ステレオ/モノラルポップノイズ検出法の縦列接続による話者照合のための声の生体検知*

◎塩田さやか（首都大）, Villavicencio Fernando, 山岸順一,
小野順貴, 越前功（NII）, 松井知子（統数研）

1　はじめに

パスワードや暗証番号を用いた認証にはユーザ自身が忘れてしまうことや，他人に盗まれてしまうという問題があった．そこでこれらの問題が起こりにくいという理由から生体認証技術が様々なデバイスのセキュリティとして用いられるようになってきた [1, 2]．近年，生体情報として指紋や顔画像，光彩などを用いた生体認証が普及しつつある．また，声を用いた生体認証技術である話者照合も，導入が簡単であるという理由から注目されつつある．これまで声は年齢や体調による変動が大きいことから実用化が遅れていたが，i-vector [3] や PLDA [4] といった最先端の手法を用いることですでに実用化にも十分な識別性能を得られることも報告されている．

一方、音声合成 [5, 6] や声質変換 [7] といった声を作る技術の性能も非常に高くなってきている．なかでも目標話者の声が数文章あれば目標話者の声を自由に作ることができる話者適応技術 [8] は福祉やエンターテイメントなどに活用される重要な技術である．しかしながら，ある特定話者の音声を高い精度で自由に作成可能であるということは話者照合の観点からは非常に難易度の高いなりすまし音声が作成可能であるということを意味する．実際，[9–11] において最先端の話者照合システムにおいても音声合成技術を使ったなりすましが可能であるという報告がされている．これは話者照合の分野では重要な問題となっており，Interspeech2015 では ASVspoof2015 と題したスペシャルセッションでなりすまし攻撃に対する対策に関するコンペティションが開催されたほどである [12]．これらのなりすまし攻撃への対策は使用する特徴量に音声スペクトルだけでなく基本周波数や位相情報など様々な情報を用いること [13, 14] や精度の高いモデルを用いること [15] で行われてきた．しかし，話者照合及び話者適応どちらの技術も話者性を表現するという点で目的は同じであるため，今後どちらの技術も性能が上がっていけばいくほどに話者性を表す部分でなりすまし音声か本人かを判定することが困難になることは想像に難しくない．

そこで，著者らはなりすまし攻撃を話者照合とは別のモジュールで検出する声の生体検知という枠組みを提案してきた [16]．検出すべきなりすまし音声が用意される手法は録音再生や音声合成，声質変換など様々なものが想定されるが，すべてスピーカーで再生されることに着目し，入力音声が実際に人間が発声したものなのかスピーカーで再生された音なのかを判定する枠組みが声の生体検知である．声の生体検知を実現するために人間の発声では自然に発生し，かつスピーカーでは再現不可能な現象を捉えることを目指した．そこで，人間の呼気の影響を受けてマイク内部で発生するポップノイズの有無を生体音声か否かを判定を行うポップノイズ検出に着目し，ポップノイズの検出法としてモノラルおよびステレオチャンネルの手法をそれぞれ提案してきた．各手法において高いポップノイズ検出精度を得たもののそれぞれに利点と問題点があった．そこで本稿では，各手法が持つ問題点を補間するためにステレオおよびモノラルポップノイズ検出を縦列接続したポップノイズの検出法を提案する．また，実験結果より従来の単独のシステムよりもポップノイズの検出精度が大幅に改善することも示す．

