社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

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

塩田さやか† FernandoVillaviencio†† 山岸　順一†† 小野　順貴††

越前　功†† 松井　知子††† 貴家　仁志†

† 首都大学東京システムデザイン学部　東京都日野市旭が丘 6-6
†† 情報学研究所　東京都千代田区一ツ橋 2-1-2
††† 統計数理研究所　東京都立川市緑町 10-3

あらまし　本稿では，話者照合に対する再生音声によるなりすまし攻撃を防ぐための枠組みである声の生体検知のためのポップノイズの音素情報の影響について調査する．近年，声を用いた生体認証法である話者照合において合成音声や録音音声の再生によって受けるなりすまし攻撃が重大な問題となっている．なりすまし攻撃への対応として，入力音声が実際に人間が発話したものなのか，スピーカーで再生されたものなのかを判定する声の生体検知が提案された．声の生体検知では，入力音声に息の吹かれにより発生するポップノイズが含まれるかどうかを判定する．このポップノイズ検出法によってなりすまし攻撃を高精度で防ぐことが確認されているが，一方，ポップノイズの有無の判定だけによる検出では悪意のある攻撃者に破られてしまう可能性が高いという問題がある．そこで，本研究では声の生体検知と話者照合の性能評価と同時にポップノイズが含まれる音素について分析しポップノイズの音素情報が声の生体検知の頑健性向上に有用であるか調査し報告する．

キーワード　話者照合，声の生体検知，ポップノイズ，周波数特性, 音素情報

Voice liveness detection using phoneme information for speaker verification

Sayaka SHIOTA†, Fernando VILLAVIENCIO††, Junichi YAMAGISHI††,

Nobutaka ONO††, Isao ECHIZEN††, Tomoko MATSUI†††, and Hitoshi KIYA†

† Tokyo Metropolitan University　6-6, Asahigaoka, Hino, 191-0065, Tokyo
†† National Institute of Informatics　2-1-2 Hitotsubashi, Chiyoda-ku, 101-8430, Tokyo
††† The Insitute of Statistical and Mathematics　10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

Abstract　This paper investigates the effectiveness of phoneme information for voice liveness detection. While the performance of speaker verification systems are recently improved , that of speech synthesis techniques use for spoofing attacks have been improved, too. To overcome this problem, a framework for voice liveness detection has been proposed. By using this technique, the input can be classified as being generated by an actual speaker by means of pop noise detection method. To increase the robustness of the pop noise detection we investigate into an existing correlation between the phonetic context and pop noise events. The experimental results suggest us that the use of phonetic information is effective to increase the robustness of the pop noise detection.

Key words　Speaker verification, voice liveness detection, pop noise, feature characteristics, phoneme information

1. ま え が き

ユーザ認証方法として一般的な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．そのような問題が起こりにくく，他人と共有できない生体認証技術が新たなユーザ認証法として用いられるようになってきた [1,2]．近年，生体情報として特に指紋や顔画像，光彩などを用いた生体認証が普及しつつある．同時に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他の生体認証と比べて認証性能が十分ではないとされ実用化が遅れていたが，i-vector [3] や

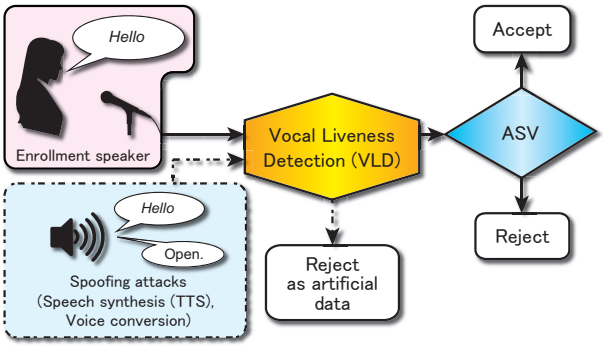

図 1　声の生体検知を含む話者照合システムの全体図

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が報告されている．

一方、音声合成 [5, 6] や声質変換 [7] といった任意の話者の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技術として評価されている．しかしながら，ある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システムに対して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も意味する．実際，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 [9–11]．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，なりすまし攻撃を話者照合とは別のモジュールで識別するために声の生体検知という枠組みを提案された [16]．検出すべきなりすまし音声が作成される手法として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．声の生体検知では人間の発声では自然に起こり，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標としている．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に着目したポップノイズ検出法が提案された．しかし，これらのポップノイズ検出法は入力音声にポップノイズのような信号が含まれているか否かのみを判定しているため頑健性としては十分ではなかった．そこで，本稿では声の生体検知技術の性能向上のために音素情報が有効となるかを調査し報告する．

