

THE IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS (JAPANESE EDITION)

IEICE | **電子情報通信学会**
D | **論文誌** 情報・システム

VOL. J101-D NO. 3

MARCH 2018

本PDFの扱いは、電子情報通信学会著作権規定に従うこと。

なお、本PDFは研究教育目的（非営利）に限り、著者が第三者に直接配布することができる。著者以外からの配布は禁じられている。

情報・システムソサイエティ

一般社団法人 **電子情報通信学会**

THE INFORMATION AND SYSTEMS SOCIETY

THE INSTITUTE OF ELECTRONICS, INFORMATION AND COMMUNICATION ENGINEERS

話者照合のための音素情報を考慮したポップノイズ検出法による
声の生体検知望月紫穂野^{†a)} 塩田さやか[†] 貴家 仁志[†]Voice Liveness Detection Based on Phoneme Information-based Pop-noise Detector
Shihono MOCHIZUKI^{†a)}, Sayaka SHIOTA[†], and Hitoshi KIYA[†]

あらまし 本論文では、音素情報を考慮したポップノイズ検出法による声の生体検知を提案する。近年、話者照合システムに登録話者の録音した声や合成音声スピーカー再生するなりすまし攻撃が問題となってきた。既になりすまし攻撃に対処するための手法が幾つか提案されているが、それらの手法は様々な音響的特徴量を用いるものが主であり精度が十分ではなかった。そこで、なりすまし攻撃に対する根本的な解決策の一つとして、声の生体検知という入力音声スピーカーで再生されたものなのか人間が実際に話したものなのかを識別する仕組みが提案された。声の生体検知の実現手法の一つとして、入力音声にポップノイズが発生しているかを検出する方法が有用であることが報告されている。しかし、なりすまし攻撃においてもポップノイズとして検出されてしまう部分がありポップノイズ検出だけでは誤って判定してしまうことがあった。そこで、ポップノイズの発生頻度と音素には依存傾向があることを利用し、本研究では音素情報を考慮したポップノイズ検出法を提案する。生体認証実験及び話者照合実験を行うことで、なりすまし攻撃に対する頑健性が向上することを報告する。

キーワード 話者照合, ポップノイズ検出, 声の生体検知, 音素情報

1. ま え が き

近年、声を用いた生体認証システムである話者照合の精度向上に伴い、入室管理や携帯電話のセキュリティシステムやスマートスピーカー等の音声対話システムにおけるユーザ個別サービスを実現する技術としての実用化が期待されてきている。しかしながら、登録話者の声を録音し、再生するなりすまし攻撃や少量の学習データから目標話者の声を作る技術である音声合成 [1], [2], 声質変換 [3] を用いて登録話者を模倣するなりすまし攻撃によって話者照合の精度が大幅に低下してしまうことも報告されている [4]。そのため、話者照合システムの課題として精度向上だけでなく、なりすまし攻撃に対する頑健性向上も重要となり、活発に研究が行われている。これまでに、Interspeech2015 においてスペシャルセッションとして Anti-spoofing

Challenge2015 という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かれ、国内外の多くの研究機関が参加していた [5]。また、Interspeech2017 において続編となる Anti-spoofing challenge2017 が開催されるなど引き続き重要な課題として注目されている。これまでに提案されてきたなりすまし攻撃への対処法は、音響的特徴量として様々な特徴量を用いるものが主であった [6]~[8]。しかし、音声合成や声質変換を用いることで、それらの特徴量をほぼ再現可能となってきた。そのため、話者照合システムのモデル表現や特徴量抽出による対策ではなく、なりすまし攻撃に対する根本的な解決策を考える必要がある。なりすまし音声を用いる際、基本的に音声をスピーカーで再生して音を流すことが考えられる。つまり、システムに入力された音声実際に人間が発声したのか否かを判定する仕組みがあれば、なりすまし攻撃を防ぐことが可能であると期待される。この入力音声を人間が実際に発声したのか否かを判定する仕組みとして声の生体検知が提案されている [9]。また、これまでに声の生体検知を実現する手法として、入力音声にポップノイズが発生しているかを検出する方法が有用

[†] 首都大学東京システムデザイン研究科, 日野市
Department of Information and Communication Systems,
Tokyo Metropolitan University, Hino-shi, 191-0065 Japan

a) E-mail: mochizuki-shihono@ed.tmu.ac.jp

DOI:10.14923/transinfj.2017PDP0017

であることが報告されている．ここでポップノイズとはマイク内部に息や風が入りこむことにより変則的に振動板を揺らしてしまうことで発生してしまうノイズのことを指す [10], [11]．しかし，ポップノイズ検出による声の生体検知を行うことでなりすまし攻撃を高い精度で検出できる一方で，人間の発声においてもポップノイズが生じない場合や再生音声に乗ってしまったノイズの誤検出により検出精度が安定しないという問題があった．

