Vol.2016-SLP-114 No.16
2016/12/20

話者照合のための整数化を用いた位相情報抽出に関する考察

仲野 詩織[1,a)]　塩田 さやか[1]　貴家 仁志[1]

概要：音声信号から特徴抽出を行う際に広く使用されるケプストラム特徴は，周波数分解時に得られる振幅スペクトルのみを用いて計算されている．一方，近年の研究報告により音声知覚で位相スペクトルが有用であることが知られるようになり，音声認識，音声合成，音声照合等，様々な分野で位相スペクトルの活用方法が検討されている．しかし，位相情報はフレーム切り出しの影響や位相情報の計算で起こる位相飛びを考慮する必要があるため有用な特徴抽出を行うことは難しい．そのため位相を正規化する手法や群遅延を位相情報として用いることで位相飛びの影響を回避する手法が提案されている．しかし，計算機で計算される位相スペクトルは計算誤差等の小さな値の変化に対しても余分なスペクトルを発生させてしまうことがある．そこで，本研究では位相情報抽出手法に関して，整数化や簡素化を用いた手法を検討した．また，検討手法によって抽出した位相に基づく特徴量と振幅に基づく特徴量を用いた話者照合実験を行った．実験の結果から，従来の振幅のみを特徴量として用いる手法よりも，振幅と位相を合わせる方が照合精度が向上することを報告する．

キーワード：話者照合，位相抽出，整数位相，UBM–GMM，i–vector

Investigation of integer–based phase information extraction for automatic speaker verification

Nakano Shiori[1,a)]　Shiota Sataka[1]　Kiya Hitoshi[1]

Abstract: Almost all automatic speaker verification (ASV) systems are based on statistical approaches (e.g., GMM, SVM, i–vector). These systems are traditionally assumed that input feature vectors are calculated from mel–frequency cepstral coefficients (MFCCs), which are extracted from a short-time magnitude spectrum. Recently, it has been reported that phase related features perform well in various research topics. However, the phase information changes considerably according to the frame position in an input speech. In addintion, the phase jump is sometimes occurred depended on some calculation methods. It is necessary to normalize the phase response with respect to the frame position. Therefore, this paper investigates the effectiveness of an integer phase extraction and a phase simplification method. The experimental results show that the system combinations of the magnitude spectrum and the phase spectrum are improved the performance than the conventional methods.

Keywords: automatic speaker verification, phase extraction, integerd phase, UBM–GMM, i–vector

1. はじめに

現在の話者照合ではメルケプストラム係数 (Mel–frequency cepstral coefficients; MFCC) のような音声の振幅スペクトルのみを用いて導出される特徴量が一般的に使用されている．一方で，MFCC の抽出過程で得られる位相スペクトルは特徴量としてほとんど使用されていなかった．これは，人間の聴覚系が位相スペクトルに鈍感であり，音声知覚には振幅スペクトルを主に使用していると考えられてきたためである [1]．しかし，音声知覚で位相スペクトルが有用であることが近年報告され [2]，様々な研究分野でその有用性が報告されている．位相スペクトルはフレーム

[1] 首都大学東京
IPSJ, 6–6 Asahigaoka, Hino–shi, Tokyo, 191–0065, Japan
[a)] nakano-shiori1@ed.tmu.ac.jp

切り出しの影響や計算で生じる位相飛びが発生してしまうことが知られており，位相スペクトルを直接使用することは難しい．そこで，位相を正規化する手法 [3] や群遅延を位相情報として用いる手法 [4], [5] などが提案されている．本稿では文献 [3] の位相情報の抽出法に関して更なる検討を行った．文献 [3] の手法で抽出した位相スペクトルは話者の情報を含むと同時に余分なスペクトル成分が発生しており，また，位相情報は計算方法から変動が激しいことが分かった．そのため本稿では整数化を用いて余分な位相を除去することや簡素化を行うことで位相情報の変動を抑えることを提案し，話者照合実験によって検討した手法の有効性を報告する．

2. 話者照合システム

話者照合システムとは入力された音声が登録された話者本人の音声か否かを識別するシステムのことである．本章では統計モデルを用いた話者照合システムとして広く用いられる UBM–GMM(Universal background model–GMM)[6] および i–vector[7] に基づく話者照合について紹介する．

