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*

☆矢口凌也, 塩田さやか, 小野順貴, 貴家仁志 (首都大)

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声を用いた生体認証技術である話者照合が普及しつつある．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]．これまでになりすまし検出のコンペティション（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．それらの多くは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 [4,5]．しかし，音響的特徴量を模倣した合成手法や未知の合成手法を用いられた場合の脆弱性も問題となっている．

また，なりすまし検出の一つとして，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 [6,7]．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，その実現手法としてポップノイズ検出法を提案している．また，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も提案された [8]．この手法では人間が発声する際，音素毎に口内の音源位置が若干異なるという現象に着目し，マイク間の入力信号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る．しかし，この手法で検出される発音位置の変動幅は極めて微小であり，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．本報告ではこのマイク間到来時間差に着目しつつ，さらに発話内容にも依存しない手法として，実発話を検出するのではなくスピーカー再生を検出する手法を考える．実発話の場合，発話のない区間は音を発していないためマイク間到来時間差を用いて音源定位を行うと音源位置が不安定になるがスピーカー再生の場合，無発話区間であっても収録時の背景雑音やスピーカーの電磁ノイズが発生するため音源定位されやすい傾向にある．そこで提案法では無発話区間におけるマイク間の相互相関値を用いたなりすまし検出を行う．評価実験において，従来法と比較し提案法は大幅な検出精度の改善が得られたことを報告する．

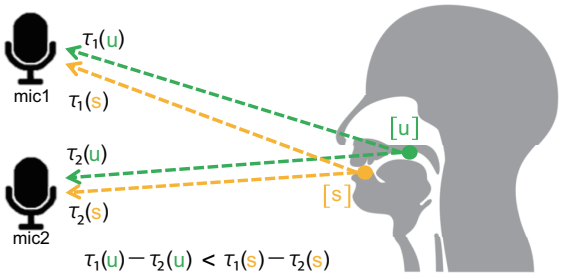

図 1　音素 u および音素 s が mic1,2 に到来するまでの時間 τ_1, τ_2 の差 $(\tau_1 - \tau_2)$ が異なる例

2　関連研究

2.1　到来時間差を用いた音源定位 [8]

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．これは，普及してきているスマートフォンやタブレット，スマートスピーカー等の機器にマイクが複数個搭載されることが多く，複数マイクによる音声収録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や舌先，歯など様々な位置で音を生成している．マイク間の到来時間差を用いる手法では図 1 に示すように，音素毎に音源定位の位置が細かく変動することを複数マイクの到来時間差の変動から検出している．具体的には，各フレームの時刻 $t = [1, ..., T]$ 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) で示される一般化相互相関関数 (Generalized Cross-Correlation function) $\mathrm{GCC}(d)$ [9] の最大値 (最大相互相関値) を求め，そこから到来時間差を算出している．

$$\mathrm{GCC}(d) = \frac{\sum_t [(\mathrm{mic}_1(t) - \overline{\mathrm{mic}_1(t)}) * (\mathrm{mic}_2(t+d) - \overline{\mathrm{mic}_2(t+d)})]}{\sqrt{\sum_t (\mathrm{mic}_1(t) - \overline{\mathrm{mic}_1(t)})^2} \sqrt{\sum_t (\mathrm{mic}_2(t+d) - \overline{\mathrm{mic}_2(t+d)})^2}}, \quad (1)$$

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

ここで，d は遅延点数，記号の上のバーは平均を表している．文献 [8] 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．

*Spoofing detection method using generalized cross-correlation between multiple channels for speaker verification. by YAGUCHI Ryoya, SHIOTA Sayaka, ONO Nobutaka and KIYA Hitoshi (Tokyo Metropolitan University)

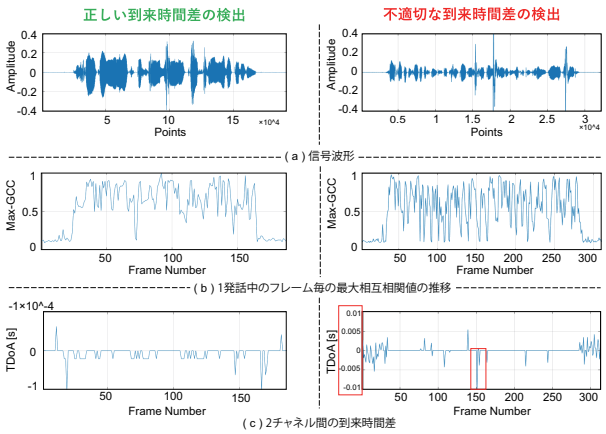

図 2　到来時間差の検出例

2.2　到来時間差検出法の問題点

前節で述べたマイク間の到来時間差を用いる手法では実発話における音素毎の微細な音源差を検出する必要があるため収録条件が限定されている．追実験の結果から正しく検出できる例とできない例を図 2 に示す．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非常に大きくなっており，音速から算出される音源の変動幅は約 3.4 m となっている．つまり音源の変動範囲が口内にとどまらず本来の目的である微細な変動を正しく検出できていないことがわかる．