ポップノイズの発生原理と人の発声器官の仕組みから，人間が発声する際にはポップノイズを発生させやすい音素と発生させにくい音素があると考えられる．そこでポップノイズ検出後にポップノイズ区間の音素の出現傾向を考慮したポップノイズ検出を行うことでポップノイズの検出精度が向上すると期待できる．本研究では音素情報を考慮したポップノイズ検出法による声の生体検知を提案する．また，提案法を用い，更にポップノイズの発生頻度のバランスを設計したプロンプト文を提示することで，従来のポップノイズ検出法と比較してポップノイズ検出法の検出精度を向上させることで，声の生体検知と話者照合を組み合わせた際のなりすまし攻撃に対する頑健性が向上することを報告する．

本論文の構成は以下のとおりである．2. では声の生体検知と従来のポップノイズ検出法について述べる．3. では提案法である音素情報を考慮したポップノイズ検出法について説明し，4. では提案法の評価実験を行う．最後に 5. で本論文をまとめる．

2. 話者照合のための声の生体検知 [9]

2.1 ポップノイズ情報を用いた声の生体検知

近年，話者照合に登録話者の声を録音した音声や合成音声などをスピーカーで再生して入力音声とする，なりすまし攻撃が問題となってきている．なりすまし攻撃に対する根本的な解決策として，声の生体検知という入力音声スピーカーで再生されたものなのか人間が実際に発声したものなのかを識別する枠組みが提案された．声の生体検知は図 1 に示すように，話者照合と組み合わせる使用することを想定している．図 1 の例では声の生体検知部で入力された音声信号が実際に人間から発せられたものか否かを識別し，生体であると判定された場合のみ後段の話者照合に入力信号を渡すというフローになっている．これまでに声の生体検知の実現手法として，入力音声にポップノイズが発

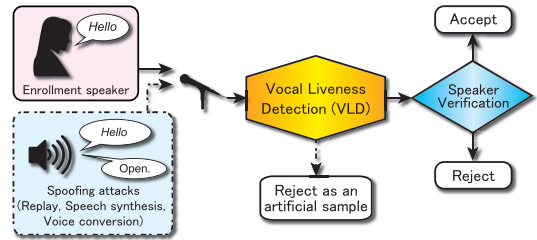


図 1 声の生体検知と話者照合システムのフロー
Fig. 1 Diagram of the voice liveness detection module and automatic speaker verification system.

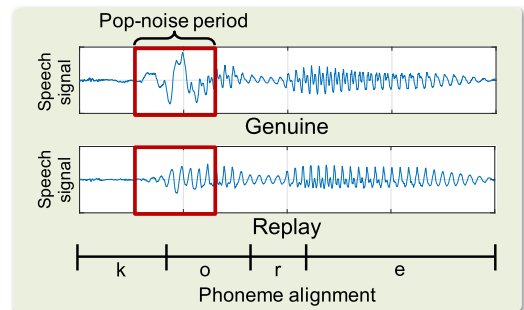


図 2 ポップノイズが発生した実発話（上）及び収録した音声を再生した音声（下）の波形
Fig. 2 Speech signals of a genuine having pop-noises and the replay attack.

生しているかを検出する方法が有用であることが報告されている．ここでポップノイズとはマイク内部に息や風が入りこむことにより変則的に振動板が揺れるために発生してしまうノイズのことを指す [10], [11]．図 2 は実際に人間が発話した音声を収録した実発話と，収録した実発話をスピーカーで再生した再生音声の音声波形をそれぞれ示している．図から，実発話で発生したポップノイズによる波形のひずみが再生音声中では発生していないことが分かる．これは人間が発声する際には呼吸を用いるが，スピーカーは振動板による空気の振動で音を発生させているからであり，スピーカーはポップノイズのような音を再生することはできても同じ現象を起こすことはできないためである．このことから，発話中のポップノイズを検出することが声の生体検知を実現する手段として適切であると考えられる．

2.2 ポップノイズ検出法

本論文では，入力音声のポップノイズを検出するためにシングルポップノイズ検出法 [9] を用いた．ポップノイズは発話内で突発的に起こるノイズのため，局所

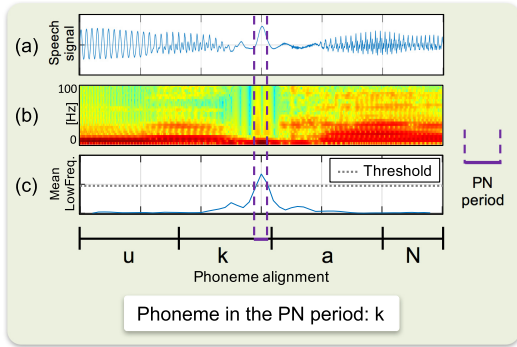


図 3 ポップノイズ区間がある音素の抽出

Fig. 3 Phoneme extraction in the PN (Pop-noise) periods.