2.1 UBM–GMM

GMM は M 個の単峰性ガウス分布 $p_i(\mathbf{X})$ と混合重み ω_i を掛け合わせた線形重ね合わせで表現される．ここで，登録話者 s を表す GMM は式 (1) のように定める．

$$p(\mathbf{X}|\lambda_s) = \sum_{i=1}^{M} \omega_i p_i(\mathbf{X}). \tag{1}$$

ここで，$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ は特徴ベクトルを表す．また，$\sum_{i=1}^{M} \omega_i = 1$ である．UBM–GMM に基づく話者照合に用いる特定話者モデルの学習はまず，登録話者 s のデータから特徴量を抽出し登録話者 s の GMMλ_s を学習する．次に，不特定話者の平均的なモデルである UBM を事前に学習しておき事後確率最大化 (Maximum a posteriori probability; MAP) 適応法 [8] を用いて登録話者 s の分布に適応させる．最後に，EM アルゴリズムを用いて最適化を行い登録話者 s のモデルを作成する．照合時には登録話者モデル λ_s に対する入力データ $\mathbf{X}$ のフレーム平均対数尤度を式 (2) のように算出する．

$$\log p(\mathbf{X}|\lambda_s) = \frac{1}{T} \sum_{i=1}^{T} \log p(\mathbf{x}_i|\lambda_s). \tag{2}$$

UBM–GMM では照合スコアとして対数尤度を用いる代わりに，特定話者モデル λ_s と不特定話者モデル λ_{ubm} の対数尤度比を用い，予め設定した閾値より大きければ登録話者の音声であると判定する．

2.2 i–vector

i–vector に基づく話者照合は因子分析を用いたモデルアプローチの 1 つである．因子分析で発話データを話者とチャネル依存の全変動 (total variavility; TV) 空間に写像するアプローチである [9]．話者 s の GMMλ_s の平均だけを結合した GMM スーパーベクトル $\mathbf{m_s}$ は因子分析によって以下のように定義される．

$$\mathbf{m_s} = \mathbf{m} + \mathbf{T} \cdot \boldsymbol{\omega}_s. \tag{3}$$

ここで，$\mathbf{m}$ は UBM から得られる話者及びチャネル非依存の GMM スーパーベクトルである．$\mathbf{T}$ は低ランクの矩形行列で，TV 空間を貼る基底ベクトルから構成される．$\boldsymbol{\omega_s}$ が与えられた発話に対する i–vector である．照合時は入力データに対して算出した i–vector($\boldsymbol{\omega_{test}}$) と, 話者モデルとして登録した i–vector($\boldsymbol{\omega_{trg}}$) のコサイン類似度が広く用いられる．

$$score(\boldsymbol{\omega_{trg}}, \boldsymbol{\omega_{test}}) = \frac{\boldsymbol{\omega_{trg}} \cdot \boldsymbol{\omega_{test}}}{|\boldsymbol{\omega_{trg}}||\boldsymbol{\omega_{test}}|}. \tag{4}$$

この照合スコア $score(\boldsymbol{\omega_{trg}}, \boldsymbol{\omega_{test}})$ が予め設定した閾値より大きければ登録話者の音声であると判定する．

3. 位相抽出手法

従来の話者照合では音声特徴量として MFCC が主に使用されており，音声に含まれている位相情報は考慮されていなかった．近年の研究により位相スペクトルも音声信号を表すために必要不可欠な要素であり，様々な研究分野の性能改善に有用な情報を持っていることがわかってきた．しかし，位相スペクトルを特徴量として用いる場合，フレーム切り出しの影響を受けてしまうことなどが知られており，扱いが難しい．そのため群遅延スペクトルを位相情報として用いる手法 [10], [11] や位相を正規化する手法 [3] が提案されている．本稿では文献 [3] をもとに位相抽出手法に関して更なる検討をおこなった．

