Vol.2016-MUS-111 No.26
2016/5/21

声の生体検知のためのポップノイズの音素バランスを考慮したプロンプト文についての考察

望月 紫穂野[†1,a)]　塩田 さやか[†1]　貴家 仁志[†1]

概要：本稿では，声の生体検知に用いられるポップノイズの音素バランスを考慮したプロンプト文設計について提案する．近年，話者照合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主に様々な音響的特徴を用いるものが主流であり，精度が十分ではなかった．そのため，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している．また，これまでにポップノイズの発生する音素に偏りがあることがすでに報告されている．そこで本研究では，実際にポップノイズの出現頻度を考慮した音素バランス文を設計し，プロンプト文として提示することで，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：ポップノイズ検出，声の生体検知，音素情報，プロンプト文，話者照合

A study of sentence design based on pop-noise balance for voice liveness detection

SHIHONO MOCHIZUKI[†1,a)]　SAYAKA SHIOTA[†1]　HITOSHI KIYA[†1]

Abstract: This paper proposes the sentence design based on the pop-noise balance for a voice liveness detection framework and speaker verification systems. In recent years, spoofing attacks, (e. g., speech synthesis, voice conversion, replay), has become a serious problem against the speaker verification systems. Some techniques have been proposed to protect the speaker verification systems from these spoofing attacks. However, since these techniques are focused on some kinds of the acoustic features, the accuracy of the robustness is not enough. Thus, the voice liveness detection (VLD) has been proposed. The VLD framework identify that an input sample is uttered by an actual human or a loudspeaker. To realize the VLD framework, the pop-noise detection methods are proposed. Additionally, it has been reported that the phoneme information is important for the pop-noise detection. Thus, this paper proposes the pop-noise balanced sentences in order to improve the pop-noise detection accuracy. The experimental results show that the pop-noise balanced sentences obtained the high performance accuracy than the conventional methods.

Keywords: Detecting pop noise, voice liveness detection, phoneme information, prompt sentence, speaker verification

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．同時に，登録話者の声を録音して再生する再生攻撃や音声合成 [1][2]・声質変換 [3] といった声を作る手法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が報告されている [4]．そこで，話者照合システムの課題として精度向上およびなりすまし攻撃に対する頑健性向上が重要となり，国内外の研究機関で活発に研究が行われている．実際，昨年の Interspeech2015 ではスペシャルセッションとして Anti-spoofing Challenge 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も開かれている [5]．これまでに提案されてきた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][7][8]．しかし音声合成や声質変換といった手法においても，様々な音響的特徴量を用いて合成を行っているため，話者照合システムが使用する音響的特徴量を用いることで簡単に話者

†1 現在，首都大学東京，システムデザイン学部
Presently with Tokyo Metropolitan Uniersity, fuculty of system design
a) mochizuki-shihono@ed.tmu.ac.jp

照合システムを破ることができてしまうという問題があった．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枠組みが提案された [11]．声の生体検知では入力音声が実際に人間から発声されたものなのか，スピーカー等で再生されたものなのかを識別するために，入力音声にポップノイズが含まれているかを検出することが有用であることが報告されている．一方，通常の読み上げ文ではポップノイズが発生しない場合もあるということから，プロンプト文について指定する必要性についても指摘があった．また，実際にポップノイズがどのような音素に含まれやすくまたは含まれにくいといった傾向があることも報告されている [12]．そこで，本研究では実際にポップノイズの出現頻度を考慮した音素バランス文を設計し，プロンプト文として提示することで，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．

2. 声の生体検知 [11]

声の生体検知とは，システムに入力された音声が実際に人間から発声されたものなのか，スピーカー等で再生されたものなのかを識別する枠組みである．図 1 に示すように，話者照合の前段として使用することで，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的としている．これまでに，声の生体検知の実現方法として入力音声にポップノイズが含まれるかを検出する手法が提案されてきた．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により発生してしまうノイズのことを指し，音声収録の際にはポップノイズが音声認識・話者照合などのモデル推定精度低下の要因となるためなるべく入らないように収録されてきた [13][14]．声の生体検知ではあえてポップノイズを収録し，その有無を検出することで高いなりすまし検出精度を得ることが報告されている．しかし，ポップノイズ検出の問題の 1 つとして，入力音声にポップノイズが生じない可能性があることが挙げられている．実際，これまでの報告においてもポップノイズが発生しやすい音素と発生しにくい音素にはそれぞれ傾向があることがわかっている [12]．また，マイクの種類により傾向が多少異なる音素もあるが，マイクの種類に依存することない音素の出現傾向も存在することも報告されている．そのため，ポップノイズ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を用いた生体検知がより頑健なものになると考えられる．