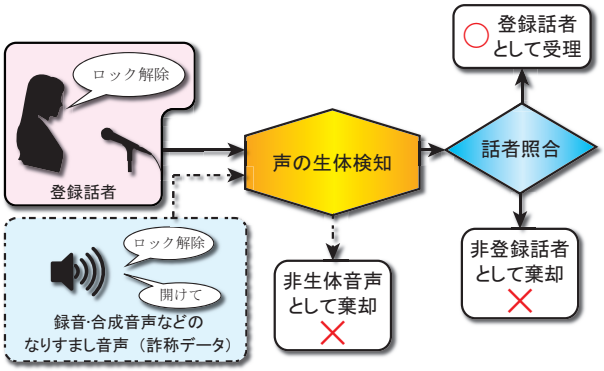

図 1　声の生体検知を含む話者照合システムの全体図

2　声の生体検知法

2.1　話者照合システムへのなりすまし攻撃

登録話者になりすます攻撃方法として主に考えられるものは以下の 3 点である．

再生攻撃　登録話者の声をあらかじめ録音する方法．テキスト依存型に有効

音声合成　任意の文章で登録話者の声を生成可能．テキスト非依存型にも対応可能．

声質変換　別の話者の声を登録話者の声に変換．音声合成と同様に任意のテキストで使用可能．

---

*A tandem double-single channel pop noise detector based voice liveness detection for speaker verification by SHIOTA Sayaka (Tokyo Metropolitan University), Villavicencio Fernando, YAMAGISHI Junichi, ONO Nobutaka, ECHIZEN Isao (Natinal Institute of Informatics) MATSUI Tomoko (The Institute of Statistical Mathematics)

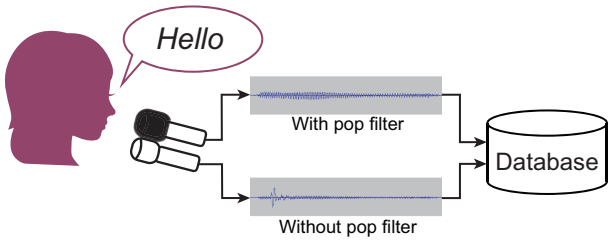

図 2　ステレオ録音時の収録フロー

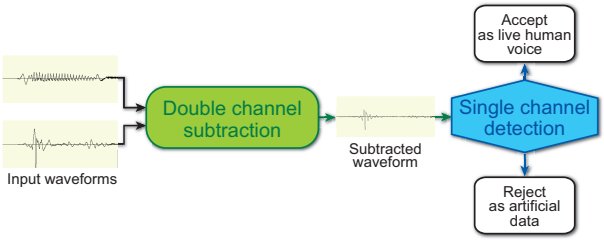

図 3　ステレオ/モノラルポップノイズ検出法の縦列接続による声の生体検知のプロセス

再生攻撃に対してはテキスト提示型の話者照合システムを用いることで防ぐことが可能である [17, 18]．また，合成音声や音声合成を用いたなりすまし攻撃に対抗する手段としてもすでに様々な手法が報告されている [12, 19–22]．しかし，当然ながらこれらの手法は合成音声の性能が人間に近づけば近づくほど識別性能が下がってしまう．そのため，話者照合システムだけで合成音声か登録話者の実音声化を判定するのではなく，別のモジュールとしてなりすまし音声なのか実音声なのかを判定する必要がある．

2.2　声の生体検知法の枠組み

なりすまし攻撃に使われる音声は基本的にスピーカーによって再生される．そこで，入力音声が実際に人間が発話したものなのか，スピーカーによって再生されたものなのかを判定することでなりすまし攻撃を防ぐ，声の生体検知という枠組みが提案された [16]．図 1 に声の生体検知を含む話者照合システムの全体図を示す．声の生体検知は入力音声がスピーカー再生か実発話による音声かを判定するモジュールと考えられるので話者照合の前段処理または，話者照合と同時に処理することが可能である．人が声を発声するメカニズムから呼気が重要な要素であることは明白である．そこで，声の生体検知を実現するために入力音声が実発話によるものなのかスピーカー再生によるものなのかを決定づける要因として呼気に着目する．