3.1 Relative phase information[3]

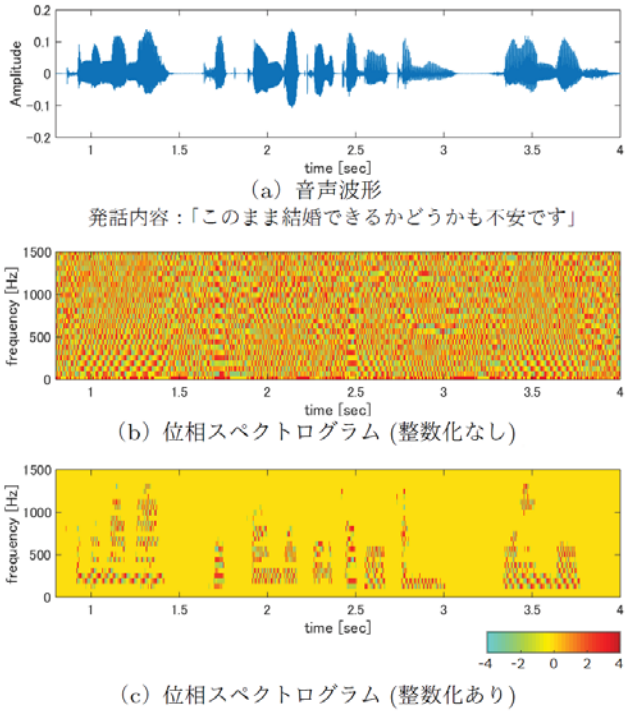
音声信号の離散フーリエ変換は以下の式で表される．

$$\sqrt{X^2(\omega+t) + Y^2(\omega+t)} \times e^{j\theta(\omega+t)}. \tag{5}$$

ここで，ω，t は周波数と時間，X，Y は実部と虚部を表す．$\sqrt{X^2(\omega+t) + Y^2(\omega+t)}$ が振幅スペクトル，$\theta(\omega+t)$ が位相スペクトルである．位相スペクトルは，同じ周波数 ω でもフレーム切り出しの位置によって値が大きく変わってしまう．そこで，式 (6) のようにある基準とする周波数 ω_b の位相を一定にして他の周波数における位相を相対的に求めることで正規化を行う．

$$\tilde{\theta}(\omega+t) = \theta(\omega+t) + \frac{\omega}{\omega_b}(A - (\omega_b + t)). \tag{6}$$

ここで，A は基準周波数 ω_b に設定した位相の値である．本稿では $A = 0$ とする．



(a) 音声波形
発話内容：「このまま結婚できるかどうかも不安です」
(b) 位相スペクトログラム (整数化なし)
(c) 位相スペクトログラム (整数化あり)

図 1　音声波形と位相スペクトログラム

Fig. 1　Speech waveform and phase spectrogram (b)Relative phase spectrogram, (c)Round relative phase spectrogram

3.2　整数化

音声信号のフーリエ変換は機械計算を用いることで手計算では発生しない周波数にも計算精度の限界などで位相情報を持ってしまうことがある．位相の値は $-\pi \sim \pi$ の間になるため，誤差であらわれる値も位相としては大きな値となることがある．この影響を抑えるために位相情報の計算をする際に値を整数化することを検討した．図 1(b) に 3.1 節で抽出した位相スペクトルを，図 1(c) に位相を整数化したとき位相スペクトルを示す．図より，(b) ではノイズ部分にも大きな値の変化が表れており，(c) では余分な部分の値が消えており，音声部分の特徴をより明確に表していることがわかる．

3.3　位相情報の簡素化

位相情報は極座標表現で表すことができる．この時，位相情報 θ は図 2 に示すように，$-\pi \leq \theta < -\frac{\pi}{2}$，$-\frac{\pi}{2} \leq \theta < 0$，$0 \leq \theta < \frac{\pi}{2}$，$\frac{\pi}{2} \leq \theta \leq \pi$ のいずれかの値域に分けることができる．位相情報の値はフレーム切り出しなどの影響による変動が大きい．そこで，3.2 節で整数化した位相特徴をさらに 4 つの領域にわけ，実際の数値を簡素な表現に変えることで位相情報の大きな変動ではなくおおまかな変動のみに着目した特徴抽出を行った．

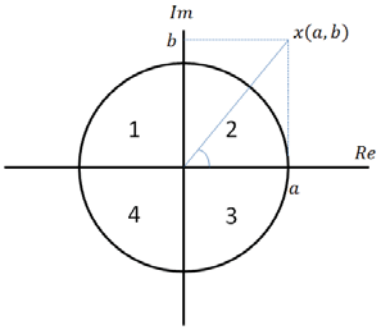


図 2　位相情報の簡素化

Fig. 2　Simplification of phase information

4.　位相情報のモデル化およびシステム統合

4.1　位相情報のモデル化

3 章で述べた抽出法を用いて抽出された位相情報は GMM によってモデル化を行う．位相特徴のみを用いて話者照合実験を行ったところ，位相に基づく GMM の対数尤度 L_{phase} の平均，分散に大きなばらつきがみられた．そこで，以下の式でスコアの正規化を行う．