的に強いエネルギー変動をもつ性質がある。そのため、シングルポップノイズ検出法ではそのエネルギー変動を捉えることで検出を行う。手順としてはまず、短時間フーリエ変換（分析窓サイズ N 、窓シフト幅 $N/4$ ）を行い、入力音声の周波数分解を行う（図 3 (b)）。

$$X(v, \omega) = \int_{-\infty}^{\infty} x(t)w(t-v)e^{-j\omega t} dt \quad (1)$$

ただし、 x は入力信号、 w は窓関数、 v は時刻、 ω は角周波数を示す。次にフレームごとに低周波領域 $[0, F]$ Hz のパワースペクトルの平均を求める（図 3 (c)）。この平均が低周波成分のエネルギーの推移を表し、フレーム間でのエネルギー変動がしきい値より大きくなる区間をポップノイズが生じている区間として検出する（図 3）。シングルポップノイズ検出法は 1 本のマイクで実現可能であり、導入コストが低く、また話者照合システムとの親和性も高いことが利点としてあげられる。

3. 音素情報を考慮したポップノイズ検出法による声の生体検知

3.1 ポップノイズと音素の依存性

2.2 で述べたポップノイズ検出法を用いた声の生体検知では、なりすまし音声の中にはポップノイズが検出されず、人間の発話にはポップノイズが必ず検出されることを前提としていた。しかしながら、実際には図 4 に示すように実発話においてはポップノイズが発生していないにもかかわらず、再生音声においてポップノイズ区間が検出されてしまう場合があった。これはポップノイズ検出法があくまで低周波領域のパワー変動の軌跡に着目しているため、再生音声からもホワ

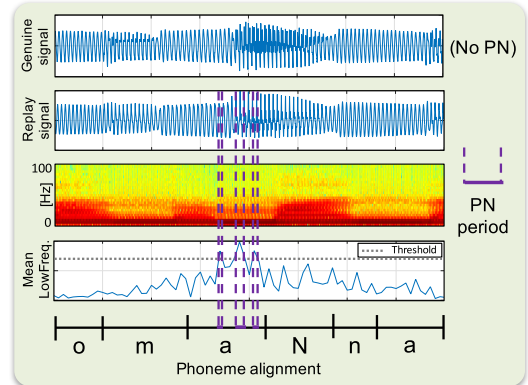


図 4 実発話ではポップノイズが発生していないにもかかわらず再生音声ではポップノイズが誤検出された例

Fig. 4 Speech signal of replay attack that PN period is detected.

イトノイズや背景雑音、不適切なポップノイズ検出しきい値の設定等によりポップノイズを誤検出してしまうことがあるためである。そこでポップノイズの誤検出を減らし、ポップノイズ検出精度を向上させる手法について考える。

ポップノイズの発生原理と人の発声器官の仕組みから、呼吸を使う破裂音や促音といったポップノイズを発生させやすい音素と、無声音のような呼吸をあまり使わない、つまりポップノイズを発生させにくい音素があると考えられる。そこでポップノイズ検出後にポップノイズ区間内の音素の出現傾向を考慮することで、偶発的に若しくは恣意的に発生したポップノイズを棄却することができ、ポップノイズ検出がより高精度になると期待できる。

3.2 ポップノイズの発生頻度と音素の傾向分析

声の生体検知のために収録されたデータベースである VLD データベース [9] を用いてポップノイズとして検出された区間にある音素の傾向を調査した。VLD データベースには、風防カバーを装着しないで収録した音声データが収録されており、風防カバーなしのマイクで収録した音声データにはポップノイズが比較的多く発生している状態を想定している。そこで、傾向調査には風防カバーなしで収録したデータを用いた。ポップノイズ区間にある音素を抽出するための手順は以下に示すとおりである。

- 1: 音声データに対して音声認識を行い、音素アライメントを取得。
- 2: 音声データに対してシングルポップノイズ検出法

を用い、ポップノイズ区間のアライメントを取得。

3:手順1, 2 で得られたアライメント情報を比較して、ポップノイズ区間にある音素を抽出 (図3)。

ここで、ポップノイズを発生させやすい音素を EPN (Easily caused Pop-noise; EPN) 音素、ポップノイズを発生させにくい音素を HPN (Hardly caused Pop-noise; HPN) 音素とする。

