複数チャネル間の相互相関関数を用いたなりすまし検出法の雑音環境下における評価*

☆矢口凌也, 塩田さやか, 小野順貴, 貴家仁志 (首都大学東京)

1 はじめに

近年，生体認証技術の精度向上に伴い，スマートフォンやネットバンキング，入国管理など様々な場所で生体認証システムの導入が進んできている．生体認証技術の一つである話者照合は声を生体情報として用いている．話者照合はスマートスピーカーや音声対話システムとの親和性も高く，導入しやすいという利点がある．しかし，他の生体認証技術と同様に，なりすまし攻撃を考慮する必要がある．話者照合のシステムへのなりすまし攻撃としては，音声合成や録音音声をスピーカー再生することが挙げられている．このなりすまし攻撃を検出することが重大な課題として着目され，近年特に活発に研究が行われている [1,2].

これまでに，話者照合のためのなりすまし検出の枠組みの一つとして，声の生体検知という入力された音声が人間による発声かスピーカー再生かを判別する枠組みが提案された [3]．声の生体検知では，人間が発話する際に必然的に表れる特徴を検出することに着目しており，その 1 つとして人間の呼気を検出するポップノイズ検出法が提案されている．また，別の実現手法として，口内の音源定位位置の変動を検出する手法も提案された [4]．しかし，これらの手法では発話内容や収録環境に認証精度が依存しやすいという問題があった．

筆者らはこれまでに，発話内容に依存しない，複数マイクを用いたなりすまし検出法を提案してきた [5](図 1)．提案した手法は，実発話固有の現象を検出するのではなくスピーカー再生特有の現象を検出することに着目している．この手法では特に，実発話では発話の前後やショートポーズなどの無音部で話者からは音が発せられていないが，なりすまし攻撃の場合にはなりすましに用いる音声の収録時の背景雑音やスピーカー再生時に含まれる電磁ノイズなどを発している点に着目している．これまでは，この手法の有効性を初期実験により評価したもののみを報告してきた．そこで，本報告では更に提案法がどのような環境下においてどの程度頑健性を持つのかについて実験を行い，提案法の有効性と特徴を調査した．

2 複数チャネル間の相互相関関数を用いたなりすまし検出法

2.1 到来時間差を用いたなりすまし検出 [4]

これまでに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．具体的に

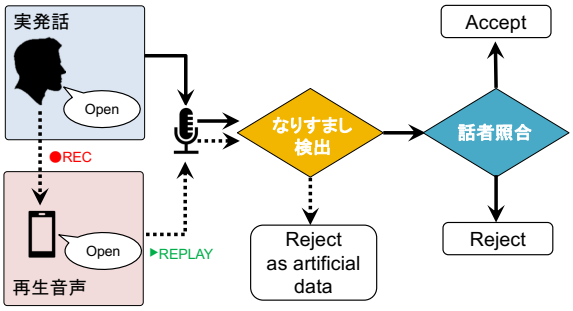

図 1 なりすまし検出のシステムフロー

は，2 本のマイクがある場合に，2 チャネル信号 $r_1(t)$, $r_2(t)$ (t はフレームの時刻 $t = [1, ..., T]$) において，式 (1) で示される相互相関関数 CC(d) [4,6] の最大値 (最大相互相関値) を求め，そこから音声のマイク間到来時間差 Δt を算出することで音源位置の推定を行っている．

$$\mathrm{CC}(d) = \frac{\sum_t [(r_1(t) - \mu_{r_1}) * (r_2(t+d) - \mu_{r_2})]}{\sqrt{\sum_t (r_1(t) - \mu_{r_1})^2} \sqrt{\sum_t (r_2(t+d) - \mu_{r_2})^2}}, \quad (1)$$

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

ここで，d は遅延点数，μ_{r_1}，μ_{r_2} はそれぞれ対応する信号の平均を表している．人間は発声時に口内の様々な位置から音を発するため，人間の場合にはこの到来時間差が微細に変動するがスピーカーの場合には変動しないことを利用した検出法となっている．文献 [4] 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固定されている状況で収録された音声に関して高い検出精度を得られることが報告されている．

