社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－SLP－74（30）
2008 / 12 / 10

位相情報を利用した話者識別・照合法の評価

王　龍標† 南　和江†† 山本　一公†† 中川　聖一††

† 静岡大学 システム工学科, 〒432-8561 静岡県浜松市中区城北3－5－1
†† 豊橋技術科学大学 情報工学系, 〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘1-1
E-mail: †xiaowang@sys.eng.shizuoka.ac.jp, ††{minami,kyama,nakagawa}@slp.ics.tut.ac.jp

あらまし 従来のMFCCに基づく話者認識手法は、位相情報が有効に利用されていない。本稿では、話者識別・照合のための切り出し位置による位相のずれを一致させる位相の利用法とその改善法を提案し、従来のMFCCと組み合わせる。男性22名、女性13名が5時期に、普通、遅い、早いスピードで発声したNTTデータベースを使って提案法を評価した。学習には最初の1時期の普通速度で発声された5文（約20秒）を用いた。クリーン音声、雑音重畳音声による評価実験を行い、提案する位相情報は単独ではMFCCに劣るが、話者認識の能力をもつことが分かった。MFCCと組み合わせることで、より高い話者識別・照合率を得ることができた。改善した位相情報が従来の位相情報よりも有効であることが分かった。

キーワード 話者識別、話者照合、位相情報、MFCC、組み合わせ手法

Evaluation of speaker identification/verification method using phase information

Longbiao WANG†, Kazue MINAMI††, Kazumasa YAMAMOTO††, and Seiichi NAKAGAWA††

† Department of System Engineering, Shizuoka University
3-5-1, Johoku, Naka-ku, Hamamatsu, Shizuoka, 432-8561, Japan
†† Department of Information and computer Sciences, Toyohashi University of Technology
1-1, Hibarigaoka, Tempaku-cho, Toyohashi, Aichi, 441-8580, Japan
E-mail: †xiaowang@sys.eng.shizuoka.ac.jp, ††{minami,kyama,nakagawa}@slp.ics.tut.ac.jp

Abstract In conventional speaker recognition methods based on MFCC, phase information has been ignored. In this paper, we propose original phase information and a modified feature parameter obtained from the original phase information, and integrate the phase information with MFCC on a speaker identification/verification method. The speaker identification/verification experiments were performed using NTT database which consists of sentences uttered by 35 Japanese speakers (22 males and 13 females) with normal, fast and slow speaking modes on five sessions. Each speaker uttered only 5 utterances at a normal speaking mode (about 20 seconds in total) for training data. Although the phase information based method worked worse than MFCC based method, it had some ability of speaker identification. By integrating the phase information with the MFCC, the speaker identification error rate was remarkably reduced for all speaking rates in comparison with a standard MFCC-based method. The proposed modified phase information was more robust than the proposed original phase information for all speaking modes.

Key words speaker identification, speaker verification, MFCC, phase information, GMM

1. はじめに

GMMを用いた話者認識に関する研究としては，Reynoldsらは，GMMを用いたテキスト独立型話者識別法を提案している[1]．識別法に関する研究としては，学習と認識時の音響環境の変動に対して，背景話者モデルを用いた尤度正規化による頑健な話者認識法[2]，話者モデルの分散の補正による頑健な話者認識法[3]などがある。特徴量に関する研究としては、話者性が多く存在する周波数帯域をRASTAにより強調した話者照合法[4]，ピッチやLPC予測残差を併用する方法が提案されている[5]。モデル化に関する研究としては，音声中の動的な話者性を考慮する方法として，フレーム間の連鎖確率を導入したGMMによる話者識別法[6]，スペクトルとピッチをモデル化する方法として，多空間上の確率分布に基づくGMMによる話者認識法が提案されている[7]。また、GMMを用いた話者照合における尤度の正規化法の比較[8]などが試みられている。GMMと話者適応化モデルの併用による話者認識の研究も行われている[9][10]。

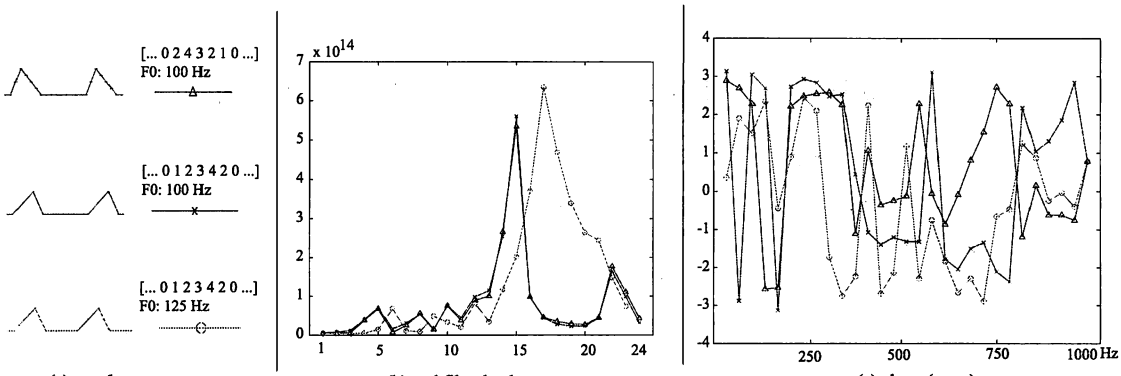

図 1　合成音声の音源波形、メルフィルタバンク出力と位相

以上のように GMM による手法はテキスト独立型話者認識において一般的な手法として使われている。この手法では識別対象の話者ごとに GMM で音声をあらかじめモデル化する。ただこの GMM で用いる特徴パラメータは従来は MFCC を用いている。しかし、MFCC は時間領域の音声フレームのフーリエ変換の振幅だけを利用するので、位相情報が無視されている。MFCC は声道の情報が捉えるが、音源の情報は直接捉えない。音源情報は話者認識に有効であることが報告されてきた [5], [11]〜[14]。これらのほとんどは LPC 分析に基づいており限界がある。位相情報は音源情報を有する可能性があるので、話者認識に有用すると期待される。いくつかの研究は位相の情報を直接にモデル化され、音声処理へ統合されている [15], [16]。音声認識での位相の重要性が文献 [17], [18] に報告されている。