3.3 提案法

音素情報を考慮したポップノイズ検出法を用いた声の生体検知について説明する。フローを図5に示す。はじめにシングルポップノイズ検出法を用いて入力音声のポップノイズを検出する。入力音声にポップノイズが発生しているならばその音声を生体による音声として受理する。発生していないならば非生体による音声として棄却する。次に誤受理してしまった再生音声を棄却するために、ポップノイズ区間にある音素情報を用いたポップノイズ検出法を行い、ポップノイズ区間に EPN 音素があるかどうかの判定を行う。3.2 で述べた手順により、ポップノイズ区間に EPN 音素があるならば、それは人による発話によって発生したポップノイズと想定されるため生体として受理する。逆に EPN 音素がないならば、なりすまし攻撃と想定されるため非生体として棄却する。ここで、EPN 音素部分に背景雑音等でポップノイズが発生してしまった場合、なりすまし攻撃を誤受理してしまうことが想定される。このような誤受理を減らすために、EPN 音素情報で生体として受理された音声のポップノイズ区間に、HPN 音素があるかどうかで更に生体検知を行う。人間が発声した場合、HPN 音素の場所ではポップノ

イズが非常に発生しづらい。つまりポップノイズ区間に HPN 音素があることは、人間の発声としては不自然である。そこでポップノイズ区間に HPN 音素がある場合は再生音声として棄却し、HPN 音素がない場合は実発話として受理する。このように音素情報を用いることで、従来のポップノイズ検出で人間の発話を誤棄却しないようしきい値を低く設定した場合に誤受理されてしまう再生音声を棄却することができ、ポップノイズ検出法の検出精度が向上すると考えられる。

4. 評価実験

4.1 音素情報を考慮したポップノイズ検出法による声の生体検知の性能評価

提案法である音素情報を考慮したポップノイズ検出法による声の生体検知の性能を評価するため、人間の発話及び収録した実発話をスピーカーで再生し収録した再生音声を用いて生体検知実験を行った。評価のためポップノイズ検出を従来手法として比較実験を行った。

4.1.1 実験条件

評価のために、実発話となりすまし攻撃用にスピーカーによる再生音声を用意した。収録環境は次のとおりである：

- 収録場所：屋外
- マイク：AKG P170
- 音量：各話者ごとに調節
- 再生用スピーカー：ELECOM LBT-SPP300
- 話者数：5名
- サンプリング周波数：48 kHz

テストデータの発話内容については、ポップノイズの発生頻度を考慮せず JNAS データベース [12] から5文章抜粋し用いた。ただし、5文章全てに EPN 音素が含まれており、4文章に HPN 音素が含まれている。また、発話内容については全話者で共通である。ポップノイズ検出の条件を表1に示す。検出時に用いるしきい値については、実発話が全て受理される最大値で固定した。音素アライメントの抽出には汎用大語彙連続音声認識エンジン Julius [13] のディクテーション

表1 ポップノイズ検出条件

Table 1 Experimental conditions of single PN detection.

Frequency band	[0,50] Hz
# of FFT point	960 points
Window width (N)	20 msec
Window shift (N/4)	5 msec

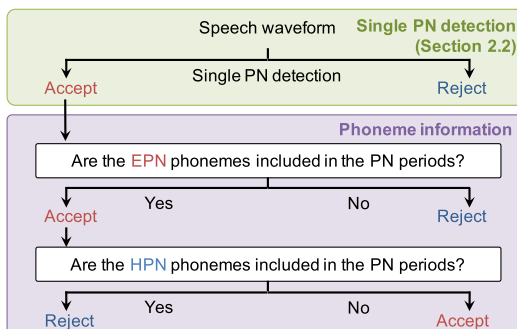


図5 音素情報を考慮したポップノイズ検出法による声の生体検知のフロー

Fig.5 Flow of voice liveness detection based on PN detector considering phoneme information (EPN:easily caused PN, HPN:hardly caused PN).

キット (DNN-HMM 版) を使用した. 実際のシステム運用時を想定し, 認識誤りが含まれている場合にもそのままの認識結果を用いて提案法を行った. ただし, 本実験ではポップノイズが発生している部分に認識誤りはなく, ポップノイズ以外の部分における認識誤りによるポップノイズ検出精度への影響はなかった.

声の生体検知に用いる EPN 音素及び HPN 音素の選択方法は以下のとおりである. VLD データベースの 510 文 (17 話者) に対して 2.2 のポップノイズ検出を行い, 検出したポップノイズ区間にある音素とその音素の全発話中に出現する総数からその音素がポップノイズ区間にある割合を求め, ランキングを作成した. ランキングの上位は特に呼気を使って発音する音素であり, 下位はほぼ呼気を使わないで発音する音素であることを意味する. そこで EPN 音素にはランキング上位 11 音素を選択し, HPN 音素にはランキング下位 3 音素を選択した. 実際に用いた EPN 音素は “t, ky, hy, b, s, sh, k, o:, e:, u:, o”, HPN 音素は “ry, i:, m” である. 評価尺度には以下の生体受率率を用いた.

$$\text{生体受率率} = \frac{\text{生体として受理されたサンプル数}}{\text{全サンプル数}} \quad (2)$$

図 6 に手法ごとの実験フローを示す. 各手法の詳細は以下のとおりである.