2.2 無発話区間におけるスピーカー特性

2.1 節で述べた到来時間差の変動を用いたなりすまし検出は収録条件や発話内容に制限があった．そこで，それらの条件に依存しにくい手法として，複数チャネル間の最大相互相関値を用いたなりすまし検出法を提案した [5]．この手法は，人間が発声する際，発話の前後やショートポーズなどの無音部では音が発されないため最大相互相関値が安定しないが，スピーカー再生ではなりすましに用いる音声の収録時の背景雑音や再生系の電磁ノイズを発するため無発話区間においても最大相互相関値が高くなることに着目している．この特徴を観測モデルを用いて説明する．まず，人間の発話を 2 本のマイク a,b で収録する場合の時間周波数領域における観測モデルを以下

*Evaluation of spoofing countermeasure using generalized cross-correlation between multiple channels under noise environment. by YAGUCHI Ryoya, SHIOTA Sayaka, ONO Nobutaka and KIYA Hitoshi (Tokyo Metropolitan University)

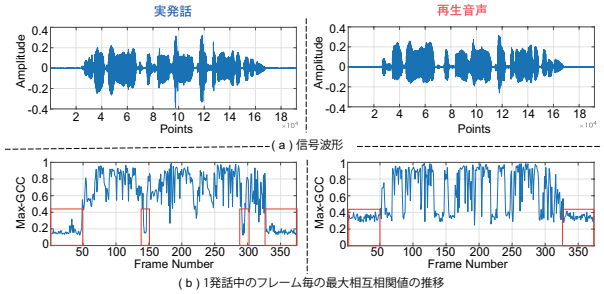

図 2　フレーム毎の最大相互相関値の推移
(発話内容:自称女優の美しいイボンヌに恋する物語だ)

の式で表す.

$$M_a(t,f) = H_a(f)S(t,f) + N_a(t,f), \quad (3)$$
$$M_b(t,f) = H_b(f)S(t,f) + N_b(t,f), \quad (4)$$

ここで，M_a, M_b がそれぞれマイク a, b での観測信号，S が音声信号，H_a, H_b が発話位置からマイク a, b までの伝達特性，N_a, N_b が背景雑音を表す．また t，f は時間，周波数のインデックスを表す．無発話区間，すなわち音声信号 $S(t,f) = 0$ の区間での観測モデルは，

$$M_a(t,f) = N_a(t,f), \quad (5)$$
$$M_b(t,f) = N_b(t,f), \quad (6)$$

のようにそれぞれの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．

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

この録音音声をスピーカーで再生した場合，マイク a,b における観測信号はそれぞれ，

$$M_a(t,f) = H'_a(f)(M_p(t,f) + N_s(t,f)) + N_a(t,f), \quad (8)$$
$$M_b(t,f) = H'_b(f)(M_p(t,f) + N_s(t,f)) + N_b(t,f), \quad (9)$$

と表せる．$H'_a(f)$, $H'_b(f)$ はスピーカーからマイク a, b までの伝達特性，$N_s(t,f)$ は再生系で生じる雑音である．また，無発話区間の観測モデルは，

$$M_a(t,f) = H'_a(f)(N_p(t,f) + N_s(t,f)) + N_a(t,f), \quad (10)$$
$$M_b(t,f) = H'_b(f)(N_p(t,f) + N_s(t,f)) + N_b(t,f), \quad (11)$$

となる．つまり，$S(t,f) = 0$ 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，$H'_a(f)$, $H'_b(f)$ で決まる音源位置に定位されることになる．

図 2 に実発話と収録した実発話を再生した音声の波形と，各フレームにおけるマイク間の最大相互相関値の推移を示す．赤枠の無発話区間を見ると，実発話では最大相互相関値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値をとることがわかる．これは，無発話区間においてもスピーカーから微弱な電子音や収録された背景雑音が再生されるため，2 チャンネル間信号は実発話の無音区間に比べ最大相互相関値が高くなるためである．