3. ポップノイズの音素のバランスを考慮したプロンプト文の設計

ポップノイズの発生頻度を考慮したプロンプト文を設計するために，まずポップノイズを含む音素の傾向を調査する必要がある．そのためにポップノイズが発生する区間

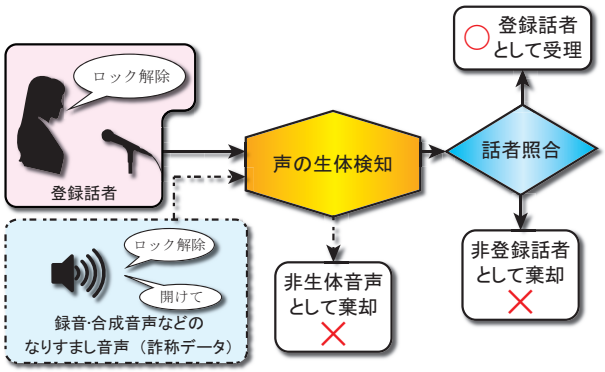

図 1 話者照合システムに対する声の生体検知の概要

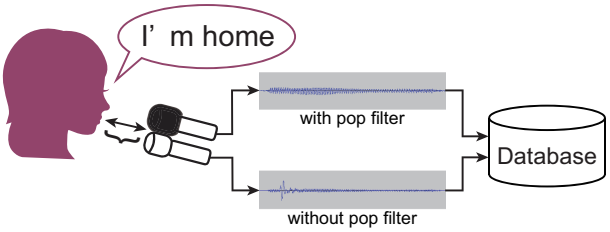

図 2 ステレオ収録のフロー

の音素情報を以下の手順で抽出した．ただし，本研究では図 2 のように 2 本の同じ種類マイクを用意し，片方には風防カバーを装着，もう片方には風防カバーを装着しないで収録したステレオ収録の音声データを使用することを想定している．そのため風防カバーありの音声データに関してはポップノイズがほぼ発生しておらず，風防カバーなしの音声データに関してはポップノイズが発生している状態であり，また，2 チャンネルのデータは同時に収録を行うため時間のずれは生じない．

手順 1： 汎用大語彙連続音声認識エンジン Julius[15] を用いて風防カバーありのマイクにより収録した音声データに対して音声認識を行い，認識結果よりトライフォンの音素アライメントを取得．

手順 2：風防カバーありおよび風防カバーなしの音声データを用いてステレオポップノイズ検出を用い，閾値を調整することでポップノイズを含む区間のアライメントを取得．

手順 3：手順 1，2 それぞれ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．

以上の手順により得られたポップノイズを含みやすいトライフォンをさらに先行音素，中心音素，後続音素それぞれに分けて傾向を調査したところ，先行音素と中心音素に関しては “f，p，sh，z，ts，s”などの息を吐き出して発声する音素に偏る傾向にあり，後続音素には大きな傾向が得られなかった．また，ポップノイズ検出に用いる閾値を下げていった際に，それでもポップノイズとして検出されない音素についても傾向を調査した結果，ポップノイズの発生しにくい音素として “m，d，e:，y，j，a:”などのあまり息

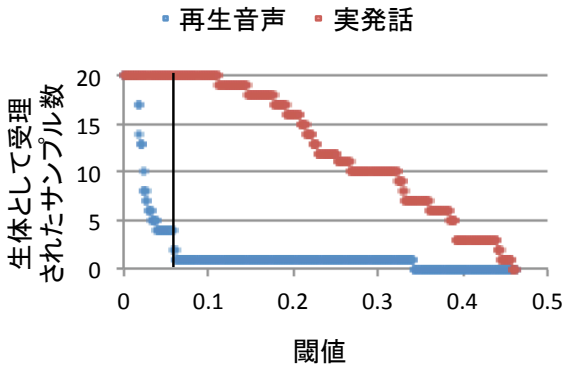

図 3 ポップノイズが発生しやすい文に対してポップノイズを検出する際の閾値と生体として受理されたサンプル数の関係（黒線は再生音声の 10%を生体と誤認識してしまう閾値）