(A) ポップノイズ検出: ポップノイズの有無のみで判定. (従来法)

(B) EPN 音素検出: ポップノイズ検出後に EPN 音素情報を用いて判定.

(C) HPN 音素検出: ポップノイズ検出後に HPN 音素情報を用いて判定.

(D) EPN-HPN 音素検出: ポップノイズ検出後に EPN 音素情報を用いて生体検知し, その後 HPN 音素情報を用いて判定.

4.1.2 実験結果

図 7 は各手法ごとの生体受率率を示している. 実発話 (赤) の割合が高く, 再生音声 (青) の割合が低い方が理想的な状態を表す. まず従来法であるポップノイズ検出 (A) の結果をみると, 再生音声の誤受率率が 56% となっていることから, 全ての実発話を受理するようなしきい値のとき, ポップノイズ検出のみでは 6 割弱の再生音声を誤受理してしまうことが分かる. 次に EPN 音素検出 (B) の結果を見ると, ポップノイズ検出に比べ実発話の受率率が 12 ポイント減少した.

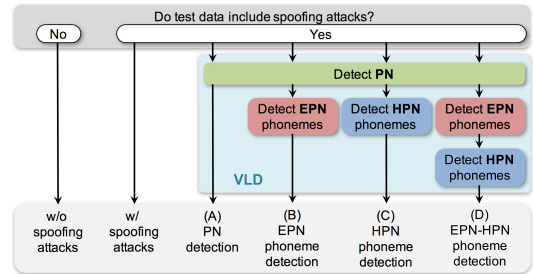


図 6 実験フロー
Fig. 6 Flow of experiments.

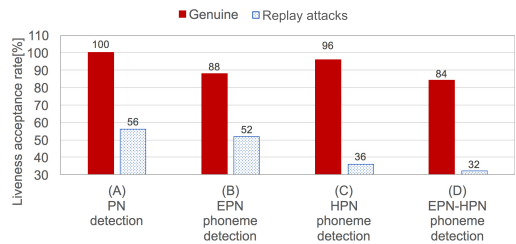


図 7 各手法ごとの生体受率率
Fig. 7 Liveness acceptance rate of each methods.

これはポップノイズ区間に EPN 音素がなく実発話を誤棄却してしまったためである. しかし, 再生音声の誤受率率が 4 ポイント低下したことから, EPN 音素情報を用いることで従来のポップノイズ検出よりも多くの再生音声を棄却できたと分かる. 次に HPN 音素検出 (C) を見ると, EPN 音素よりも生体受率率が高くなっている. また, 再生音声の誤受率率も更に改善されていることが分かる. これらの結果より, ポップノイズの有無による判定よりも音素情報を併せて確認する方がより高い精度で判定可能であることが分かる. 更に, EPN-HPN 音素検出 (D) の結果を見ると, 再生音声の誤棄却率が更に改善されている. この結果より, EPN 音素と HPN 音素どちらも用いることでどちらか一方の音素検出だけでは棄却できなかった再生音声を, 両方の音素情報を用いて検出を行うことで棄却できたことを示す. しかし, EPN-HPN 音素検出においては実発話の受率率も低下しているため, なりすまし攻撃に対する頑健性が向上する一方で, ユーザの利便性が低下してしまう. 実発話の受率率が下がってしまう要因として, 実験で用いたテストデータに設定した HPN 音素や EPN 音素が入っていない文章が含まれていたことが挙げられる. つまり, 音素情報を考慮したポップノイズ検出法による声の生体検知を適切に用いるためには, 読みあげる文章に EPN 音素及び

HPN 音素どちらも適切なバランスで入っている必要がある。

4.2 ポップノイズの発生頻度を考慮した音素バランス文の使用による声の生体検知性能の改善
次に、EPN 音素及び HPN 音素が適切なバランスで入っているプロンプト文を用いた生体検知実験と話者照合実験を行った。評価には声の生体検知による生体受率と話者照合システムとの連結実験による等価エラー率を用いた。

4.2.1 実験条件

比較評価のために、ポップノイズの発生頻度を考慮していないプロンプト文及び考慮しているプロンプト文を用意した。以降、従来プロンプト文及び提案プロンプト文とする。それぞれのプロンプト文ごとに、人が実際に発話した音声を取録した実発話となりすまし攻撃用にスピーカーによる再生音声を用意した。従来プロンプト文には VLD データベースを使用した。提案プロンプト文には 3.2 の予備実験で得られた傾向を元に、EPN 音素及び HPN 音素両方の音素が必ず入る文を設計したものをを用いた。ただし短すぎない身近な読み上げ文となるように配慮した。設計した提案プロンプト文の例を付録 (表 A-1) に示す。提案プロンプト文を用いた取録は次のとおり行った：