3　ポップノイズ検出法

通常，音声を収録するマイクは呼気などの声以外のノイズが入らないようポップフィルタやウィンドスクリーンというフィルタをつけている．マイク内部に呼気が入り込んでしまうと空気の振動を電気信号に変換するための振動板が呼気によって音響的な空気の振動とは異なる振動をしてしまい，ポップノイズと呼ばれるノイズを発生させてしまう．つまり，ポップノイズは人間が発声したからこそ発生してしまうものであると言える．そこで，声の生体検知を実現するための具体的な手法の一つとしてポップノイズ検出法を提案してきた [16]．本節では，始めにこれまで提案してきたモノラルおよびステレオチャンネルによるポップノイズ検出法について紹介し，次に提案法である縦列接続法について述べる．

3.1　モノラルポップノイズ検出法

モノラルマイクによるポップノイズ検出法では，ポップノイズを拾いやすくするためにあえてポップフィルタを装着せずに音声を収録する．ポップノイズは 20〜100msec 程度の短い区間で発生し，特に低周波成分に強いエネルギーを持つ傾向にある．つまり，低周波成分の突発的なエネルギーの変化を検知するでポップノイズの検出が可能となる．これまでの実験において，モノラルポップノイズ検出法は高い検出性能を得られることを報告してきた．しかし，低周波成分の急激なエネルギー変化はポップノイズだけでなく背景雑音などでも起こる可能性があるため，この手法だけではなりすまし攻撃を防ぐことは難しい．

3.2　ステレオポップノイズ検出法

ステレオチャンネルによるポップノイズ検出法では，図 2 のようなポップフィルタを装着したマイクと装着しないマイクを並べた音声のステレオ収録を行う．その結果，ポップフィルタを装着したマイクではポップノイズの影響が少ないクリーンな音声が，ポップフィルタを装着しないマイクではポップノイズを含む音声が収録される．ここで，ポップフィルタあり、なしの録音チャンネルの短時間フーリエ変換（STFT）をそれぞれ $F_x(b,w), F_p(b,w)$ と表す．b は時間フレーム，w は角周波数である．それぞれのマイクの STFT の差分を

$$D(b,\omega) = F_p(b,\omega) - C(\omega)F_x(b,\omega), \qquad (1)$$

と定義すると，差分信号 $D(b,\omega)$ は音声や 2 チャンネル共通で含まれるノイズをほぼ含まずポップノイズ成分がより明確に含まれる信号となる．ただし，$C(w)$ は，ポップフィルタありなしの 2 チャンネル間での周波数特性の違いを補償するフィルタである．得られた差分信号に逆 STFT を行い振幅波形へ戻し，あらかじめ定めた閾値より大きい振幅が存在する区間をポップノイズが存在する区間として検出する．これまでの実験において，ステレオポップノイズ検出法は，モノラルチャンネルより検出精度が低い一方で，非定常雑音やチャンネルノイズなどの低周波成分に乗りやすい雑音を除くことが可能であるため，ポップノイズだけをより正確に検出することが可能であることがわかっている．

表 1　実験条件

データベース	VLD データベース [16]
性別/人数	女性 17 名
生体データ	各話者 30 文章
詐称データ	各話者 31 文章
サンプリング周波数	48kHz
量子化ビットレート	16bit

3.3　ステレオ/モノラルポップノイズ検出法の縦列接続による声の生体検知

ステレオポップノイズ検出法とモノラルポップノイズ検出法はそれぞれ異なる性能と特徴を持つ．そこで，それぞれの特徴を合わせることでより頑健なポップノイズ検出を行うことを提案する．提案法のフローを図 3 に示す．図 3 のように，ステレオ/モノラル両手法を縦列接続することでステレオポップノイズの過程で得られる差分信号をモノラルポップノイズ検出の入力として使用する．このように二つの手法を縦列接続することで，ポップノイズ以外の背景雑音やチャンネルノイズなどの影響を緩和した信号を元にモノラルポップノイズ検出を行うことができるため，よりポップノイズ検出精度を向上させることが可能となる．