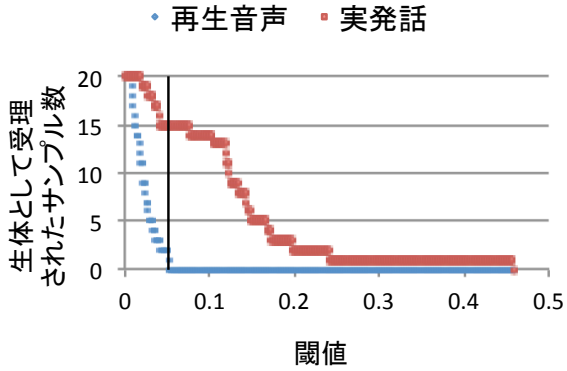

図 4 ポップノイズが発生にくい文に対してポップノイズを検出する際の閾値と生体として受理されたサンプル数の関係（黒線は再生音声の 10%を生体と誤認識してしまう閾値）

を吐き出さなかったり，口を閉じて発声したりする音素が挙げられた．これらの傾向を元に，ポップノイズが発生しやすい音素を多く取り入れた文およびポップノイズが発生しにくい音素を多く用いた文を設計した．その際，短すぎない身近な読み上げ文となるように考慮している．設計したプロンプト文の一例を挙げる．ポップノイズが発生しやすい文：“伸びる散水ホース破損相次ぐ設備不備，”“フレッシャーズ応援新春スーツ割”．ポップノイズが発生しにくい文：“暖冬により青森でウグイス鳴く，”“マニュアルなぜ 5 年後見つかる”．

4. 実験

4.1 実験条件

設計したプロンプト文の性能を評価するために生体検知実験を行った．まずはじめに設計したプロンプト文を用いた音声収録を静かな室内で行った．ポップノイズ検出法にステレオ検出法 [11] を用いたため，マイク（AKG P170）を 2 本使用し，1 本は風防カバーを装着し，1 本は風防カバーを装着しない 2 チャンネル同時収録を行った．マイクの音量は各話者毎マイク毎に調節した．また再生音声に今

表 1 従来法および提案法の FRR，FAR

従来法		提案法		
FRR	FAR	音素	FRR	FAR
5%	10%	先行	5%	5%
		中心	5%	10%

回収録した音声をスピーカー（ELECOM LBT-SPP300）で再生したものを収録したものを用いた．マイクと口およびマイクとスピーカーの距離は約 5 センチとした．使用したデータ数は男性 3 名女性 1 名，文章数は各話者につき肉声 5 文/再生音声 5 文である．サンプリング周波数は 16kHz とした．Julius に用いた音声認識のモデルには，新聞読み上げ文で構成されたディクテーションキットのものを用いた．これは今回設計したプロンプト文が新聞記事に近いためである．音素認識率は 8 割程度であった．ポップノイズ検出の条件としては窓関数にハミング窓（分析窓長：65536 点，窓シフト幅：32768 点）を使用した．評価尺度には生体による音声を非生体として拒絶してしまう生体拒否率（FRR）と非生体による音声を生体として受理してしまう非生体受入率（FAR），FRR 曲線と FAR 曲線が交わる点である等価エラー率（EER）を用いた．比較手法として従来法と提案法の 2 つを用いた．従来法では閾値のみを用いて FRR，FAR を算出した．提案法では従来法で用いた閾値により生体として受理されたサンプルに対して，さらにポップノイズ区間に含まれる音素を用いて生体か否かの識別をした上で FRR，FAR を算出した．その際，ポップノイズ区間に含まれる頻度の高い 10 個の音素を先行音素，中心音素毎に設定し，サンプルのポップノイズ区間にその中の 1 つ以上該当する音素を含む場合，そのサンプルを生体として受理し，そうでない場合は非生体として棄却した．

