4.2 再生機器別実験

4.2.1 実験条件

次に収録マイクの種類を 4.1 節で用いた micA に固定し，スピーカー別に各手法の精度を調べる．収録条件を表 2 に示す．スピーカーについては SONY SRS-ZR7（SONY），ELECOM LBT-SPP300（ELECOM），iPhone6s の 3 種類を使用した．SONY はハイレゾ音源対応のスピーカーであり，低音域から高音域まで再現力のある据え置き型スピーカーである．ELECOM はポータブルスピーカー，iPhone6s はスマートフォンであることから音質や音の再現力としてはどちらも SONY より劣る．実験に用いた文章数は実発話 5 文，スピーカー毎に 5 文である．

4.2.2 実験結果

図 5 に再生機器別の生体検知実験の EER を示す．TDOA の識別性能がいずれの再生機器を用いた場合にも悪いこと

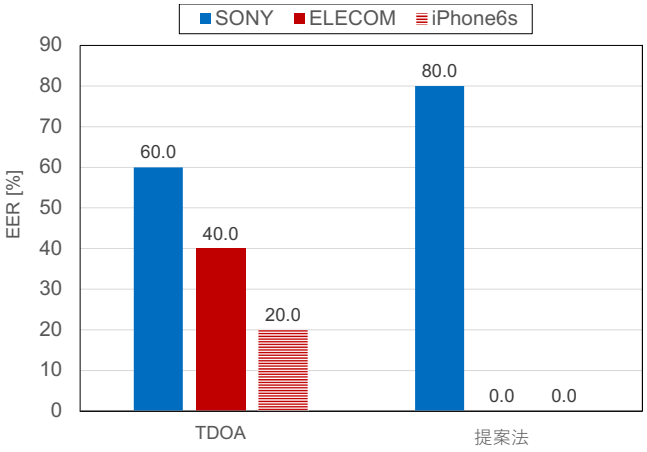

図 5　再生機器別の EER

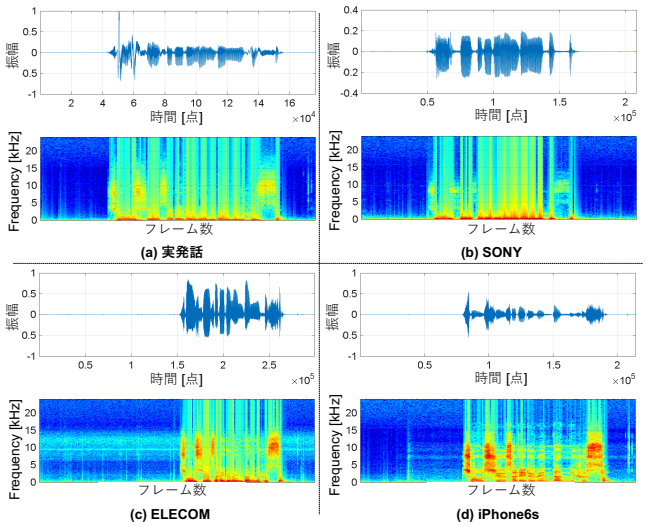

図 6　実発話と各種再生機器における音声波形とスペクトログラム

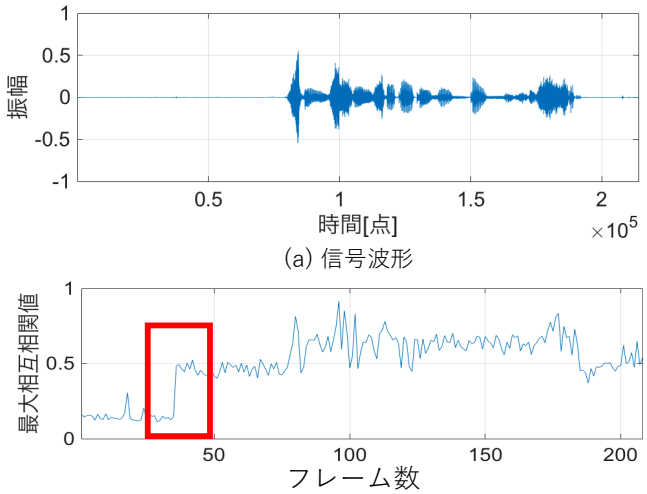

図 7　無発話区間における最大相互相関の変化

がわかる．これは収録環境が前実験とは異なり静かではあるものの背景雑音が存在することが影響していると考えられる．実際に誤受理，誤棄却した文章を確認したところ適切な到来時間差の検出ができていなかったことから，収録環境への依存が大きいことがわかる．一方，提案法では SONY 以外のスピーカーを用いた場合に識別誤りがなく高い精度であることが確認できる．SONY は高音質のスピーカーであり無発話区間ではほぼ電子音を発していないため低い性能となっている．実際になりすまし攻撃を行う場合，詐称者が再生機器を選ぶことができる一方で，収録マイクはシステム構築の際に決めることが可能であると想定できるため，提案法はポータブルスピーカーについては有用であると期待できる．