4　評価実験

提案法の有効性を示すために従来のステレオ，モノラルそれぞれ単独の検出法と提案法でのなりすまし音声の検出精度を比較した．

4.1　実験条件

実験条件については図 1 に示す．使用したマイクと名称は以下の通りである．

- Camcorder（SONY ECM-DM5P）：ピンマイクタイプのエレクトリックコンデンサマイク
- Voice recorder（SONY ECM-XYST1M）：ビデオカメラなどの外付け用マイク
- Headset (SHURE SM10A-CN)：ヘッドセット用マイク

口からマイクまでの距離は Headset，Camcorder，Voice recorder の順に離れていく．なりすまし音声には HMM に基づく音声合成 [6] における話者適応技術として変分ベイズ法に基づく線形回帰法を用いた [23]．適応データには全話者で共通の発話内容となる 50 文章をポップフィルタ付きのヘッドセットマイクを用いて収録したものを用いた．作成されたなりすまし音声をスピーカ (BOSE 111AD) によって再生したものを音声収録時と同条件で収録したものを詐称データとして用いた．

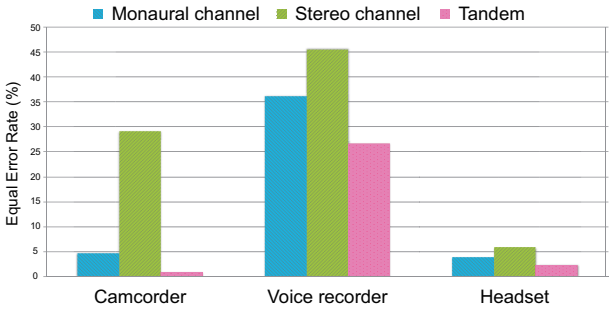

図 4　各手法を使った時の EER

4.2　実験結果

図 4 に従来の単独手法 2 つ（Monaural channel, Stereo channel）及び提案法 (Tandem) の 3 手法において各マイクロフォンを用いた時の等価エラー率 (Equal Error Rate; EER) を示す．ここで，EER とは実際に人間が発話した音声が入力データであるのになりすまし音声として棄却してしまう生体拒否率となりすまし音声を実際に人間が発話した音声として受け入れてしまう再生音声受入率が等価になる点を指す．図 4 より，提案法である縦列接続型のポップノイズ検出法はどのマイクを用いた場合でも従来手法と比較して EER が大きく低下していることがわかる．特にモノラル単独の手法と提案法を比べると，Camcorder を用いた場合 3.78%EER が低下している．従来法においては，手法に限らずマイクに Headset を使用した際の EER がもっとも低くなっていた．その理由として，収録時のマイクと口の距離がもっとも近いことが考えられてきた．しかし，今回の実験において提案法の Camcorder の EER がもっとも低かった．Camcorder はピンマイクタイプであり，ノイズに非常に敏感な性質があったため，実音声の中のポップノイズを検出する精度は非常に高かった．一方で，詐称データにおいてもポップノイズ以外のノイズをポップノイズとして拾ってしまう傾向にあった．そのため，なりすまし攻撃を生体音声として受理してしまうという誤検出が起こってしまい性能が十分に出ていなかったと考えられる．提案法では差分信号を求める際にそれらのチャンネル間で共通に拾えるノイズなどを除くことができ，実際に発生したポップノイズのみを検出することが可能となったことから高い精度を得られるようになったと考えられる．

5　おわりに

本稿では，話者照合のための声の生体検知の改善として，ステレオ/モノラルポップノイズ検出法の縦列接続法について提案した．従来のステレオおよびモノラルポップノイズ検出法では，それぞれ異なる利点と問題点があったが縦列に接続することで，それぞれの問題点を補間した手法となった．生体検知実験においても従来法よりも大幅に EER が低下するこ

とを確認した．今後の課題としては，より大規模な実験データを使用することおよび再生攻撃等のなりすまし攻撃を想定した実験を行うことが挙げられる．また，提案したポップノイズ検出法では発話内にポップノイズの有無だけを判定しているが，ポップノイズの入る音素についても考慮することが挙げられる．

謝辞 本研究の一部は科学研究費基盤（B）26280066 による．

参考文献

[1] A. Jain, P. Flynn, and A. Ross, “Handbook of biometrics,” 2007.

