Vol.2017-MUS-115 No.57
2017/6/18

話者照合のための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た声の生体検知

望月 紫穂野[†1,a)]　塩田 さやか[†1]　貴家 仁志[†1]

概要：本稿では，ポップノイズの発生頻度のバランスを考慮したプロンプト文を用いた声の生体検知について提案する．近年，話者照合システムに登録話者の登録した声や合成音声などをスピーカーで再生したものを入力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処するために様々な手法が提案されているがそれらの手法は様々な音響的特徴を用いるものが主であり精度が十分ではなかった．また，なりすまし音声と登録話者の音響的特徴量の差は今後ますます減っていくことが考えられている．そこで，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が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枠組みが提案された．声の生体検知の実現手法の一つとしてポップノイズ区間に含まれる音素情報を用いた検出方法を提案し，高いなりすまし検出精度が得られることを報告した．本研究では，この音素情報を考慮したプロンプト文を実際に提示することで，声の生体検知および話者照合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．

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、プロンプト文

Phoneme information-based pop-noise detection using designed sentence for voice liveness detection and anti-spoofing countermeasure

Shihono Mochizuki[†1,a)]　Sayaka Shiota[†1]　Hitoshi Kiya[†1]

Abstract: This paper proposes a phoneme information-based pop-noise (PN) detection method using designed sentences for voice liveness detection. Recently, many countermeasures against several spoofing attacks (e.g., replay, speech synthesis) have been reported for automatic speaker verification. These techniques use some kinds of acoustical features to capture characteristics of a genuine voice. However, the accuracy of the robustness is not enough. Moreover, the acoustic differences between genuine speech signal and spoofing speech signals expected to become gradually smaller in the near future. As one of fundamental solutions, voice liveness detection (VLD) framework has been proposed. The VLD framework identifies whether an input sample is uttered by a genuine voice or played via a loudspeaker. To realize the VLD framework, PN detection methods have been proposed. However, the conventional PN detection contains vulnerability for wind or arbitrary breathing. In this paper, a phoneme information-based PN detection is proposed to reduce the vulnerability of the conventional PN detection method. Additionally, phoneme balanced-designed sentences are used for prompted-sentences of the VLD module in order to improve the accuracy. As a result, the proposed method can provide better performance than the conventional PN detection methods.

Keywords: Speaker verification, pop-noise detection, voice liveness detection, phoneme information, designed sentence

1. はじめに

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実用性が高まってきている．しかしながら，

[†1] 現在，首都大学東京，システムデザイン学部
Presently with Department of Information and Communication Systems, Tokyo Metropolitan University
[a)] mochizuki-shihono@ed.tmu.ac.jp

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] といった声を作る技術を用いて登録話者を模倣するなりすまし攻撃によって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は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．実際に，Interspeech2015 では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]．これまでに提案されてきたなりすまし攻撃に対処するための手法は，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，話者照合で広く用いられる特徴量をほぼ再現可能となっている．そこで，話者照合システム内でのモデル学習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として入力音声が実際に人間が発声したのか否かを判定する声の生体検知という枠組みが提案された [9]．声の生体検知を実現する手法としてポップノイズ検出法が提案されたが，ポップノイズ検出による声の生体検知では，なりすまし攻撃にはポップノイズが生じていないことを前提としていた．そのため，風などによってポップノイズが発生した再生音声が入力された場合，ポップノイズ検出のみではなりすまし音声を生体として誤受理してしまう可能性があった．これまでに，声の生体検知の頑健性向上のため，ポップノイズ区間内に含まれる音素情報を用いた声の生体検知を提案し，なりすまし検出精度が向上することを報告した [10]．しかし，通常の読み上げ文章ではポップノイズが発生しない場合や，音素情報による判定に用いる音素が文に含まれない場合があった．そこで本研究では，この音素情報を考慮したプロンプト文を実際に提示することで，声の生体検知および話者照合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．

2. 話者照合のための声の生体検知

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．そこで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が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枠組みが提案された．この声の生体検知は図 1 に示すように，話者照合と組み合わせて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れば後段の話者照合に入力信号を渡し，棄却されれば話者照合システムに渡さないというフローになっている．このようにしてなりすまし攻撃に対する話者照合システムの頑健性を向上させることを目指している．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含まれ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人間がマイクに向かって発声する際にマイク内部に息や風が入りこむことで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 [11,12]．