$$L'_{phase} = \frac{L_{phase} - m}{\alpha V}. \tag{7}$$

ここで，m，V はそれぞれ L_{phase} の平均，分散を表す．また，α は正規化後の分散を補正するパラメータである．

4.2　スコアの統合

本稿では，MFCC を用いた UBM–GMM または i–vector と位相を用いた GMM，2 つのシステムを統合して用いる．話者照合を行う際には，UBM–GMM または i–vector から得られた照合スコアと位相を用いた GMM から得られた対数尤度を以下の式のように線形結合し，統合スコア L^s_{comb} を得る．

$$L^s_{comb} = (1-\beta)L^s_{MFCC} + \beta L^s_{phase}. \tag{8}$$

ここで，L^s_{MFCC} と L^s_{phase} はそれぞれ話者 s の照合スコアと対数尤度であり，β は重み係数である．

5.　実験条件

検討した位相特徴抽出手法の話者照合における有効性に関して考察するために，UBM–GMM および i–vector による話者照合実験を行った．実験結果の比較には算出された照合スコアから本人拒否率と他人受け入れ率を計算し，全話者共通の閾値を設定して求めた等価エラー率（EER）を用いた．話者照合実験では VLD データベース [12] のヘッドセットマイクで収録された音声データを用いて実験を行った．1 回目の収録から 2 回目の収録までの期間は約 3 週間となっている．1 回目の収録データを時期 A，2 回目の収録データを時期 B とする．

表 1　UBM–GMM および i–vector に基づく話者照合の実験条件
Table 1　Experimental conditions for UBM–GMM and i–vector based speaker verification systems

登録話者データベース	VLD データベース (女性のみ)
学習データ （特定話者モデル）	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
i–vector の次元数	400
サンプリング周波数	16 kHz
フレーム長/フレームシフト	25 msec / 10 msec
特徴量	MFCC 19 次+ Δ + ΔΔ

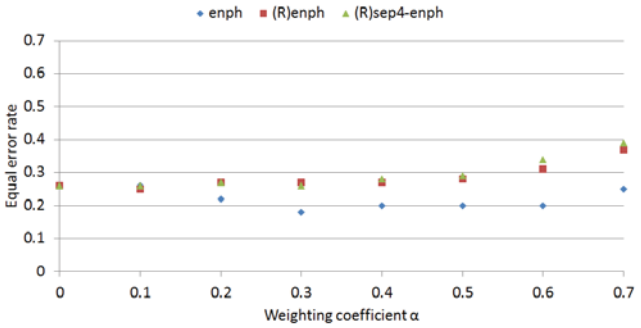
表 2　位相特徴抽出および GMM モデル化の実験条件
Table 2　Experimental conditions for phase feature extraction and GMM modeling

登録話者データベース	VLD データベース
学習データ （特定話者モデル）	70 文章 × 17 名 （計 1190 文章）
テストデータ	30 文章 × 17 名 （計 510 文章）
GMM 混合数	1
サンプリング周波数	16 kHz
使用周波数帯域	60–700Hz

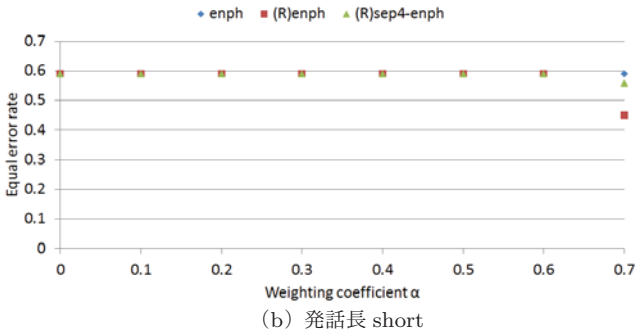
表 3　位相特徴抽出に使用したフレーム長とフレームシフト（msec）
Table 3　Frame length and frame shift used for phase feature extraction(msec)

	フレーム長	フレームシフト
frameleg0	12.5	5
frameleg1	50	25
frameleg2	75	37.5
frameleg3	100	50
frameleg4	500	100