[2] N. Poh and J. Korczak, “Hybrid biometric person authentication using face and voice features,” in *Audio- and Video-Based Biometric Person Authentication*, ser. Lecture Notes in Computer Science, J. Bigun and F. Smeraldi, Eds. Springer Berlin Heidelberg, 2001, vol. 2091, pp. 348–353.

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, May 2011.

[4] S. Prince and J. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, Oct 2007, pp. 1–8.

[5] A. Hunt and A. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1, May 1996, pp. 373–376 vol. 1.

[6] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039 – 1064, 2009.

[7] Y. Stylianou, “Voice transformation: A survey,” in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, April 2009, pp. 3585–3588.

[8] J. Yamagishi and T. Kobayashi, “Average-voice-based speech synthesis using hsmm-based speaker adaptation and adaptive training,” *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 2, pp. 533–543, 2007.

[9] N. Evans, T. Kinnunen, and J. Yamagishi, “Spoofing and countermeasures for automatic speaker verification,” in *Interspeech*, 2013, pp. 925–929.

[10] N. W. D. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. De Leon, *Speaker recognition anti-spoofing*. Book Chapter in ”Handbook of Biometric Anti-spoofing”, Springer, S. Marcel, S. Li and M. Nixon, Eds., 2014, June 2014.

[11] N. K. Ratha, J. H. Connell, and R. M. Bolle, “Enhancing security and privacy in biometrics-based authentication systems,” *IBM Systems Journal*, vol. 40, no. 3, pp. 614–634, 2001.

[12] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilci, M. Sahidullah, and A. Sizov, “Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge,” in *INTERSPEECH*, 2015, pp. 2037–2041.

[13] T. B. Patel and H. A. Patil, “Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech,” in *Proc. Interspeech*, 2015, pp. 2062–2066.

[14] X. Xiao, X. Tian, S. Du, H. Xu, E. S. Chng, and H. Li, “Spoofing speech detection using high dimensional magnitude and phase features: The ntu approach for asvspoof 2015 challenge,” in *Proc. Interspeech*, 2015, pp. 2052–2056.

[15] N. Chen, Y. Qian, H. Dinkel, B. Chen, K. Yu, and S. J. Tong, “Robust deep feature for spoofing detection — the sjtu system for asvspoof 2015 challenge,” in *Proc. Interspeech*, 2015, pp. 2097–2101.

[16] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, “Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification,” in *Interspeech*, 2015, pp. 239–243.

[17] T. Matsui and S. Furui, “Concatenated phoneme models for text-variable speaker recognition,” in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 2, April 1993, pp. 391–394 vol.2.

[18] D. Delacretaz and J. Hennebert, “Text-prompted speaker verification experiments with phoneme specific mlps,” in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2, May 1998, pp. 777–780 vol.2.

[19] L.-W. Chen, W. Guo, and L.-R. Dai, “Speaker verification against synthetic speech,” in *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*, Nov 2010, pp. 309–312.

[20] Z.-Z. Wu, C. E. Siong, and H. Li, “Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition.” in *INTERSPEECH*, 2012.

[21] M. Faundez-Zanuy, M. Hagmller, and G. Kubin, “Speaker verification security improvement by means of speech watermarking,” *Speech Communication*, vol. 48, no. 12, pp. 1608 – 1619, 2006.

[22] M. Nematollahi, S. Al-Haddad, S. Doraisamy, and M. Ranjbari, “Digital speech watermarking for anti-spoofing attack in speaker recognition,” in *Region 10 Symposium, 2014 IEEE*, April 2014, pp. 476–479.

[23] S. Watanabe, A. Nakamura, and B.-H. Juang, “Structural bayesian linear regression for hidden markov models,” *Journal of Signal Processing Systems*, vol. 74, no. 3, pp. 341–358, 2014.