2.3　無発話区間における最大一般化相互相関値を用いたなりすまし検出

前節の結果より，無発話区間において，マイク間信号の最大相互相関値はなりすまし音声の場合に安定して高く，実発話は定位する音がない無発話時に小さくなると想定される．そこで，無発話区間の最大相互相関値を検出し，その値によるなりすまし検出を行う．また，本研究では一般化相互相関関数を用いて最大相互相関値を求めた．一般化相互相関関数は，振幅を白色化して位相情報のみで相関をもとめる GCC-PHAT と呼ばれるものであり，τ を時間差，t を時間フレーム ($t = [1, ..., T]$)，L をフレーム長として以下のように表せる．

$$\phi_g(\tau;t) = \frac{1}{L}\sum_f \frac{M_1^*(t,f)M_2(t,f)}{|M_1^*(t,f)M_2(t,f)|} e^{j2\pi f\tau/L}. \quad (12)$$

また，各フレームの一般化相互相関関数の最大値 (最大一般化相互相関値) は以下のように表せる．

$$\phi_{max}(t) = \max_\tau \phi_g(\tau;t). \quad (13)$$

なりすまし検出の手法のために 2 種類の無発話区間について考える．1 つは発話中に現れるショートポーズなどの無発話区間，もう一つは発話の前後の無発話区間である．前者の場合，発話区間の始端から終端までの間の最大一般化相互相関値の最小値に着目して判定を行う．後者の場合，発話の始端より前および終端より後の無発話区間の全フレームの最大一般化相互相関値の平均を用いて判定を行うことを考える．t_s を発話開始フレーム，t_e を発話終了フレームとすると，それぞれ以下のように表せる．

$$\Phi_{min} = \arg\min_{t_s \le t \le t_e} \phi_{max}(t), \quad (14)$$

$$\Phi_{ave} = \frac{1}{K} \sum_{1 \le t < t_s,\ t_e < t \le T} \phi_{max}(t). \quad (15)$$

ただし，K は発話区間前後のフレーム総数を表す．

3　評価実験

スマートスピーカーや音声対話システムの使用環境としては，静かな室内のような雑音の少ない環境や，テレビや空調の動作音といった生活騒音が存在する環境が想定される．また，なりすましに用いる音声を収録する際にも，同様の環境が考えられる．さらに，なりすまし収録に用いられるマイクや再生に使用されるスピーカーには様々な組み合わせが考えられる．そこで本実験では，なりすまし音声の収録環境と，なりすまし検出時の環境を様々な組み合わせで行うことで，なりすまし検出法の頑健性調査を行った．

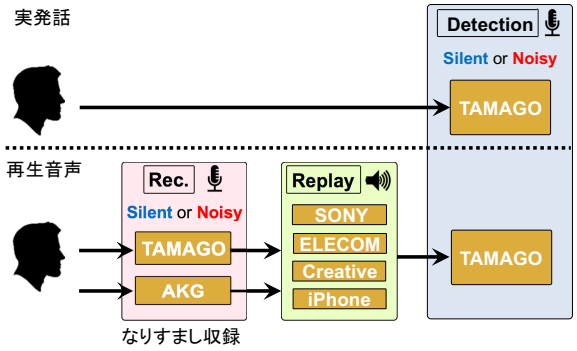

図 3　テストデータの収録フロー
(TAMAGO:8ch マイクロフォンアレイ,
AKG:単一指向性コンデンサーマイク)

3.1　収録環境

本実験のテストデータの収録フローを図 3 に示す．なりすまし収録の環境となりすまし検出のテスト時の環境にそれぞれ静かな環境 (silent)，背景雑音のある環境 (noisy) を想定した．silent では空調などの雑音がない状況，noisy ではテレビや空調など非定常音と定常音どちらも含むような状況を用意した．したがって，なりすまし収録の環境 (2 通り) となりすまし検出のテスト時の環境 (2 通り) の組み合わせの 4 通りが存在する．なりすまし収録時に使用するマイクは AKG (P170) と TAMAGO (TAMAGO-03) の 2 種とした．AKG は指向性があり，TAMAGO は音声入力時の発話者位置に自由度を持たせるため指向性の弱い MEMS マイクを複数搭載している．本実験では AKG の場合は AKG を 2 本同じ向きに平行に設置した．TAMAGO の場合には 16 チャンネルのマイクのうち 2 つを用いて収録を行った．