UBM–GMM および i–vector に基づく話者照合の実験条件を表 1，位相特徴抽出および GMM モデル化の実験条件を表 2 にそれぞれ示す．位相特徴は 3 章で示した Relative phase information(enph), 整数化を用いた Relative phase information((R)enph), 整数化および簡素化を用いた Relative phase information((R)sep4–enph) の 3 種類で抽出を行った．また，それぞれの特徴抽出法に対して 5 種類のフレーム長で特徴抽出を行った．そのため，位相特徴は計 15 種類である．位相特徴抽出に使用したフレーム長を表 3 に示す．テストデータには 2 種類の発話長を使用した．データベースのもともとの発話長 (約 4 秒) を original として，発話区間の秒数がおよそ 1 秒となるようにカットした short を作成した．MFCC および位相それぞれの特徴量



（a）発話長 original



（b）発話長 short

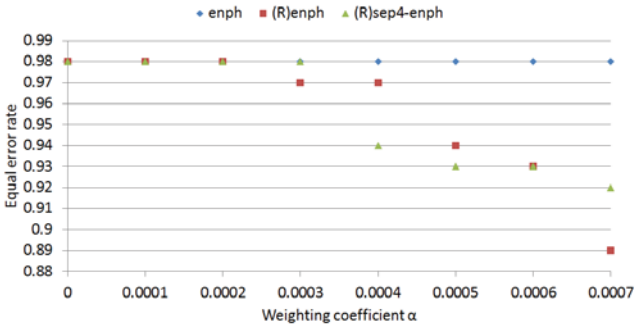
図 3　統合システムの EER(UBM–GMM と位相)
Fig. 3　EERs of integrated systems (UBM–GMM and phase)

を用いて特定話者モデルを学習し，テストデータには学習データと同じ時期と異なる時期に収録したデータを用いた．4.2 節で示した方法で MFCC を用いた UBM–GMM または i–vector から得られた照合スコアと各位相特徴抽出手法に基づく GMM から得られた対数尤度のスコア統合を行い, 文章単位の EER と比較を行った．スコア統合の前処理として位相特徴 (R)enph 及び (R)sep4–enph には 4.1 節で示した方法でスコアの正規化を行った．正規化に用いたパラメータ α はそれぞれ 0.25，0.1 である．

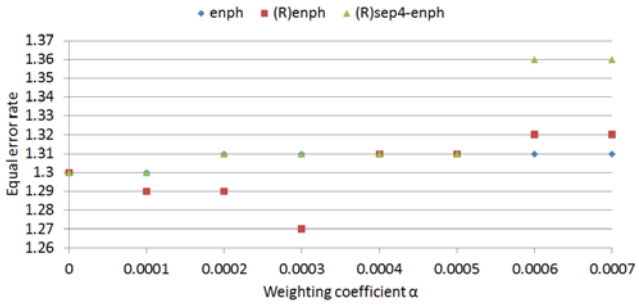
6. 実験結果

6.1 位相特徴と発話長

各発話長に対して MFCC を用いた UBM–GMM を 1 種類，MFCC を用いた i–vector を 1 種類，位相特徴抽出手法に基づく GMM を 3 種類を学習し，各モデルにテストデータを入力して照合スコアを算出した．学習データ，テストデータに共に時期 A を使用している．MFCC を用いた UBM–GMM または i–vector から得られた照合スコアと各位相特徴抽出手法に基づく GMM から得られた対数尤度から統合スコアを算出した．図 3 に UBM–GMM から得られた照合スコアと各位相抽出手法に基づく GMM から得られた対数尤度の統合スコアを用いた際の EER を示す．スコア統合に使用するパラメータ β は 0.1 ～ 0.9 まで 0.1 刻みで変化させた．また，図 4 に i–vector から得られた照合スコアと各位相抽出手法に基づく GMM から得られた対数尤度の統合スコアを用いた際の EER を示す．スコア統合



（a）発話長 original



（b）発話長 short

図 4　統合システムの EER(i–vector と位相)
Fig. 4　EERs of integrated systems (i–vector and phase)

表 4　各位相特徴抽出手法における最小の EER(%)
Table 4　Minimum EER for each phase feature extraction method(%)

（a）UBM–GMM と位相の統合結果

		テストデータ（時期 A）	
		original	short
学習データ（時期 A）	MFCC	0.26	0.59
	MFCC+enph	**0.18**	0.59
	MFCC+(R)enph	0.25	**0.45**
	MFCC+(R)sep4–enph	0.26	0.56