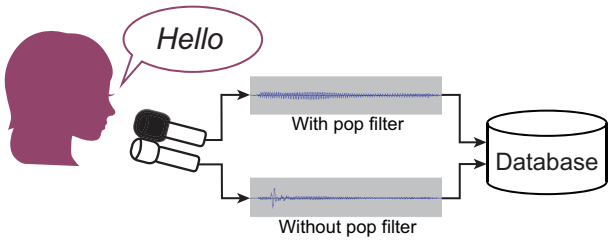

図 2　ステレオ録音時の収録フロー

2.　話者照合のための声の生体検知法

2.1　話者照合システムに対するなりすまし攻撃

話者照合システムへのなりすます攻撃として主に考えられるものを以下に挙げる．

- 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．

再生攻撃に対してはテキスト提示型や発話内容を指定しない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，合成音声を使用する手法は合成音声の話者性が登録話者に近づけば近づくほど話者照合システムとしてはなりすまし攻撃として拒否することが難しくなるため識別性能が下がってしまう．そこで，話者照合システムだけでなりすまし音声なのか登録話者の実音声か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

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ことが多い．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

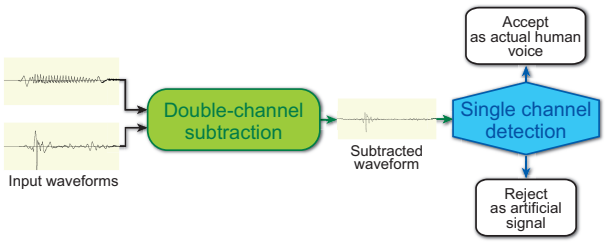

図 3　ステレオ/モノラルポップノイズ検出法の縦列接続による声の生体検知のプロセス

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，モノラル検出法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可能とする縦列接続法を考える．縦列接続法のフローを図 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.　音素情報を利用したポップノイズ検出

前章の手法によるポップノイズ検出では，入力音声にポップノイズが含まれるか否かを判定することだけに着目したが，詐称者がわざと息を吹き替けたり，風によってポップノイズが発生してしまうことも想定される．そこで本研究では，ポップノイズ検出の頑健性を高めるためにポップノイズを含む音素情報の傾向を調査し，ポップノイズの音素情報から詐称者か否かを判定可能であるか分析する．音素情報を考慮するためには入力音声が認識可能であることが望ましい．そこでまず，ステレオポップノイズ検出で使用するポップフィルタを使用したマイクで収録した音声を使用し，以下の手順でポップノイズの音素情報による判定を行う．

（1）　音声認識による音素アライメント付き音素認識結果を作成（ポップフィルタ使用マイク）

（2）　ポップノイズの区間検出

（3）　検出したポップノイズ区間に含まれる音素を抽出

手順 3 で得られる結果に含まれる音素の傾向を調査することで，ポップノイズが発生しやすい音素に傾向が従っているかまた、ポップノイズがほぼ含まれない音素が出現していないかを調査する．

5.　評 価 実 験

5.1　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．

表 2　マイク毎の各ポップノイズ検出法における等価エラー率

	Voice	Headset	Camcorder
モノラル	36.06%	3.95%	4.73%
ステレオ	45.52%	5.88%	29.11%
縦列接続	26.61%	**2.35%**	**0.95%**

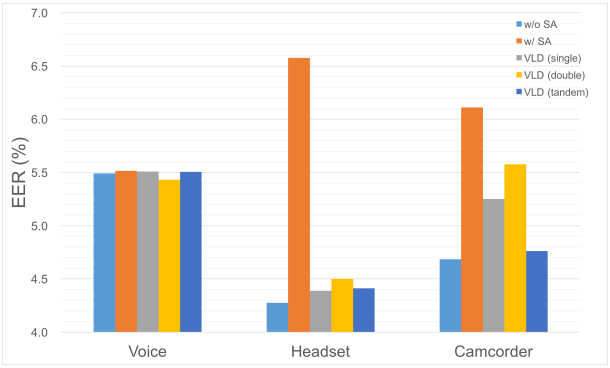

図 4　なりすまし攻撃 (Spoofing Attacks: SA) に対する話者照合および声の生体検知との組み合わせによる話者照合の等価エラー率

- 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
- 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

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]．また，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適応データとした．なりすまし攻撃には作成された合成音声をスピーカ (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使用した．

5.2　実験結果

5.3　ポップノイズ検出法の比較

はじめに，ポップノイズ検出法の性能比較および声の生体検知と話者照合の性能評価について検証する．表 2 に各ポップノイズの検出法の性能比較をマイク毎行った等価エラー率を示す．この表での等価エラー率は人間の発話を再生音声と判定する生体拒否率および再生音声を人間の発話と判定する非生体受理率が等価となる値を指す．結果より，ステレオ/モノラルの縦列接続法の等価エラー率がもっとも低くなっており，縦列接続法の有効性が確認できる．特に Camcorder を使用した際の性能はモノラル法よりさらに 3.78%等価エラー率が低くなっていることがわかる．モノラル法およびステレオ法どちらにおいても，マイクの性能比較では Headset が一番精度が高く，次に Camcorder，そして Voice recorder と続いていた．この順番は各マイクと話者の口との距離に比例していたが，縦列接続法では Camcorder が一番性能が高い．これは Camcorder が 3 種類のマイクの中で一番ノイズに対して敏感であることが原因だと考えられる．縦列接続法は，ポップノイズ検出の前段で 2 チャンネル間の差分を差し引くためチャンネル間共通のノイズを打ち消すことが可能となり，ポップノイズのみを抽出することが可能となり性能が向上したと考えられる．