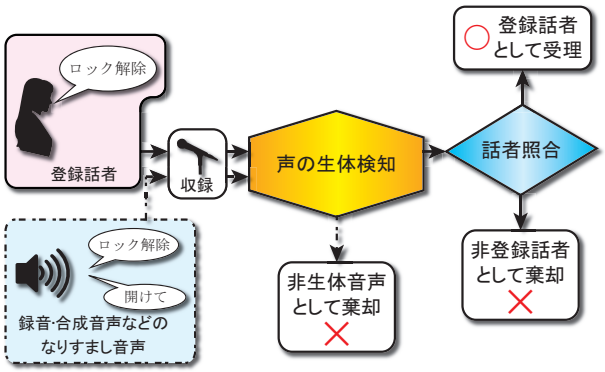

図 1 話者照合システムと声の生体検知のフロー

2.2 ポップノイズ検出法

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おこるノイズのため，突発的な強いエネルギー変動を持つ性質がある．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を行い，入力音声の周波数分解を行う．次にフレーム毎のパワースペクトルの低周波領域のみの平均を求める．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閾値より大きくなる区間をポップノイズとして検出する．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. ポップノイズに含まれる音素情報を用いた声の生体検知

3.1 ポップノイズの音素依存関係

前章で述べたポップノイズ検出法による声の生体検知では，なりすまし音声の中には偶発的にポップノイズが生じていないことを前提としていた．そのため，風などによってポップノイズが発生した再生音声は生体として誤受理されてしまう可能性がある．この問題に対応したより頑健な手法を考える必要がある．

ポップノイズの発生現象と人の発声器官の仕組みから，発声する際にポップノイズを発生させやすい音と発生させにくい音には傾向があると考えられる．そこでポップノイ

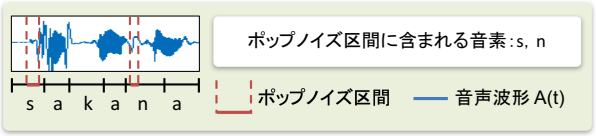

図 2 ポップノイズに含まれる音素の抽出

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した上で，生体音声か再生音声かを判定することでポップノイズ検出がより頑健になると考えられる．

3.2 ポップノイズに含まれる音素の抽出と傾向分析

VLD データベース [9] を用いてポップノイズ区間に含まれる音素の傾向を調査した．ここで 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顕著に発生している状態を想定している．ポップノイズ区間内の音素を抽出するための手順は以下に示す通りである．

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（図 2）．
3： 手順 1，2 で得られたアライメント情報を用いて，ポップノイズ区間に含まれる音素を抽出（図 2）．

調査結果よりポップノイズを発生させやすい音素と，ポップノイズを発生させにくい音素の傾向が得られた．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．調査結果より，EPN 音素を “t，ky，hy，b，s，sh，k，o:，e:，u:，o” とし，HPN 音素を “ry，i:，m” とした．

3.3 ポップノイズに含まれる音素による判定

ポップノイズに含まれる音素情報を用いた声の生体検知について説明する．フローを図 3 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含まれるならばその音声を生体による音声として受理する．含まないならば非生体による音声として棄却する．次にシングルポップノイズ検出法にて生体として受理された音声に対し，ポップノイズが生じた再生音声を棄却するためにポップノイズ区間内に EPN 音素を含むかどうかで生体検知を行う．2.4 節で述べた手順により，もしポップノイズ区間に EPN 音素を含むならば，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．逆に含まないならば，それは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．しかし，EPN 音素部分にポップノイズ区間が生じた再生音声が入力された場合，誤受理してしまうことが想定される．そのような再生音声を棄却するために，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内に，HPN 音素を含むかどうかで更に生体検知を行う．もし HPN 音素を含むならば，そのポップノイズは人による発話と考えにくいため非生体として棄却する．逆に含まないならば生体による音声として受理する．

