非音響ノイズを用いた話者照合の検討*

◎塩田さやか（首都大），松井知子（統数研），貴家仁志（首都大）

1　はじめに

近年の生体認証技術では，導入時に特別な機器を必要としないことから音声による生体認証技術である話者照合が注目されている．しかしながら，他の生体認証技術と比べ識別性能に課題があるため，識別性能の向上を目的に様々な研究機関が話者照合の研究を盛んに行っている．話者照合の研究において標準データセットとして広く用いられている NIST SRE [1] データベースは低いサンプリング周波数，低ビット数で収録され，かつ非定常な雑音を含む大規模な電話での会話音声となっている．そのためこれまでの研究では，雑音除去や音声強調といった雑音の影響を軽減する手法や，統計モデルを用いた汎化性能の向上を目指すことで話者照合の性能向上を目指すものが多かった．

一方，音声合成の分野ではブレスノイズと呼ばれるノイズには話者性があると考え，話者ごとのブレスノイズも含めて音声を合成することを行っている [2]．このように，背景雑音や BGM，チャンネルノイズといった従来のノイズとは別に，話者本人が発生源となる話者固有の非定常ノイズが存在することが知られている．そこで本研究では，話者本人が発声させる雑音を話者性を含む非音響ノイズとして捉え話者照合に取り入れることで話者照合手法にそれらがどのような影響を与えるのか検討した．

2　非音響ノイズの話者性

息継ぎやリップノイズといった話者自身が無意識的に発生させてしまうノイズがある．このノイズを音響的特徴を含んでいない非音響ノイズと定義する．背景雑音や風により発生するノイズとは違い，話者自身が発生させることから非音響ノイズに話者性が含まれていると考えられる．本稿では，特に破裂音や促音など息の吹かれ方によって収録時に発生するマイク内の信号のぶれによって起こるノイズに注目する．これまでの音声認識や話者照合等の分野では，ノイズを削減することや音声強調を行うことで性能改善をはかってきたが，本稿ではあえてこのノイズを積極的に活用することを前提として調査を行う．

従来の音声収録では，非音響ノイズや環境ノイズ，風等の影響を軽減するためにマイクにスポンジや風防ポップフィルタを使用することが多かった．今回は非音響ノイズを収録するためにこれらのノイズ除去用カバー（以下風防カバー）を使用しない状況で音声収録を行った．図 1 が同じマイクで風防カバーありなしそれぞれの実際に音声を収録した際の波形である．図 1 の下の波形のとおり，風防カバーがない場合

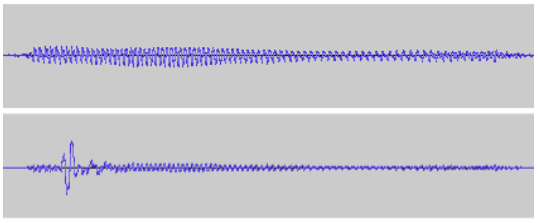

図 1　風防カバーあり（上）となし（下）のステレオマイクで音声を同時に収録した際の波形

には話者の喋り方の影響を受けて波形にノイズのような歪みが発生している．この図のようなノイズを非音響ノイズとし，本収録ではこの非音響ノイズが十分収録可能となるよう収録環境を設定した．話者照合実験としては UBM-GMM を用いたシステムを使用し非音響ノイズを含むモデルと含まないモデル，二つのモデルを結合させてモデルを尤度し，話者照合の実験を行った．二つのモデルの混合比に関しては式 (1) のように設定した．

$$P(X|\Theta_{c+n}) = aP(X|\Theta_c) + (1-a)P(X|\Theta_n). \quad (1)$$

ここで，X は入力特徴量，Θ_c, Θ_n はそれぞれカバーあり，なしモデル，a はモデルの混合比を表す．

3　評価実験

非音響ノイズの話者照合に与える影響について調査するための実験を行った．

3.1　非音響ノイズを含むデータベース

非音響ノイズを含むデータベースを構築するため，本収録ではエレクトレットコンデンサマイクロホン（SONY ECM-DM5P），ステレオマイクロホン（SONY ECM-XYST1M）及びヘッドセットマイク (SHURE SM10A-CN) の 3 種類のマイクを用意した．さらに各種類につき 2 本ずつ用意し，1 本は風防カバーを装着、1 本はカバーを装着しないで収録した．口からマイクまでの距離はそれぞれ 7 センチ，12 センチ，3 センチとした．収録文章は JNAS データベースから音素バランスを考慮した全話者共通の 50 文章と話者ごとにランダムに選択した 50 文章で，各話者合計 100 文章ずつ収録した．話者数は女性のみの 17 名となっている．収録音声に非音響ノイズを含む割合はコンデンサマイクが一番大きく，次にステレオマイク，ヘッドセットマイクという順番になっている．

3.2　実験条件

実験条件等は表 1 のとおりである．収録チャンネル

*Non-acoustical noise for speaker verification by SHIOTA Sayaka (Tokyo Metropolitan University), MATSUI Tomoko (The Institute of Statistical Mathematics), KIYA Hitoshi (Tokyo Metropolitan University)