4.2 実験結果

図 3，4 はそれぞれポップノイズが発生しやすい文およびポップノイズが発生しにくい文の閾値と生体として受理されたサンプル数の関係を示している．図 3，4 を比較すると，ポップノイズが発生しやすい文よりも発生しにくい文の方が，生体として受理されるサンプル数が減少する傾向が，実発話と再生音声で近いことがわかる．またポップノイズが発生しやすい文の方が実発話の生体として受理されたサンプル数の減少がゆるやかである．この結果よりポップノイズの出現傾向は音素に依存していることが確認できた．また，ポップノイズが発生しやすい文の方が，閾値によって肉声と再生音声を区別することが容易になることが確認できた．

次に閾値を再生音声のうち 10%を生体と誤認識してしまう値に定め（図 3，4 中の黒線の値），従来法および提案法の実験した．表 1 に実験結果を示す．提案法の先行音素を

用いた場合と従来法の FRR を比較すると，FRR が減少していることがわかる．これは閾値のみでは受理されてしまっていた非生体音声を，さらに音素情報を用いて判定することで非生体として棄却することができたためである．また，FRR が変化していないことから，生体として受理された生体音声を音素情報によって誤って棄却することがなかったことを示す．次に，先行音素と中心音素の FRR を比較すると先行音素の方が FRR が低いことから，ポップノイズが現れる位置は中心音素よりも先行音素に依存しているがわかった．これらの結果からプロンプト文設計が声の生体検知に有効であることが確認できた．

5. おわりに

本稿では，声の生体検知に用いられるポップノイズの音素バランスを考慮したプロンプト文設計について提案した．実験結果より設計したプロンプト文を収録に用いることで生体か否かをより効果的に識別できることがわかった．この結果より，なりすまし攻撃に脆弱であるテキスト依存型の話者照合に対しても今回提案したプロンプト文を用いることで頑健性が保たれることが期待される．今後の課題として音声データのサンプル数を増やすことやポップノイズが発生しにくい音素の活用方法の検討，またテキスト依存型の話者照合システムとの連結などが挙げられる．

参考文献

[1] A. Hunt and A. Black, "*Unit selection in a concatenative speech synthesis system using a large speech database,*" in Proc. ICASSP, vol. 1, pp. 373-376 vol. 1, May 1996.

[2] H. Zen, K. Tokuda, and A. W. Black, "*Statistical parametric speech synthesis,*" Speech Communication, vol. 51, no. 11, pp. 1039-1064, 2009.

[3] Y. Stylianou, "*Voice transformation: A survey*" in Proc. ICASSP, pp.3585-3588, April 2009.

[4] N. Evans, T. Kinnunen, and J. Yamagishi, "*Spoofing and countermeasures for automatic speaker verification,*" in Proc. Interspeech, pp. 925-929, 2013.

[5] Z. Wu, et al., "*Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge,*" in 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,*" in 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,*" in Proc. Interspeech, pp. 2052-2056, 2015.

[8] Sergey Novoselov, Alexandr Kozlov, Galina Lavrentyeva, Konstantin Simonchik,Vadim Shchemelinin, "*STC Anti-spoofing Systems for the ASVspoof 2015 Challenge,*" in Proc. ICASSP, pp.5475-5479, 2016.

[9] 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.

[10] 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.

[11] 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 Proc. Interspeech, pp. 239-243, 2015.

[12] 塩田 さやか, フェルナンド ビリャビセンシオ, 山岸 順一, 小野 順貴, 越前 功, 松井 知子, 貴家 仁志, "音素情報を考慮した話者照合のための声の生体検知の検討，" 電子情報通信学会　音声研究会，vol.115, no.523, (no.2015-156), pp.347-351, 2016 年 3 月．

[13] G. Elko, J. Meyer, S. Backer, and J. Peissig, "*Electronic pop protection for microphones,*" in Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on, Oct 2007, pp. 46-49.

[14] Y. Hsu, "*Spectrum analysis of base-line-popping noise in mrheads,*" Magnetics, IEEE Transactions on, vol. 31, no. 6, pp. 2636-2638, Nov 1995.

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