話者照合のための発話長を考慮した位相整数化に関する検討*

☆仲野詩織, 塩田さやか, △貴家仁志 (首都大東京)

1 はじめに

話者照合システム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抽出される特徴量が一般的に使用されている．近年，音声知覚において位相スペクトルが有用であることが報告され [1]，様々な研究分野でその有用性が報告されてきている．しかし，位相スペクトルはフレーム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 [2] や群遅延を位相情報として用いる手法 [3, 4] などが提案されている．本稿では文献 [2] の位相情報の抽出法をもとに更なる位相抽出手法について検討を行った．文献 [2] 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，位相情報は計算誤差などの微小な値に対しても変動が大きいことが分かった．これまでに位相の整数化や簡素化により位相情報の変動を抑えることが提案されたが [6]，本稿ではさらに，発話長や発話時期，フレーム長などに対する影響について調査し，報告する．

2 位相抽出手法

これまでの話者照合では入力特徴量として振幅から得られる 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あまり考慮されていなかった．近年の研究報告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，音声認識や音声合成などの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など，扱いが難しいことが知られている．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や位相を正規化する手法 [2] が提案されている．本稿では文献 [2] をベースラインとし，位相抽出手法とその性能について調査を行った．

2.1 Relative phase information[2]

音声信号の離散フーリエ変換は以下の式で表される．

$$\sqrt{X^2(\omega,t)+Y^2(\omega,t)}\times e^{j\theta(\omega,t)}. \qquad (1)$$

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 (2) 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．

$$\tilde{\theta}(\omega,t)=\theta(\omega,t)+\frac{\omega}{\omega_b}(A-(\omega_b,t)). \qquad (2)$$

ここで，A は基準周波数 ω_b に設定した位相の値である．

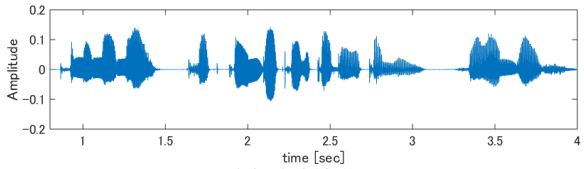
2.2 整数化

音声信号のフーリエ変換は機械計算を用いることで本来は値が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．特に，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 2.1 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の位相スペクトルを示す．図より，(b) では音声区間以外にも大きな値の変化が表れているが，(c) では音声区間のみに情報が現れており，音声部分の特徴をより明確に表していることがわかる．

2.3 位相情報の簡素化

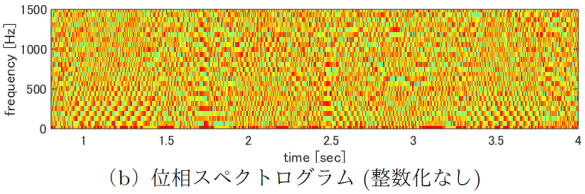
位相情報は極座標表現で表すと図 2 のように表現される．ここで，全領域を $-\pi \le \theta < -\frac{\pi}{2}$，$-\frac{\pi}{2} \le \theta < 0$，$0 \le \theta < \frac{\pi}{2}$，$\frac{\pi}{2} \le \theta \le \pi$ の 4 つの領域に分けることを考える．前節までで述べたように位相情報の値はフレーム切り出しなどの影響による変動が大きい．そこで，2.2 節の整数化位相をさらに上記の 4 つの領域にわけ，実際の

*Investigation of integer-based phase considering short utterance for automatic speaker verification.
by NAKANO, Shiori, SHIOTA, Sayaka, KIYA, HITOSHI (Tokyo Metropolitan University)

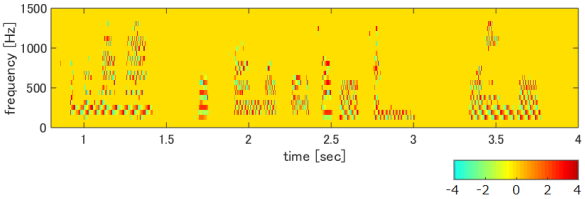


(a) 音声波形

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)