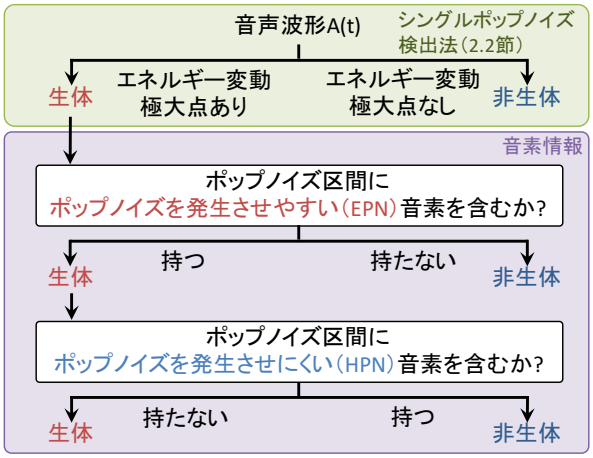

図 3 ポップノイズに含まれる音素情報を用いた声の生体検知のフロー

3.4 ポップノイズの音素バランスを考慮したプロンプト文の設計

ポップノイズに含まれる音素情報を用いた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 音素および HPN 音素が含まれている必要がある．そこで本研究では，声の生体検知の判定に用いる音素の発生頻度を考慮した音素バランス文をプロンプト文として提示することで，ポップノイズに含まれる音素情報を用いた声の生体検知の頑健性を向上させることを目指す．設計したプロンプト文は 3.2 節の予備実験で得られた傾向を元に，EPN 音素および HPN 音素両方の音素が必ず入る文となっており，また，短すぎない身近な読み上げ文となるように配慮した．設計したプロンプト文の例：“大通りに面したまま睡眠，” “図書館は百万冊．”

4. 評価実験

4.1 実験条件

設計したプロンプト文を評価するために声の生体検知および話者照合実験を行った．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および考慮しているプロンプト文それぞれで，人が実際に発話した音声を収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．以降，従来プロンプト文および提案プロンプト文とする．従来プロンプト文には VLD データベースを使用した．提案プロンプト文を用いた収録は次の通り行った：

表 1 ポップノイズ検出および話者照合における実験条件

ポップノイズ検出	
周波数帯域	(0,10] Hz
周波数分解能	5 Hz
分析窓幅	200 msec
窓シフト幅	25 msec
閾値 θ	実発話を 100%受理する値
話者照合	
サンプリング周波数	16 kHz
量子化ビット数	16 bit
特定話者モデル学習データ	各話者 60 文章
UBM データベース	JNAS（女性のみ）
UBM 学習データ	165599 文章
分析窓幅	25 msec
窓シフト幅	10 msec
特徴量	MFCC19 次+Δ+ΔΔ

- 収録場所：防音室
- マイク：SONY ECM-XYST1M（ [9] と共通.）
- 音量：各話者毎に調節
- 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
- 提案プロンプト文数：各話者 40 文

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は話者 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．ただし使用した音声データは，従来プロンプト文・提案プロンプト文ともに同じマイクで収録されたものであり，実発話だけではなく再生音声にもポップノイズが生じている．ポップノイズ検出および話者照合における実験条件を表 1 に示す．話者照合に用いた UBM は，JNAS の音声および，JNAS の音声に電子協騒音データベース [13] 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．また，ポップノイズ区間に含まれ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 [14] を使用し，モノフォンの音素アライメントを取得した．また，ポップノイズが生じた区間についても認識誤りはほぼ発生しておらず，声の生体検知の誤りとなるものはなかった．声の生体検知に用いる EPN 音素および HPN 音素は 3.2 節と同じものを用いた．話者照合の評価尺度には本人棄却率（False rejection rate; FRR）と他人受理率（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．図 4 に手法毎の実験フローを示す。各手法の詳細は以下の通りである：

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，声の生体検知をせずに話者照合し 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，声の生体検知をせずに話者照合し EER を算出.

ポップノイズ検出：なりすまし攻撃を含むテストデータに対し，ポップノイズの有無のみで生体検知した後，話者照合し EER を算出.

EPN 音素検出：なりすまし攻撃を含むテストデータに対し，ポップノイズ検出後に EPN 音素情報を用いて生体検知した後，話者照合し EER を算出.