次に，収録した実発話音声とそれぞれのスピーカーにより再生した音声のスペクトログラムを図 6 に示す．図 6（a）の実発話と図 6（b）の SONY を使用した場合のスペクトログラムを比較すると，ほぼ同様の周波数成分を表現できていることが確認できる．特に無発話区間において最大相互相関値の平均値が実発話とほぼ同等であったため SONY を用いた場合に提案法の識別性能が悪くなったと考えられる．また，図 6（c）の ELECOM と図 6（d）の iPhone6s の場合において，スペクトログラムの 10 kHz 付近の周波数帯に定常的に成分があることが確認できる．この定常的に現れている成分によって提案法では無発話区間においても音源定位が安定し，相互相関の平均が高い値となったと考えられる．本実験では試験的に話者照合システムが収録を開始する前から音声獲得終了時までスピーカーによる再生が常に行われることを想定して実験を行った．しかし，実際には詐称者が再生を開始するタイミングは話者照合の認証開始後であると考えられる．図 6（d）の iPhone6s のみ収録が始まってからスピーカー再生を始めた場合の結果となっている．0.4 秒付近のスペクトログラムに見える若干の変化がスピーカー再生を始めたタイミングになっている．ここでさらに図 7 に iPhone6s の波形と最大相互相関の推移を示す．図 7 を見ると 0.4 秒付近で相互相関の値が最も高くなっていることが確認できる．つまり，なりすまし攻撃の際の再生開始のタイミングが遅いほど無発話区間の最大相互相関の平均は，実発話に近づくという問題があり，今後の課題の一つだと考えられる．

5. おわりに

本稿では，2 チャネル信号間の相互相関関数に着目した声の生体検知法を検討した．提案法では無発話区間の相互相関の平均を見ることでなりすまし攻撃を高い精度で検出できることを報告した．今後の課題として，環境雑音に対する提案法の評価および無発話区間抽出手法の改善，より大規模な評価実験等が挙げられる．

謝辞　本研究の一部は科学研究費若手研究（B）16757733 による．

参考文献

[1] Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi., “Spoofing and Countermeasures for Automatic Speaker Verification,” in Proc. Interspeech, pp.925–929, 2013.

[2] ASVspoof2015, http://www.asvspoof.org/index2015.html

[3] ASVspoof2017, http://www.asvspoof.org

[4] Sayaka Shiota, Fernando Villavicencio, Junichi Yamagishi, Nobutaka Ono, Isao Echizen, Tomoko Matsui., “Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification,” in Proc. Interspeech, pp.239–243, 2015.

[5] Linghan Zhang, Sheng Tan, Jie Yang, Yinging Chen., “VoiceLive:A phoneme localization based liveness detection for voice authentication on smartphones,” Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp.1080–1091, 2016.

[6] Sarfaraz Jelil, Rohan Kumar Das, S.R.M.Prasanna, Rohit Sinha., “Spoof Detection Using Source, Instantaneous Frequency and Cepstral Features,” in Proc. Interspeech, pp.22–26, 2017.

[7] Roberto Font, Juan M.Espin, Maria Jose Cano., “Experimental analysis of features for replay attack detection-Results on the ASVspoof 2017 Challenge,” in Proc. Interspeech, pp.27–31, 2017.

[8] Xianliang Wang, Yanhong Xiao, Xuan Zhu., “Feature selection based on CQCCs for automatic speaker verification spoofing,” in Proc. Interspeech, pp.32–36, 2017.

[9] Sayaka Shiota, Fernand Villavicencio, Junichi Yamagishi, Nobutaka Ono, Isao Echizen, and Tomoko Matsui., “Voice liveness detection for speaker verification based on a tandem single/double-channel pop noise detector,” in Proc. The Speaker and Language Recognition Workshop Odyssey, pp.259–263, 2016.

[10] 矢口凌也, 塩田さやか, 貴家仁志., “話者照合のための低周波および高周波成分情報を用いた声の生体検知の提案,” 日本音響学会春季大会, no.2-Q-3, pp.137–140, 2018.

[11] 望月紫穂野, 塩田さやか, 貴家仁志., “話者照合のための話者性を考慮した音素情報に基づくポップノイズ検出法を用いたテキスト依存型声の生体検知,” 電子情報通信学会音声研究会, vol.117, no.517, (no.SP2017-94) pp.57–62, 2018.

[12] Jian Liu, Gorkem Kar, Yingying Chen, Jie Yang, Marco Gruteser., “Snooping keystrokes with mm-level audio ranging on a single phone,” in Proc. ACM MobiCom, pp.142–154, 2015.