（b）i-vector と位相の統合結果

		テストデータ（時期 A）	
		original	short
学習データ（時期 A）	MFCC	0.98	1.30
	MFCC+enph	0.98	1.30
	MFCC+(R)enph	**0.86**	**1.27**
	MFCC+(R)sep4–enph	0.91	1.30

に使用するパラメータ β は 0.0001 ～ 0.001 まで 0.0001 刻みで変化させた．また，表 4 に図 3，4 の結果で，MFCC のみを用いた場合の EER と MFCC と統合した各位相特徴抽出手法で最も低い EER を示す．

まず，MFCC のみを特徴量として用いた場合と，MFCC と位相の両方を用いた場合の違いを比較する．表 4（a），（b）より，UBM–GMM，i–vector ともに MFCC 単体よりも位相特徴を統合することで EER が改善していることから位相情報の有用さが確認できる．

次に，テストデータの発話長に関して比較する．発話長 original の場合，表 4（a）より UBM–GMM との統合では enph が最小の EER となっている．一方で，表 4（b）から i–vector との統合では (R)enph が最小の EER となっており，enph は改善が見られなかった．発話長 short の場合，表 4（a），(b) ともに (R)enph が最小の EER となっている．このことから，発話長が短いと位相のばらつきが影響してしまうことが考えられる．また，位相の整数化を行うことで余分な成分を除去することができ，少量のデータでも安定したモデル化が可能となったことで EER が改善したと考えられる．

表 5　各フレーム長において最小の EER(%)
Table 5　Minimum EER for each frame length(%)

（a）学習データとテストデータが同じ時期（時期 A-A）

		テストデータ（時期 A）	
		位相抽出手法	EER
学習データ（時期 A）	MFCC	–	0.26
	frameleg0	enph	**0.18**
	frameleg1	enph	**0.18**
	frameleg2	enph	0.21
	frameleg3	enph	0.23
		(R)enph	0.23
	frameleg4	enph	0.23
		(R)sep4–enph	0.23
		(R)sep4–enph	0.23

（b）学習データとテストデータが異なる時期（時期 A-B）

		テストデータ（時期 B）	
		位相抽出手法	EER
学習データ（時期 A）	MFCC	–	1.37
	frameleg0	enph	1.27
	frameleg1	(R)enph	1.31
	frameleg2	(R)enph	1.31
	frameleg3	(R)enph	**1.24**
	frameleg4	(R)enph	1.31

6.2　収録時期およびフレーム長

位相のフレーム長に対する影響を調査するために UBM–GMM による話者照合実験を行った．15 種類 (3 手法×frameleg0～4) の位相特徴抽出手法に基づく GMM を学習し，各モデルにテストデータを入力して対数尤度を算出した．UBM–GMM から得られた照合スコアと各位相特徴抽出手法に基づく GMM から得られた対数尤度から統合スコアを算出し，各フレーム長で最も低い EER を表 5 に示す．スコア統合に使用するパラメータ β は 0.1 ～ 0.9 まで 0.1 刻みで変化させた．表 5（a）は学習データに時期 A，テストデータに時期 A を，表 5（b）は学習データに時期 A，テストデータに時期 B を用いた EER を示している．表中の MFCC の行には UBM–GMM のみでの EER を示している．“位相抽出手法” は統合スコア計算によって MFCC と統合した位相抽出手法のうち最も精度の高かった手法を示している．

まず，MFCC のみを特徴量として用いた場合と，MFCC

と位相の両方を特徴量として用いた場合の違いを比較する．表 5（a），（b）ともに，すべてのフレーム長で MFCC のみよりも位相特徴を統合した場合の方が EER が低くなっている．これは前述の実験と同様の傾向であり，位相情報が特徴として有用であることが確認できる．

次に，フレーム長の種類に関して比較する．位相はフレーム切り出しによって影響を受けるため，フレーム長が長いほどその影響を低減できると考えられる．しかし，表 5（a），(b) から，フレーム長の長さが EER の改善と比例していないことがわかる．一方で，フレーム長の長さとそのとき最小の EER をとった位相抽出手法との関係をみると，特にフレーム長が長い場合（frameleg4）に，整数化や 4 値化した際の位相を用いたものが EER が一番低くなる傾向にある．このことから，位相抽出手法によって適切なフレーム長が異なることが考えられる．