表 1　実験条件

学習データ （話者依存モデル）	70 文章× 17 名 （計 1190 文章）
テストデータ	30 文章× 17 名 （計 510 文章）
UBM 用データベース	JNAS(女性のみ)
UBM 学習データ	23657 文章
GMM 混合数	1024
サンプリング周波数	8kHz
フレーム長	25msec
フレームシフト	10msec
特徴量	MFCC 19 次+ Δ + ΔΔ

としてはコンデンサマイク (C)，ステレオマイク (S)，ヘッドセットマイク (H) の各マイクに対して風防カバーあり, なし（1,0）の 2 本が存在するためチャンネル数の合計は 6 チャンネルとなる．表記例としてはコンデンサマイクの風防カバーなしの場合 C0，ヘッドセットマイクの風防カバーありの場合 H1 となる．話者識別の基準としては全特定話者モデルにおいて共通の閾値 T を用いる．式 (1) における二つのモデルの混合比は $0 \leq a \leq 1$ で設定した．

3.3　実験結果

表 2 はヘッドセットマイクの風防カバーあり，なしそれぞれのデータを使って学習した特定話者 GMM を用いた，UBM-GMM による話者照合結果の等価エラー率（EER）である．モデルと入力の組合せが H1-H1 の場合が従来の話者照合手法と等しい．また，H1 のモデルに対して H0 の入力すると，非音響ノイズの影響で EER が低下することがわかる．一方，モデルと入力の組合せが H0-H0 の場合，従来法（H1-H1）よりも EER が高くなっている．このことから話者照合において話者が発生させる非音響ノイズを含んだ情報が話者照合の性能向上に一定の効果があることが考えられる．しかしながら，他のマイクによる識別結果を見ると，特に風防カバーなしの場合 EER が大幅に低下していることがわかる．この要因の一つとして，ステレオマイクやコンデンサマイクは非音響ノイズの影響が非常に大きく話者性を捉える以上にただのノイズとして扱われてしまっていることが考えられる．実際，モデルに S0 や C0 を用いた場合，入力のマイク条件がモデルと同じであっても EER が低下していることから非音響ノイズのノイズ成分が強すぎてモデル化が十分に出来ていないと考えられる．そのため，非音響ノイズは話者性を含むことが期待されるが活用するためにはどのような収録条件が必要なのか，どのような特徴抽出やモデル化が適切なのかを十分に検討する必要がある．

また，表 3 にヘッドセットマイクの風防カバーあり、なしそれぞれのモデルを一定の割合で混ぜて話者照合を行った際の結果である．表 2 で識別結果の低かったステレオマイク，コンデンサマイクを入力データとした．モデルの混合比 a を 0 から 1 まで 0.1 刻みで変更し，最も識別率がよかった時の EER とその際の a を示している．結果より，非音響ノイズを含まないモデルのみを用いる場合よりある程度の割合で非音響ノイズを含むモデルを用いる方が識別率が向上していることからも，非音響ノイズが話者性を含むことが考えられる．ただし，上昇の幅は小さいため，やはりテスト環境と収録環境のマイクの性質が大幅に異なる時は非音響ノイズの話者性を十分に捉えられる設定を検討する必要がある．

表 2　話者モデルをヘッドセットマイクで収録したデータで学習した際の等価エラー率（EER）

		入力					
		H0	H1	S0	S1	C0	C1
モデル	H0	**3.77**	4.85	11.64	5.89	14.18	7.00
	H1	4.87	4.28	10.51	4.85	13.17	6.88

表 3　H0,H1 のモデルを混合して用いた等価エラー率（EER）とその際のモデルの混合比

	入力			
	S0	S1	C0	C1
EER	10.38	4.85	12.97	6.67
(a)	(0.8)	(1.0)	(0.5)	(0.2)

4　おわりに

本稿では，非音響ノイズを用いた話者照合の検討として，背景雑音ではなく話者自身が発生する非音響的なノイズを話者照合で識別する要素の一部に用いることを検討した．そのために，話者の呼気により発生するノイズをあえて収録した新たなデータベースを収録し話者照合での有効性と問題点について確認した．今後は，非音響ノイズを適切にモデル化するための特徴量の検討や i-vector を用いた照合への適用などが考えられる．

謝辞　本研究の一部は科学研究費スタート支援 25880026 および科学研究費基盤（B）26280066 による．また，非音響ノイズに関して助言を頂いた国立情報学研究所の小野順貴氏，山岸順一氏に感謝する．

参考文献

[1] NIST Speaker Recognition Evaluation (SRE) http://www.nist.gov/itl/iad/mig/sre.cfm

[2] S. Sundaram, et al., “AN EMPIRICAL TEXT TRANFORMATION METHOD FOR SPONTANEOUS SPEECH SYNTHESIZERS,” in Proc. Interspeech 2003, pp1221–1224, 2003-9.