- 取録場所：防音室
- マイク：AKG P170
- 音量：各話者ごとに調節
- 再生用スピーカー：ELECOM LBT-SPP300
- マイクとの距離：約 7cm
- 話者数：15 名
- サンプリング周波数：48 kHz

テストデータには従来プロンプト文を用いたデータベースから話者 17 名それぞれに対し実発話 40 文/再生音声 5 文、提案プロンプト文を用いたデータベースからは話者 15 名それぞれに対し実発話 40 文/再生音声 40 文を用意した。発話内容は全話者共通である。ただし使用した音声データは、実発話だけではなく再生音声にもポップノイズが発生している。ポップノイズ検出条件は 4.1 の実験と同様 (表 1) であり、EPN 音素及び HPN 音素も 4.1.1 と同様である。また、ポップノイズ区間にある音素の抽出に用いる音声認識には汎用大語彙連続音声認識エンジン Julius を使用し、モノフォンの音素アライメントを取得した。

話者照合実験には GMM-UBM (Gaussian Mixture Model-Universal Background Model) に基づく話者

表 2 GMM-UBM の実験条件

Table 2 Experimental conditions of speaker verification.

Sampling rate	16 kHz
Bit rate	16 bit
Window width	25 msec
Window shift	10 msec
Feature extraction	MFCC19+E+ Δ + $\Delta\Delta$
UBM database	JNAS (female only)
# of speakers of UBM	153 female speakers
Training data of UBM	165599 sentences
# of mixtures	1024

照合システムを用いた [14]。特徴量抽出及びモデル学習に用いた実験条件は表 2 にまとめてある。また特定話者モデルに関しては上述の新たに収録したデータベースを用いており、話者数 15 名、学習データには上述のテストデータに用いる 40 文章とは異なる 60 文章を用いた。UBM の学習データとしては、JNAS の原音声及び、JNAS の音声に電子協騒音データベース [15] の展示会場の雑音を SN 比が 0, 5, 10, 15, 20, 30dB となるよう重畳した音声を用いて学習した。話者照合の評価尺度には本人棄却率 (False rejection rate; FRR) と他人受率 (False acceptance rate; FAR) が等しくなる点である等価エラー率 (Equal error rate; EER) を用いた。声の生体検知の実験フローは 4.1 と同様である (図 6)。話者認識実験での比較手法の詳細は以下のとおりである。

なりすまし攻撃なし：なりすまし攻撃を含まないテストデータに対し、声の生体検知を行わずに話者照合し、EER を算出。

なりすまし攻撃あり：なりすまし攻撃を含むテストデータに対し、声の生体検知を行わずに話者照合し、EER を算出。

声の生体検知との組み合わせる手法は 4.1 の各手法 (A)～(D) を行った後、話者照合を行い EER を算出した。

4.2.2 実験結果

図 8, 9 に各プロンプト文を用いたときの生体受率を示す。図 8 はポップノイズの発生頻度を考慮していない従来プロンプト文を用い、図 9 はポップノイズの発生頻度を考慮している提案プロンプト文を用いている。まず、従来プロンプト文の生体受率 (図 8) と 4.1 の実験結果 (図 7) を比べたとき、両方ともポップノイズの発生頻度を考慮していない読み上げ文を用い、またポップノイズ検出のしきい値を全ての実発話が受率される最大値に設定しているにもかかわらず

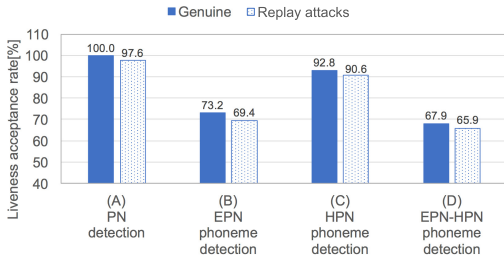


図 8 ポップノイズの発生頻度を考慮していないプロンプト文の生体受率

Fig. 8 Liveness acceptance rate by using non-designed sentences.

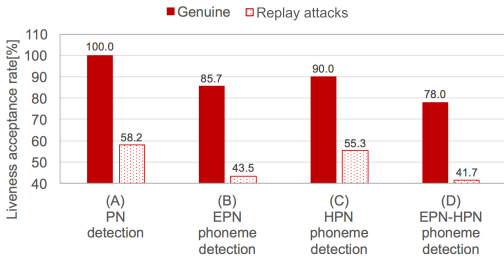


図 9 ポップノイズの発生頻度を考慮したプロンプト文の生体受率

Fig. 9 Liveness acceptance rate by using designed sentences.