Fig. 1　音声波形と位相スペクトログラム

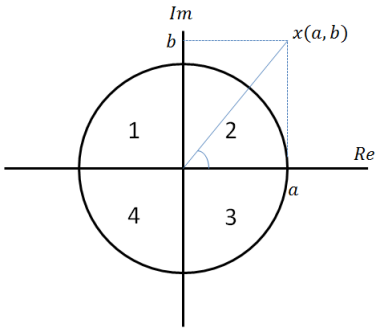


Fig. 2　位相情報の簡素化

数値をさらに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．

3　モデル学習およびシステム統合

3.1　位相情報のモデル学習

2 章で述べた抽出法を用いて抽出された位相情報をもとに GMM によるモデル学習を行う．照合時には位相に基づく GMM に対する入力音声のフレーム平均対数尤度 L_{phase} の正規化を行い，照合スコアとして用いる．

$$L'_{phase} = \frac{L_{phase} - m}{\alpha V}. \qquad (3)$$

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．

Table 1　UBM–GMM に基づく話者照合システムの実験条件

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
サンプリング周波数	16 kHz
フレーム長	25 msec
フレームシフト	10 msec
特徴量	MFCC 19 次+ Δ + ΔΔ

3.2　スコア統合

本稿では，振幅スペクトルから抽出される MFCC を用いた UBM–GMM に基づく話者照合システムと位相を用いた GMM に基づく話者照合システム，2 つのシステムの照合スコアを統合して用いる．話者 s に対して照合を行う際には，UBM–GMM から得られた照合スコア L^s_{MFCC} と位相を用いた GMM から得られた照合スコア L^s_{phase} 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．

$$L^s_{comb} = (1-\beta)L^s_{MFCC} + \beta L^s_{phase}. \qquad (4)$$

ここで，β は重み係数である．

4　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，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 L^s_{comb} 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．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．特定話者モデルの学習には VLD データベース [5] のヘッドセットマイクで収録された音声データを用いた．収録は 2 回に分かれており，1 回目と 2 回目の間の期間は約 3 週間となっている．本稿では 1 回目を時期 A，2 回目を時期 B とする．

UBM–GMM 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 モデル学習の実験条件を表 2 に示す．位相特徴は 2 章で示した Relative phase information(E), 整数化を

Table 2　位相特徴抽出および GMM モデル学習の実験条件

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
正規化パラメータ	$A = 0, \omega_b = 1000$

Table 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）

	フレーム長	フレームシフト
frameleng0	12.5	5
frameleng1	50	25
frameleng2	75	37.5
frameleng3	100	50
frameleng4	500	100

用いた Relative phase information(RE), 整数化および簡素化を用いた Relative phase information(RSE) 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し，GMM の学習を行った．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 3 種類の発話長を使用した．データベース本来の発話長 (約 4 秒) を original として，発話区間の秒数がおよそ 1 秒とした short，3 発話を連結した long を作成した．MFCC の抽出では前処理として音声区間検出を行ったが，位相特徴抽出では行っていない．学習では時期 A のデータを使用し，MFCC および位相それぞれの特徴を抽出して特定話者モデルを学習し，テスト時には時期 A，B それぞれを用いて MFCC および位相の抽出を行った．また，スコアの正規化に用いたパラメータ α は RE，RSE でそれぞれ 0.25，0.1 である．

5　実験結果

5.1　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話者照合実験を行った．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 から得られた照合スコアから統合スコアを算出し，各フレーム長で最も低かった EER を表 4 に示す．表中の “MFCC のみ” の行は UBM–GMM のみでの EER を示している．また，表中の括弧内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．全ての位相抽出手法で精度が等しい場合には表記を省略してある．

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 4 より，すべてのフレーム長および発話長で 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．このことから，位相情報が話者性を表現する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期待していた．しかし，表 4 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（frameleng4）に，整数化や簡素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，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 4 より，学習データとテストデータの時期が同じ場合には frameleng0 または frameleng1 が最小の EER となったが，学習データとテストデータの時期が異なる場合には発話長によって最小の EER となるフレーム長が異なった．発話長 long では frameleng0，frameleng1，frameleng2 が最小の EER となった．これは，発話長が十分に長いため発話時期の変動に左右されにくく，短いフレーム長であっても安定して位相抽出が可能なためであると考えられる．一方で，発話長 original および short では frameleng3 で最小の 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できているためだと考えられる．