なりすまし音声を再生するスピーカーには SONY (SRS-ZR7)，ELECOM (LBT-SPP300)，Creative (INSPiRE2.0 1300)，iPhone (Apple iPhone6s) の 4 種類を用いている．SONY および Creative は持ち運びには向かないものの低音域から高音域までの再現力のある据え置き型スピーカーである．ELECOM は持ち運び用であり，通電すると電磁ノイズを発生する傾向があり，iPhone は目立つ電磁ノイズは発生しないが元の音声よりもこもった音を再生する特徴がある．マイクと口の距離，マイクと再生機器の距離は約 15cm で統一して収録した．

話者は男性 2 名，女性 2 名の計 4 名で，文章数は各話者につき 5 文章の計 20 文を実発話とし，それらを 2 種のマイクでそれぞれ収録し，4 種のスピーカーで再生し再収録した計 160 文 (20 × 2 × 4) をなりすまし音声とした．実発話はなりすまし検出のテスト時、なりすまし音声はなりすまし収録とテスト時の収録環境計 4 種類のそれぞれにおいて収録した．発話内容は日常会話で，発話長は約 3 秒である．なお，収録条件としてはサンプリング周波数 16kHz，量子化 bit 数 16bit を用いた．

表 1　各条件下における EER

		なりすまし収録マイク	
収録-検出環境	手法	TAMAGO	AKG
noisy-silent	CQCC	37.97%	23.41%
	GCC(min)	3.89%	5.91%
	GCC(avg)	**0.00%**	2.33%
noisy-noisy	CQCC	42.86%	25.49%
	GCC(min)	2.50%	4.29%
	GCC(avg)	**0.00%**	2.00%
silent-silent	CQCC	37.30%	23.01%
	GCC(min)	4.72%	12.00%
	GCC(avg)	**0.00%**	7.00%
silent-noisy	CQCC	42.51%	25.12%
	GCC(min)	4.00%	4.21%
	GCC(avg)	2.44%	**1.15%**

3.2　実験条件

本実験の比較手法として ASVspoof2017 のベースラインになった Constant Q ケプストラル係数 (CQCC) [2] を特徴量として用いた GMM に基づく手法を用いた [7]．GMM の学習には VLD データベース [3] から実発話，再生音声それぞれ 900 文を用いた．その他の CQCC の抽出条件は ASVspoof2017 のベースラインに準じている．比較手法は以下の 3 つである．

CQCC
テストデータの CQCC を抽出し，学習した GMM から対数尤度スコアを計算し判定する．

GCC(min)
2.3 節の手順に従い，テストデータから式 (14) の照合スコアを算出し判定する．この際発話前後の無発話区間は用いず，発話区間中のショートポーズの検出に着目した手法となっている．

GCC(avg)
2.3 節の手順に従い，テストデータから式 (15) の照合スコアを算出し判定する．特に発話区間前後の無音区間に着目した手法となっている．

本実験では無発話区間の検出にエネルギーに基づく発話区間検出を利用している．また，発話区間検出において始端・終端が間違っている場合にのみ手修正を行った．提案法の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．

3.3　実験結果

表 1 に環境毎，なりすまし収録のマイク毎のなりすまし検出の EER を示す．結果より，いずれの収録-検出環境，なりすまし収録に TAMAGO，AKG どちらのマイクを使った場合においても CQCC の精度は低いことがわかる．CQCC は ASVspoof2017 の結果からも録音再生攻撃の検出に対しては精度が高く出な

表 2　noisy-noisy 環境下で口-マイク間を 15cm・1m とし音源-マイク間距離を 15cm とした時の EER

	なりすまし収録マイク	
手法	TAMAGO	AKG
CQCC	22.33%	22.33%
GCC(min)	1.67%	2.50%
GCC(avg)	0.00%	1.67%

いことが報告されており妥当な結果と言える．一方，GCC(min) と GCC(avg) はどの環境においても比較的高い精度で検出しており，さらに GCC(min) よりも GCC(avg) の精度が高いことがわかる．この結果より，発話間のショートポーズにおける検出は精度が GCC(avg) よりも安定しないことがわかる．