ず、従来プロンプト文の方は再生音声をほとんど棄却できていない。これは再生音声において、ポップノイズが発生していないにもかかわらず、低周波領域のエネルギー変動がある区間をポップノイズとして誤検出してしまったためである(図4)。次に、従来プロンプト文と提案プロンプト文の結果を比較する。結果より従来プロンプト文を用いた場合より、提案プロンプト文を用いた方が再生音声の誤受率率が大幅に低下していることが分かる。特にポップノイズ検出(A)による生体検知においては、従来プロンプト文を用いたときあまり再生音声を棄却できていないが、提案プロンプト文の方は多くの再生音声を棄却できていることが分かる。音素情報を用いた声の生体検知(EPN音素検出(B)、HPN音素検出(C)、EPN-HPN音素検出(D))については、再生音声の誤受率率が低くなっており全体的に高い精度を示している。また実発話に対しては、EPN音素検出とEPN-HPN音素検出のとき従来プロンプト文よりも提案プロンプト文の方が生体受率率が高い。これらの結果より、提案プロンプト文を用いることで声の生体検知の頑健性及び利便性が向上することが分かった。

図10は従来プロンプト文、提案プロンプト文それ

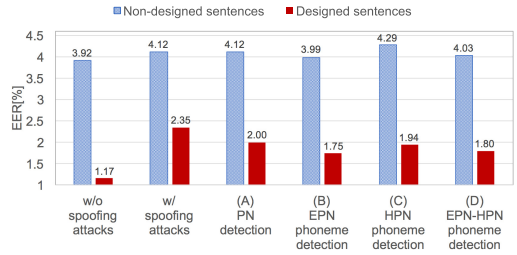


図 10 ポップノイズの発生頻度を考慮していないプロンプト文及び考慮しているプロンプト文の EER

Fig. 10 Respective EERs of speaker verification by using non-designed and designed sentences.

ぞれを用い、声の生体検知と話者照合を行ったときの EER である。青色の棒グラフが従来プロンプト文での EER、赤色の棒グラフが提案プロンプト文での EER を示している。はじめに、なりすまし攻撃なしの EER となりすまし攻撃ありの EER を比較する。従来プロンプト文と提案プロンプト文のどちらにおいてもなりすまし攻撃なしの EER に比べて、攻撃ありの EER の方が高くなっている。このことから、話者照合システムは登録話者の音声を録音再生するなりすまし攻撃に対して脆弱であることが確認できる。次になりすまし攻撃ありの EER とポップノイズ検出による EER を比較する。従来プロンプト文では、攻撃ありの EER とポップノイズ検出(A)による EER では変化がない。一方で、提案プロンプト文ではポップノイズ検出により EER が約 0.27 ポイント改善した。これは、従来プロンプト文では実発話と再生音声間で生じなかったポップノイズ発生との差が、提案プロンプト文では生じたため、ポップノイズ検出により再生音声を棄却することができたためである。次に EPN 音素検出(B)及び HPN 音素検出(C)による EER に着目する。従来プロンプト文を用いたとき、EPN 音素検出の EER が他の手法と比べて最も低い EER が得られ、ポップノイズ検出の EER から約 0.13 ポイント改善した。一方、HPN 音素検出による EER はベースラインであるポップノイズ検出の EER よりも悪化した。これは従来プロンプト文に入っている音素と HPN 音素検出に用いた音素リストが合わず、再生音声だけでなく実発話まで棄却してしまうなど正しく生体検知できなかったためである。提案プロンプト文でも EPN 音素検出の EER が他手法の中で最も低い EER が得られ、ポップノイズ検出の EER と比較して約 0.25 ポイント改善した。また HPN 音素検出による EER も、EPN 音素検出に比べて改善は少ないものの、ポップノ

イズ検出による EER よりも低下した。これはプロンプト文と音素リストが合っていたため、実発話を棄却しすぎることなく、話者照合に影響を与える再生音声を棄却できたことを示している。また、なりすまし攻撃ありの EER から声の生体検知によって最も改善された EER は、従来プロンプト文で約 0.13 ポイント、提案プロンプト文で約 0.6 ポイントであることから、提案プロンプト文を用いることで、声の生体検知と話者照合を統合したシステムがより頑健になるといえる。最後に EPN-HPN 音素検出 (D) のときの EER を見ると、ポップノイズ検出の EER と比較して、従来プロンプト文では約 0.08 ポイントしか改善していないのに対し、提案プロンプト文では約 0.2 ポイント改善している。以上より、ポップノイズの発生頻度を考慮したプロンプト文を用いることで、声の生体検知と話者照合を組み合わせた際のなりすまし攻撃に対する頑健性が向上するといえる。