最後に，位相特徴量の種類とフレーム長に関して比較していく．学習データとテストデータの時期が同じ場合には，発話長 long の場合には全てのフレーム長で全位相抽出手法で等しい EER となっている．発話長 original の場合には従来手法である E が全フレーム長の中で最小の EER と

Table 4　各フレーム長において最小の EER(%)

			テストデータ（時期 A）			テストデータ（時期 B）		
			long	original	short	long	original	short
学習データ（時期 A）	MFCC のみ		0.00	0.26	0.59	0.04	1.37	3.92
	MFCC + phase	frameleng0	0.00	**0.20** (E)	**0.45** (RE)	**0.00** (RSE)	1.27 (E)	3.92
		frameleng1	0.00	**0.20** (E)	0.59	**0.00** (E,RSE)	1.31 (RE)	3.73 (E)
		frameleng2	0.00	0.21 (E)	0.59	**0.00** (E)	1.31 (RE)	3.92
		frameleng3	0.00	0.23 (E,RE)	0.59	0.04 (RE,RSE)	**1.24** (RE)	**3.66** (E)
		frameleng4	0.00	0.23 (E,RSE)	0.59	0.04 (RE,RSE)	1.31 (RE)	3.92

なっているが，フレーム長が長い場合では検討した位相特徴 (RE と RSE) も同程度の EER となっている．発話長 short の場合には frameleng0 のみ RE が最小の EER となっている．一方で，学習データとテストデータの時期が異なるときの結果を見ると，発話長 long ではフレーム長が短い場合に E および RSE が，フレーム長が長い場合に RE または RSE が最小の EER となった．発話長 original では frameleng0 を除く全フレーム長で RE が最小の EER となった．発話長 short では E が最小の EER となっているが，frameleng4 での結果のように RE または RSE を統合することで精度を下げてはいないことがわかる．これらのことから，RE と RSE は E よりも特徴点が少ないにもかかわらず従来法と同等に位相の特徴を表せており，また，提案法である RE は特にフレーム長が長い場合，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．

5.2　MFCC と位相情報のフレーム長

表 3 に示した位相特徴抽出に使用したフレーム長を用いて MFCC の抽出を行い，UBM–GMM の学習を行った．各フレーム長の UBM–GMM に対応するフレーム長のテストデータ (発話長 original) を入力して照合スコアを計算し，各位相特徴抽出に基づく GMM から得られた照合スコアから統合スコアを算出した．実験結果より，学習データとテストデータの時期が異なる場合には MFCC のフレーム長が短く（frameleng0）かつ，位相特徴のフレーム長を長めに設定し（frameleng2），整数化を行った RE が 0.60% と高い精度となった．

6　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話者照合実験を行った．実験結果では位相の整数化や簡素化による性能改善が得られた．今後の課題としては，位相抽出手法の改善や位相特徴抽出における音声区間検出の適用，位相のモデル学習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] Paliwal *et al.*, “Usefulness of Phase Spectrum in Human Speech Perception,” Proc. Eurospeech, 2003, pp. 2117–2120, (2003).

[2] Wang *et al.*, “Relative phase information for detecting human speech and spoofed speech,” Proc. Interspeech, 2015, pp. 2092–2096, (2015).

[3] 山本 *et al.*, “長時間分析に基づく位相情報を用いた音声認識の検討,” 電子情報通信学会技術研究報告（SP）, vol. 110, No. 143, 2010, pp. 31–36, (2010).

[4] Correia *et al.*,“Preventing converted speech spoofing attacks in speaker verification,” Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on, Proc. IEEE, 2014, pp. 1320–1325, (2014).

[5] Shiota *et al.*, “Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification,” Proc. Interspeech, 2015, pp. 239–243, (2015).

[6] 仲野 *et al.*, “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,” 研究報告音声言語情報処理（SLP）, vol. 2016–SLP–114, No. 16, 2016, pp.65–70, (2016).