表 3　ポップノイズが発生しやすい音素・発生しにくい音素

	発生頻度高	発生頻度低
Camcorder	f, hy, ky, h, sh	a:, i:, e:, gy
Voice	f, hy, o: ,b, ky	a:, i:, e:, g
Headset	f, a, i, u, ts	a:, i:, ny, p

次に声の生体検知と話者照合を統合した際の話者照合システムの性能について比較する．表 4 は，話者照合の際の等価エラー率を示す．ここでの等価エラー率は他人受入率と本人拒否率の等価となる点を指す．まず始めになりすまし攻撃（Spoofing Attacks; SA）を含まない話者照合の性能をマイク毎に w/o SA として示す．本実験ではなりすまし攻撃に使用した音声データは Headset で収録されているため，Headset および Camcorder ではなりすまし攻撃が含まれる w/ SA の結果は大幅に等価エラー率が上昇している．しかしながら，声の生体検知 VLD を用いた場合にはポップノイズ検知法に問わず性能が大幅に改善することがわかる．これは Camcorder でも同様の傾向となっている．Voice に関してはなりすまし攻撃に用いたデータとのチャンネルの違いからなりすまし攻撃が有効ではなかったため，VLD の性能も大きな変化はなかったが，なりすまし攻撃を含む場合でも VLD を使った場合は等価エラー率が若干改善していることがわかる．

5.4　音素情報に関する調査

次に，ポップノイズを含む音素情報に関する分析し，考察する．人間の発話とスピーカー再生それぞれの入力データからポップノイズを含むトライフォンを抽出し，先行音素，中心音素および後続音素それぞれの出現傾向を分析した．音素の出現傾向はポップノイズ検出法および閾値設定にも依存するが，実音声の発話を用いた場合には中心音素や先行音素，後続音素に”f”や”hy”といった息を強く吐き出すように発声する音素にポップノイズが多く含まれる傾向にあることがわかった．他にも発声頻度の高い音素および低い音素を表 3 に示す．スピーカー再生音声についても閾値を下げ，誤ってポップノイズとして検出されてしまう音素を抽出したが音素毎の発生頻度に傾向がないため，頑健なポップノイズ検出に音素情報が有用であると期待できる．表 2 において誤って受け入れてしまったなりすまし音声に対して，音素情報を用いた判定を行った結果，Voice recorder については数文章の棄却が可能となり，頑健性が向上することがわかった．しかし，Headset および Camcorder については受け入れてしまったなりすまし音声を音素情報を用いて棄却することはできなかった．その理由として，収録に用いている発話内容が音素バランスを考慮した新聞読み上げ文章であるため，受け入れてしまったなりすまし音声の発話内容にポップノイズが発生しやすい音素が含まれていなかったことがあげられる．そのため，より頑健なポップノイズ検出法のためには音素バランスではなく，ポップノイズの発生頻度を考慮した文章をプロンプト文として提示する必要がある．

6. むすび

本稿では，話者照合のための声の生体検知の頑健性向上のために，ポップノイズを含む音素情報の傾向を分析し調査した．従来のポップノイズ検出は入力音声にポップノイズがあるか否かだけの判定を行っていたが，さらに音素情報を用いることで適切な位置にポップノイズが含まれているかを判定する．分析の結果，音素の発声頻度が高い音素と低い音素の傾向があることが確認できたが実際のポップノイズ検出においては必ずしもポップノイズを発生させやすい音素が含まれいないという問題があり，十分な性能評価ができなかった．しかし，傾向が確認できているため，適切に使用することでよりポップノイズ検出が頑健になることが期待できる．今後の課題として，ポップノイズの発生頻度のバランスを考慮した文章を読み上げ文章として設計することおよびデータ数を増やすことがあげられる．

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. 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*. Springer Berlin Heidelberg, vol. 2091, pp. 348–353, 2001.

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, 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, ICCV. IEEE 11th International Conference on*, pp. 1–8, Oct 2007.

[5] 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.

[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 *Proc. ICASSP*, pp. 3585–3588, April 2009.

[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 *Proc. Interspeech*, pp. 925–929, 2013.

[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.

[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, and M. S. A. Sizov, “Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge,” in *Proc. Interspeech*, pp. 2037–2041, 2015.

[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*, pp. 2062–2066, 2015.

[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*, pp. 2052–2056, 2015.

[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*, pp. 2097–2101, 2015.

[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 *Proc. Interspeech*, pp. 239–243, 2015.

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

[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, pp. 777–780 vol.2, May 1998.

[19] L.-W. Chen, W. Guo, and L.-R. Dai, “Speaker verification against synthetic speech,” in *Proc. ISCSLP*, pp. 309–312, Nov 2010.

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

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

[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*, pp. 476–479, April 2014.

[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.