次に，なりすまし収録のマイクに TAMAGO を用いた場合に着目し，4 種の収録-再生環境について比較する．なりすまし収録時に雑音が入り込んでいる noisy-silent と noisy-noisy の時に特に GCC(avg) の性能が高い．これはなりすまし収録時に，目的音声とともにテレビや空調による雑音が入り込むことにより，再生時に雑音部を再生することで無発話区間においても全体的に最大一般化相互相関値が高く現れるためである．照合時に雑音が存在する noisy-noisy の時，テスト時の背景雑音によって再生された雑音の最大一般化相互相関値が影響を受けることも考えられたが，なりすまし収録時に再生された雑音が最大一般化相互相関値の算出において影響力を持っており高い精度を保てることがわかった．また，silent-silent について見ても GCC(avg) では完全に識別している．これは，なりすまし収録時にテレビや空調による雑音が存在しないが，なりすまし再生時においても雑音が存在しないため，スピーカーから発される微弱な電磁ノイズなどを定位しているためであると考えられる．また，本実験で用いたスピーカーのうち，ELECOM 以外のスピーカーでは電磁ノイズを人間の耳では認識できない程度しか発生していなかったにも関わらず正しく検出できていたのは，非可聴音にもスピーカー再生の特徴があったと考えられる．silent-noisy が silent-silent に比べ精度が多少低下したのは，silent-noisy ではテスト時にテレビや空調による雑音が存在するため，実発話であっても最大一般化相互相関値が高く出てしまったためであると考えられるが，大幅な精度低下ではなくなりすまし検出が頑健に行われたと言える．

次になりすまし収録のマイクに AKG を用いた場合について考察すると，各環境に対し TAMAGO と同様の傾向が伺えるが，TAMAGO に比べ全体的に精度が低い．特に silent-silent では精度の差が顕著に現れている．これは TAMAGO に搭載されているマイクは小型の MEMS マイクであり，利用シーンとして音声入力時の場所に自由度を持たせるように設計されているため，マイク指向性が低く目的音声以外の雑音が入りやすい．一方で，AKG マイクは指向性の高いコンデンサーマイクであり雑音が入りづらいため，スピーカーで再生される雑音が少なく，最大一般化相互相関値が高くならなかったためであると考えられる．

本研究で想定されている環境は，スマートスピーカーや音声対話システムとの対話におけるなりすまし検出であるため，人間とマイク間距離が照合時およびなりすまし過程での収録時においては長いことが考えられる．そこで追加実験として，noisy-noisy の環境下においてマイクからの距離を 1m として実発話のテスト収録を行った．マイクからの距離を 1m としたテスト音声を noisy-noisy のデータに追加し，EER を算出したものが表 2 である．全体的に表 1 の結果と同様の結果になることが確認でき，実環境下においても最大一般化相互相関値を用いたなりすまし検出の頑健性が示された．

4　おわりに

本稿では，マイク間の相互相関に着目したなりすまし検出法を提案し，様々な使用環境を想定し性能を分析した．提案法では無発話区間の最大一般化相互相関値の平均を見ることで実環境においても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，より大規模なデータによる評価，統計的な枠組みの導入などが挙げられる．

謝辞　本研究の一部は科学研究費若手研究（B）16757733，科研費基盤（A）16H01735 による．

参考文献

[1] N. Evans *et. al.*, “Spoofing and countermeasures for automatic speaker verification,” in Proc. Interspeech, pp.925–929, 2013.

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

[3] S. Shiota *et. al.*, “Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification,” in Proc. Interspeech, pp.239–243, 2015.

[4] L. Zhang *et. al.*, “VoiceLive:A phoneme localization based liveness detection for voice authentication on smartphones,” in Proc. the 2016 ACM SIGSAC Conference on Computer and Communications SecurityACM MobiCom, pp.1080–1091, 2016.

[5] 矢口ら, “複数チャネル間の相互相関関数を用いた話者照合のためのなりすまし検出,” 日本音響学会秋季大会, no.3-2-4, pp.1335-1338, 2018.

[6] J. Liu *et. al.*, “Snooping keystrokes with mm-level audio ranging on a single phone,” in Proc. ACM MobiCom, pp.142–154, 2015.

[7] X. Wang *et. al.*, “Feature selection based on CQCCs for automatic speaker verification spoofing,” in Proc. Interspeech, pp.32–36, 2017.