HPN 音素検出：なりすまし攻撃を含むテストデータに対し，ポップノイズ検出後に HPN 音素情報を用いて生体検知した後，話者照合し EER を算出.

EPN-HPN 音素検出：なりすまし攻撃を含むテストデータに対し，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 音素情報を用いて生体検知した後，話者照合し EER を算出（図 3）.

また，ポップノイズ検出に用いる閾値については，実発話を 100%受理する値に設定した．本提案は誤受理率を減らすための考え方であるため，閾値は低めに設定してある．

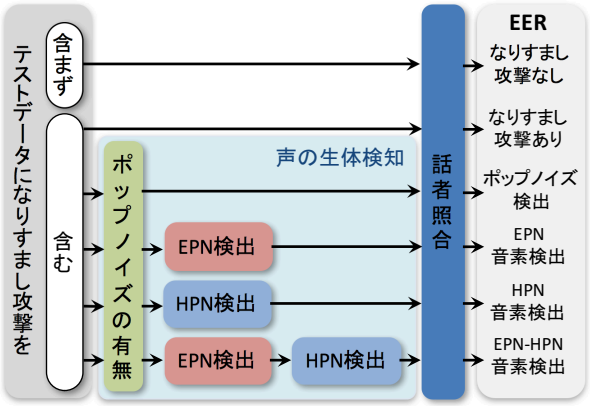

図 4 実験フロー

4.2 実験結果

図 5， 6 に声の生体検知による判定後に生体として受理された文章数の割合を示す．実発話（塗りつぶし）の割合が高く，再生音声（ドット柄）の割合が低い方が理想的な状態を表す．図 5 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 6 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．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理率が大幅に低下していることがわかる．特にポップノイズ検出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．一方，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わかる．音素情報を用いた生体検知（EPN 音素検出，HPN 音素検出，EPN-HPN

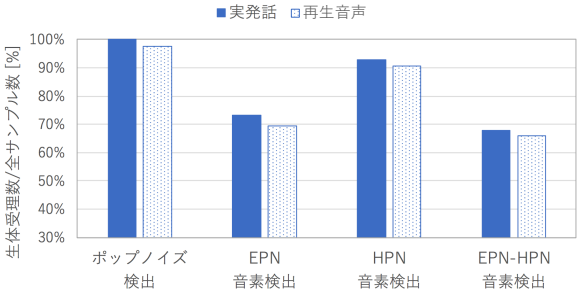

図 5 ポップノイズの発生頻度を考慮していないプロンプト文の生体受理率

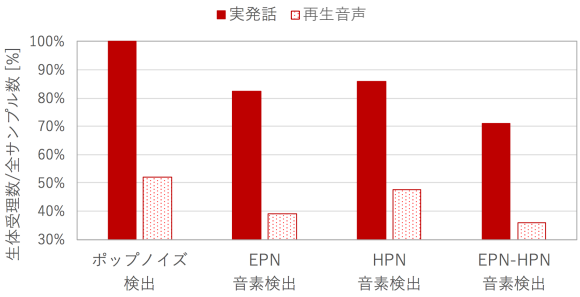

図 6 ポップノイズの発生頻度を考慮しているプロンプト文の生体受理率

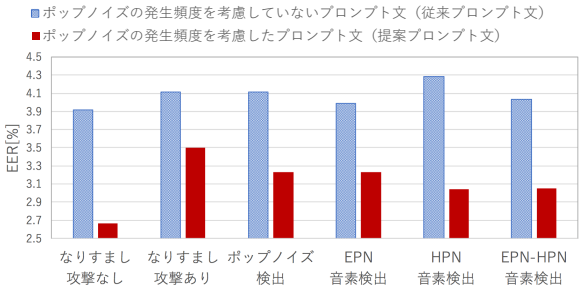

図 7 ポップノイズの発生頻度を考慮していないプロンプト文および考慮しているプロンプト文の EER

音素検出）については，再生音声の誤受理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理率が高い．これらの結果より，提案プロンプト文を用いた方が声の生体検知の精度を向上させることが可能であることを確認できた．