3　提案法

3.1　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，以下のように表すことができる．

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

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．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、

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

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．

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

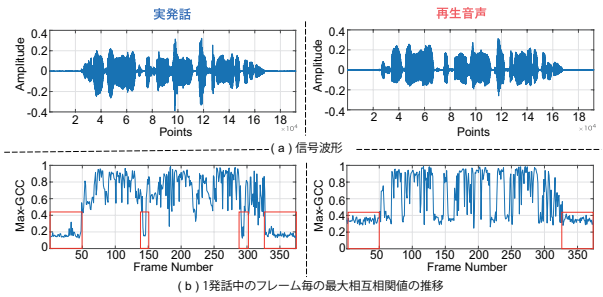

図 3　フレーム毎の最大相互相関値の推移
発話内容：「自称女優の美しいイボンヌに恋する物語だ」

この録音音声をスピーカーで再生した場合，観測信号は

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

と表せる．$H_1'(f)$, $H_2'(f)$ はスピーカーからマイク 1, 2 までの伝達特性，$N_s(t,f)$ は再生系で生じる雑音である．無発話区間の観測モデルは，

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

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，$H_1'(f)$, $H_2'(f)$ 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ことがわかる．これは，実発話の場合，無発話区間には相関のある信号がないため 1 フレーム中の相互相関の値が全体的に低くなる．全体的に低い相関値からピークの値を選択しても，無発話区間の最大相互相関値は全体的に低い値となるためである．一方，再生音声の場合全体的に高い値をとるのは，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，マイク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．

3.2　無発話区間における最大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，本研究では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

表 1　テストデータの収録条件

サンプリング周波数	48 kHz
量子化ビット数	16 bit
マイク	AKG P170 × 2 本
口-マイク距離	15 cm
スピーカー	SRS-ZR7(SONY) LBT-SPP300(ELECOM) INSPiRE2.0 1300(Creative) iPhone6s

表 2　CQCC-GMM の学習条件

データベース	VLD データベース [6]
学習データ	実発話, 再生音声各 900 文
GMM 混合数	512
周波数帯域	20Hz-24kHz
特徴量	CQCC 19 次+Δ+ΔΔ (0 次除く)
マイク	AKG P170 × 2 本
スピーカー	BOSE 111AD

し検出を行うことを提案する．手順は以下の通りである．

1. 入力音声から発話区間以外を抽出．
2. 両マイクの信号をそれぞれフレーム分割し，チャネル間の一般化相互相関をフレーム毎に算出．

ここで用いる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，L をフレーム長として

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

と表せる．提案法で扱うフレーム毎の一般化相互相関関数の時間差 τ についての最大値は以下のように表せる．

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

提案法では 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，特に発話区間内の最大相互相関値の最小値に着目して判定を行う．後者の場合，無発話区間として選択された全フレームの最大相互相関値の平均を用いて判定を行うことを考える．ここで，最大相互相関値の最小値や平均という意味は，発話全体から抽出している特徴であり，フレーム数に対する最大一般化相互相関関数の最小値もしくは平均であるのでそれぞれ

$$\Phi_{min} = \arg\min_t \phi_{max}(t), \quad (14)$$

$$\Phi_{ave} = \frac{1}{K}\sum_{t\in T_{NS}} \phi_{max}(t), \quad (15)$$

と表せる．ただし，T_{NS} は無発話区間のフレーム番号の集合，K は T_{NS} の要素数である．

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．

4.1　実験条件

提案法は学習を必要としないため，まずテストデータの収録条件について述べる．収録を行った場所は静かなエレベーターホールで背景雑音としてエアコンや人の声が同時に収録されている．収録条件は表 1 に示す通りである．テストデータには男性 2 名の実発話 22 文と収録した実発話を 4 種類のスピーカーで再生し，再収録した再生音声 88 文の合計 110 文を用いた．発話内容は日常会話で，発話長は約 3 秒である．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 はモバイル用であり，通電すると電磁ノイズを発生する傾向があり，iPhone6s は目立つ電磁ノイズは発生しないが篭った音を再生する特徴がある．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) を特徴量として用いた GMM ベースの手法を用いた [3]．GMM の学習条件を表 2 に示す．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 つとした．