次に，位相特徴量の種類に関して比較していく．表 5（a）では，enph が全ての条件の中で一番低い EER となっている．しかし，frameleg3 および frameleg4 では検討した位相特徴 ((R)enph と (R)sep4–enph) も同程度の EER となっている．つまり，(R)enph と (R)sep4–enph は enph よりも特徴点が少ないが，同様に位相の特徴を表せていると考えられる．

最後に，フレーム長とテストデータの時期の違いに関して比較する．表 5（a）より，学習データとテストデータの時期が同じ場合には frameleg0 および frameleg1 が最小の EER となったが，表 5（b）より，学習データとテストデータの時期が異なる場合には frameleg3 が最小の EER となった．これは同じ発話内容であっても発話時期による変動が大きく，フレーム長を長くとった方が安定した位相抽出が可能になるためだと考えられる．各フレーム長で最小の EER となった場合の位相抽出手法に関して比較すると，表 5（a）より，学習データとテストデータの時期が同じ場合には従来手法である enph が全フレーム長の中で最小の EER となり，フレーム長が長い場合のみ (R)enph または (R)sep4–enph が最小の EER となった．一方で，表 5（b）より，学習データとテストデータの時期が異なる場合には frameleg0 を除く全てのフレーム長で (R)enph が一番低い EER となった．このことから，提案手法である (R)enph は位相特徴の頑健性を向上させることができていると考えられる．

7. おわりに

本稿では特徴量として近年注目されている位相情報の抽出法のより適切な抽出手法について検討を行った．検討した抽出手法によって得た位相情報に基づく特徴量が有効であるかを調査するために UBM–GMM および i–vector を用いた話者照合実験を行った．実験結果では学習データとテストデータが同時期のものであれば MFCC のみを使用した場合よりも，MFCC と位相を合わせて使用した場合の方が良い結果が得られた．今後の課題としては，発話長と発話時期の違いについての検討や位相のモデル化手法の検討，他の位相抽出手法の検討などがあげられる．

謝辞　本研究の一部は科学研究費基盤 (B)26280066 および科学研究費若手 (B)93008552 による．

参考文献

[1] Zhu, D. and Paliwal, K. K.: Product of power spectrum and group delay function for speech recognition, *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, Vol. 1, IEEE, pp. I–125 (2004).

[2] Paliwal, K. K. and Alsteris, L. D.: Usefulness of phase spectrum in human speech perception., *INTERSPEECH* (2003).

[3] Wang, L., Yoshida, Y., Kawakami, Y. and Nakagawa, S.: Relative phase information for detecting human speech and spoofed speech, *Proc. Interspeech*, pp. 2092–2096 (2015).

[4] Hegde, R. M., Murthy, H. A. and Gadde, V. R. R.: Significance of the modified group delay feature in speech recognition, *IEEE Transactions on audio, speech, and language processing*, Vol. 15, No. 1, pp. 190–202 (2007).

[5] 山本一公，末吉英一，中川聖一：長時間分析に基づく位相情報を用いた音声認識の検討 (認識, 理解, 対話, 一般)，電子情報通信学会技術研究報告. SP, 音声， Vol. 110, No. 143, pp. 31–36 (2010).

[6] Reynolds, D. A., Quatieri, T. F. and Dunn, R. B.: Speaker verification using adapted Gaussian mixture models, *Digital signal processing*, Vol. 10, No. 1, pp. 19–41 (2000).

[7] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P. and Ouellet, P.: Front-end factor analysis for speaker verification, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 4, pp. 788–798 (2011).

[8] Povey, D., Chu, S. M. and Varadarajan, B.: Universal background model based speech recognition, *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, pp. 4561–4564 (2008).

[9] 小川哲司，塩田さやか：i-vector を用いた話者認識，日本音響学会誌， Vol. 70, No. 6, pp. 332–339 (2014).

[10] Yegnanarayana, B. and Murthy, H. A.: Significance of group delay functions in spectrum estimation, *IEEE Transactions on signal processing*, Vol. 40, No. 9, pp. 2281–2289 (1992).

[11] Correia, M. J., Abad, A. and Trancoso, I.: Preventing converted speech spoofing attacks in speaker verification, *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*, IEEE, pp. 1320–1325 (2014).

[12] Shiota, S., Fernando, V., Yamagishi, J., Ono, N., Echizen, I. and Matsui, T.: Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification, *in Proc. Interspeech 2015* ((accepted), 2015).