5. む す び

本論文では、音素情報を考慮したポップノイズ検出法による声の生体検知について提案した。実験結果より、音素情報を考慮しないポップノイズ検出法に比べ、再生音声の誤受率が提案プロンプト文では 16.5 ポイント低下した。またポップノイズ検出法と話者照合システムとを組み合わせた場合には従来プロンプト文より提案プロンプト文の方が全ての条件において EER が大幅に改善しており、また音素情報を考慮しないポップノイズ検出法と比較すると、EER が従来プロンプト文を用いた場合は 0.13 ポイント、提案プロンプト文を用いた場合は 0.25 ポイント低下した。以上より、提案法を用いることでポップノイズ検出法のなりすまし攻撃に対する頑健性が向上することが確認できた。

今後の課題としては話者ごとの音素の傾向調査、またポップノイズが生じた再生音声のテストデータでの実験も行う予定である。

謝辞 本研究の一部は科学研究費基盤 (B) 2628006 による。

文 献

- [1] A.J. Hunt and A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," Proc. ICASSP, pp.373–376, 1996.
- [2] H. Zen, K. Tokuda, and A.W. Black, "Statistical parametric speech synthesis," Speech Commun., vol.51, no.11, pp.1039–1064, 2009.

- [3] Y. Stylianou, "Voice transformation: A survey," Proc. ICASSP, pp.3585–3588, 2009.
- [4] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," Speech Commun., vol.66, pp.130–153, 2015.
- [5] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," Proc. INTERSPEECH, pp.2037–2041, 2015.
- [6] 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," Proc. INTERSPEECH, pp.2062–2066, 2015.
- [7] 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," Proc. INTERSPEECH, pp.2052–2056, 2015.
- [8] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik, and V. Shchemelinin, "STC anti-spoofing systems for the ASVspoof 2015 challenge," Proc. ICASSP, pp.5475–5479, 2015.
- [9] 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," Proc. INTERSPEECH, pp.239–243, 2015.
- [10] G.W. Elko, J. Meyer, S. Backer, and J. Peissig, "Electronic pop protection for microphones," Proc. WASPAA, pp.46–49, 2007.
- [11] Y. Hsu, "Spectrum analysis of base-line-popping noise in MR heads," IEEE Trans. Magnetics, vol.31, no.6, pp.2636–2638, 1995.
- [12] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," J. Acoust. Soc. Jpn. (E), vol.20, no.3, pp.199–206, 1999.
- [13] "汎用大語彙連続音声認識エンジン Julius," <http://julius.osdn.jp/>.
- [14] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted gaussian mixture models," Digital Signal Processing, vol.10, no.1–3, pp.19–41, 2000.
- [15] "音声資源コンソーシアム 電子協騒音データベース," <http://research.nii.ac.jp/src/JEIDA-NOISE.html>.

付 録

設計した提案プロンプト文の例を表 A・1 に示す。

表 A-1 提案プロンプト文の例（太字：EPN 音素，下線：HPN 音素）

Table A-1 Examples of designed sentences.

和文	大通り面したまま睡眠
音素	o: d o: r i m e N s h i t a m a m a s u i m i N
和文	図書館は百万冊
音素	t o s h o k a N w a h y a k u m a N s a t s u
和文	急遽作る民間企業
音素	ky u: ky o t s u k u r u m i N k a N k i g y o:
和文	とても仲良い練習仲間
音素	t o t e m o n a k a y o i r e N s h u: n a k a m a
和文	損失にめげない公民
音素	s o N s h i t s u n i m e g e n a i k o: m i N

（平成 29 年 6 月 7 日受付，10 月 3 日再受付，
12 月 4 日早期公開）



望月紫穂野

2016 首都大学東京システムデザイン学部情報通信システムコース卒業。現在，同大学院システムデザイン研究科情報通信システム学域博士前期課程在学中。日本音響学会学生会員。



塩田さやか（正員）

2012 名古屋工業大学創生シミュレーション工学研究科博士課程修了。同大学にて特別研究員，統計数理研究所での特任助教を経て 2014 年より首都大学東京システムデザイン学部助教。工学（博士）。音声認識，音声合成，話者照合など音声信号処理の研究に従事。日本音響学会，電子情報通信学会，情報処理学会，IEEE，APSIPA，ISCA，各会員。日本音響学会学生・若手フォーラム副代表。



貴家 仁志（正員：フェロー）

首都大・システムデザイン教授。工博。1995～1996 豪シドニー大 Visiting Fellow。信号処理，画像処理，メディアセキュリティに関する研究に従事。本会基礎・境界ソサイエティ会長・ソサイエティ編集長，IEEE 信号処理ソサイエティ日本支部委員長・Regional Director-at-Large for Region 10 を歴任。現在アジア太平洋信号情報処理学会次期会長。本会論文賞（2008）・基礎・境界ソサイエティ功労賞（2008），電気通信普及財団賞テレコムシステム技術賞（2011），メディア学会丹羽高柳賞論文賞（2012）を各受賞。映像情報メディア学会フェロー，IEEE フェロー。