CQCC
: 実発話・再生音声それぞれから CQCC を抽出し GMM を学習する．照合時はテストデータの 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)
: 3.2 節の手順に従い，各文章において式 (15) を照合スコアとして判定する．この際発話前後の無発話区間はほぼ用いず，発話区間中のショートポーズでの検出に着目した手法である．

GCC(ave)
: 3.2 節の手順に従い，各文章において式 (16) を照合スコアとして判定する．特に発話区間前後の無音区間に着目した手法である．

本実験では GCC(min) の場合に発話中のショートポーズに着目している．GCC(ave) のように発話前後の無発話区間を用いる場合，再生音声の再生開始のタイミングが遅いと無発話区間においても再生音が無く

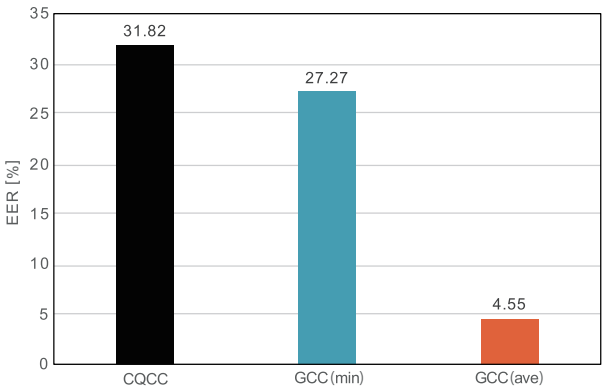

図 4　生体検知実験結果

なり，最大相互相関値の最小値も実発話と同様の値を取ることが考えられるため実発話中の無発話区間であるショートポーズを用いることとしている．

各手法の性能評価に使用する尺度は，以下に示す実発話誤棄却率 (False Rejection Rate; FRR) と再生音声誤受理率 (False Acceptance Rate; FAR) が等しくなる点である等価エラー率（Equal Error Rare; EER）を用いた．

$$\mathrm{FRR} = \frac{\text{誤棄却された実発話サンプル数}}{\text{全実発話サンプル数}}, \quad (16)$$

$$\mathrm{FAR} = \frac{\text{誤受理された再生音声サンプル数}}{\text{全再生音声サンプル数}}. \quad (17)$$

4.2　実験結果

図 4 になりすまし検出実験の各手法ごとの EER を示す．まず CQCC の結果を見ると性能があまり高くないことがわかる．学習に用いたものと同じデータベースの音声を用いて評価した際は EER が 0.23%と高い性能であったが，本実験のテストデータは学習に用いたデータベースと異なるため，収録に用いたマイクは同じであるが再生に用いたスピーカーが異なることが原因の一つであると考えられる．つまり CQCC は再生攻撃が未知のスピーカーである場合に脆弱であることがわかる．次に，CQCC と GCC(min) を比較すると，最小値を用いる手法の方が EER が低いが，十分な改善とは言えない．実発話の場合，背景雑音を定位して無発話区間でも最大相互相関値が高くなることがある．また，特にショートポーズなどの短い無発話区間では前後の発話の影響で音源定位位置があまり変動しない場合もあることが原因の一つであると考えられる．また，この精度を上げるためには発話区間でショートポーズが必ず入るよう発話内容に注意する必要があるため発話内容に依存するという問題もある．次に GCC(ave) についてみると，CQCC，GCC(min) いずれの手法よりも EER が大幅に改善していることがわかる．GCC(ave) は無発話区間の最大相互相関値のフレーム平均を用いるため，GCC(min) よりも値が安定しやすく，また背景ノイズを含む再生音声では特に発話開始前や終了後の短い時間であっても音源定位が安定して行われているため，再生機器に依存しにくい検出手法となっていることがわかる．実際に話者照合を用いる状況では学習データをあまり入手できないことも想定されるが，提案法は実際に用いる際にも頑健に動くと期待できる．

5　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た．提案法では無発話区間の最大相互相関値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，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] ASVspoof2015, http://www.asvspoof.org/index2015.html
[3] ASVspoof2017, http://www.asvspoof.org
[4] R. Font *et. al.*, “Experimental analysis of features for replay attack detection-Results on the ASVspoof 2017 Challenge,” in Proc. Interspeech, pp.27–31, 2017.
[5] X. Wang *et. al.*, “Feature selection based on CQCCs for automatic speaker verification spoofing,” in Proc. Interspeech, pp.32–36, 2017.
[6] 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.
[7] 望月ら, “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.
[8] 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 Security-ACM MobiCom, pp.1080–1091, 2016.
[9] J. Liu *et. al.*, “Snooping keystrokes with mm-level audio ranging on a single phone,” in Proc. ACM MobiCom, pp.142–154, 2015.