図 7 は従来プロンプト文，提案プロンプト文それぞれを用いた話者照合の 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両方でなりすまし攻撃なしの EER に比べ，攻撃ありの EER の方が高くなっている．このことから，なりすまし攻撃に用いられた再生音声によって話者照合システムの頑健性が低下していることがわかる．従来プロンプト文に比べて提案プロンプト文の方がなりすまし攻撃なしの EER から攻撃ありの EER への増加が大きいのは，用いた再生音声が多いためである．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 音素検出および HPN 音素検出による 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 が他手法の中で最も低い EER が得られ，なりすまし攻撃ありの 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なりすまし攻撃ありの EER よりも悪化した．これは従来プロンプト文に含まれる音素と 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はポップノイズ検出による EER と EPN 音素検出の EER とではほとんど変化がなかった．これはポップノイズ検出の段階で再生音声の多くが棄却されていたため，EPN 音素検出により棄却できた再生音声の数が少なかったためである．一方で HPN 音素検出による EER は，提案プロンプト文を用いたとき他手法の中で最も低い EER を得られ，なりすまし攻撃ありのものと比較して約 0.45 ポイント改善した．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 0.45 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 音素検出のときの EER を見ると，なりすまし攻撃ありの 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 0.45 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，声の生体検知および話者照合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. おわりに

本稿では，ポップノイズの発生頻度のバランスを考慮したプロンプト文を用いた声の生体検知について提案した．実験結果より，設計したプロンプト文を用いることで，声の生体検知および話者照合のなりすまし攻撃に対する頑健

性が向上した．本実験ではなりすまし攻撃にはポップノイズが生じていないことを前提としており，テストデータにはクリーンな再生音声のみを使用した．しかし，実際の話者照合システム運用時にはポップノイズが生じた再生音生が入力される可能性がある．そのため今後の課題として，ポップノイズが生じた再生音声のテストデータの用意，また話者毎の傾向についても調査する予定である．

謝辞 本研究の一部は科学研究費基盤 (B)2628006 による．

参考文献

[1] Andrew J Hunt and Alan W 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, pp. 373–376. IEEE, 1996.

[2] Heiga Zen, Keiichi Tokuda, and Alan W Black. Statistical parametric speech synthesis. *Speech Communication*, Vol. 51, No. 11, pp. 1039–1064, 2009.

[3] Yannis Stylianou. Voice transformation: a survey. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 3585–3588. IEEE, 2009.

[4] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. Spoofing and countermeasures for speaker verification: a survey. *Speech Communication*, Vol. 66, pp. 130–153, 2015.

[5] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov. Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. *Training*, Vol. 10, No. 15, p. 3750, 2015.

[6] Tanvina B Patel and Hemant A Patil. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. In *INTERSPEECH*, pp. 2062–2066, 2015.

[7] Xiong Xiao, Xiaohai Tian, Steven Du, Haihua Xu, Eng Siong Chng, and Haizhou Li. Spoofing speech detection using high dimensional magnitude and phase features: The ntu approach for asvspoof 2015 challenge. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] Sergey Novoselov, Alexandr Kozlov, Galina Lavrentyeva, Konstantin Simonchik, and Vadim Shchemelinin. Stc anti-spoofing systems for the asvspoof 2015 challenge. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 5475–5479. IEEE, 2016.

[9] Sayaka Shiota, Fernando Villavicencio, Junichi Yamagishi, Nobutaka Ono, Isao Echizen, and Tomoko Matsui. Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification. In *INTERSPEECH*, pp. 239–243, 2015.

[10] 望月紫穂野, 塩田さやか, 貴家仁志. 話者照合のためのポップノイズに含まれる音素情報を用いた声の生体検知の検討. 日本音響学会秋季大会, pp. 107–108, 2016.

[11] Gary W Elko, Jens Meyer, Steven Backer, and Jurgen Peissig. Electronic pop protection for microphones. In *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, pp. 46–49. IEEE, 2007.

[12] Yimin Hsu. Spectrum analysis of base-line-popping noise in mr heads. *IEEE transactions on magnetics*, Vol. 31, No. 6, pp. 2636–2638, 1995.

[13] 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.

[14] 汎用大語彙連続音声認識エンジン. http://julius.osdn.jp/.