我々は、切り出し位置による異なる位相のずれを一致させる位相の利用法を提案した [19]。しかし、従来手法である位相情報 $\tilde{\theta}$ は位相差が小さいにも関わらず大きな差として比較されてしまう欠点があった。そこで、$\tilde{\theta}$ に対応する座標値に変換した改善した位相情報 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ を用いた [20], [21]。本稿では、話者識別・話者照合のための位相情報の有効性を評価する。位相単独 GMM での実験、MFCC と位相情報の GMM の組み合わせ実験、2 つの組み合わせた GMM とさらに MFCC に基づく HMM（不特定話者の 116 音節 HMM を話者適応化したモデル）の 3 つを組み合わせた実験を行い、従来の位相情報と改善した位相情報を比較する。なお、GMM と HMM の組み合わせによる話者認識法 [22] は我々の研究室で開発されたが、最近 Jin らもその有効性を示している [23]。最後に、クリーン音声に雑音を加えて雑音環境をシミュレートして、雑音環境下における位相情報の有効性を検証する。

2. 位相情報

従来の話者認識では、MFCC(Mel-frequency cepstral coefficients) を主として用いており、音声に含まれている位相情報は一切無視されている。図 1 では、合成した音声波形の正規化された位相とパワースペクトルを示している。図 1 から位相は、音源波形の特徴によって非常に影響を受け、声道の形によっても影響されることがわかる。これに基づき、我々は位相情報の抽出法とその改善法を考案し、それを用いて GMM を作成することで、テキスト独立の話者認識・話者照合を行う方法を提案する。

2.1 異なる位相のずれを一致させる方法

16kHz でサンプリングされた音声波形において 256 個のサンプル点を切り出す。これを離散フーリエ変換することで以下の 128 個の線スペクトルを得る ($\omega = 2\pi f$)。

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

ここで、同じ周波数 ω でも切り出す位置によって位相が異なってくるという問題が生じる。この問題への対処として、ある基準とする周波数 ω 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という手法をとった。具体例として、基準周波数 ω の位相を $\pi/4$ となるようにした場合を挙げる。すなわち、

$$\sqrt{X^2(\omega,t)+Y^2(\omega,t)} \times e^{j\theta(\omega,t)} \times e^{j\theta(\frac{\pi}{4}-\theta(\omega,t))} \qquad (2)$$

とする。このとき、他の周波数 $\omega' = 2\pi f'$ では

$$\begin{aligned}\sqrt{X^2(\omega',t)+Y^2(\omega',t)} \times e^{j\theta(\omega',t)} \times e^{j\frac{\omega'}{\omega}(\frac{\pi}{4}-\theta(\omega,t))} &= \sqrt{X^2(\omega',t)+Y^2(\omega',t)} \times e^{j\tilde{\theta}(\omega',t)} \\ &= \tilde{X}(\omega',t)+j\tilde{Y}(\omega',t) \qquad (3)\end{aligned}$$

となり、ω が $\pi/4$ の場合の位相時における ω' がもつ位相を正規化することができる。このとき、(3) 式の実数成分と虚数成分はつぎのようになる。

$$\begin{aligned}\tilde{X}(\omega',t) = \sqrt{X^2(\omega',t)+Y^2(\omega',t)} \times \cos\{\theta(\omega',t) \\ +\frac{\omega'}{\omega}(\frac{\pi}{4}-\theta(\omega,t))\} \qquad (4)\end{aligned}$$

$$\begin{aligned}\tilde{Y}(\omega',t) = \sqrt{X^2(\omega',t)+Y^2(\omega',t)} \times \sin\{\theta(\omega',t) \\ +\frac{\omega'}{\omega}(\frac{\pi}{4}-\theta(\omega,t))\} \qquad (5)\end{aligned}$$

なお、今回の実験では基準周波数 $\omega = 2\pi \times 1000$Hz とした。

2.2 位相情報の利用法の改善

従来の位相パラメータ $\tilde{\theta}(\omega', t)$ の差は、$-\pi$〜π の範囲を超える場合がありえ、また $\pi - \theta'$ と $\theta'' = -\pi + \theta'$ は θ' が小さい場合は位相差が小さいにも関わらず $|\pi - \theta' - (-\pi + \theta')|$=$2\pi - 2\theta'$ と、大きな差として比較されてしまう欠点があった。そこで θ' のかわりに $\cos\theta'$ と $\sin\theta'$ に変換し、θ' に対応する座標値に変換した位相情報を用いる。この場合、上述の位相差は

$|\sin(\pi - \theta') - \sin(-\pi + \theta')|$=$2\sin(\theta') \approx 2\theta'$

$|\cos(\pi - \theta') - \cos(-\pi + \theta')|$=0　と小さくなる。

なお、θ のような周期関数の分布はフォン・ミーゼ分布で表現する方法もあるが、GMM でモデル化する方が頑健であると考えられる。

以下に今回比較した特徴量をまとめる。

(1) 従来の MFCC

(2) フーリエ変換後の正規化した対数パワースペクトル

(3) $\tilde{\theta}(\omega', t)$=$\theta(\omega', t) + \frac{\omega'}{\omega}(\frac{\pi}{4} - \theta(\omega, t))$

(4) $cos\tilde{\theta}(\omega', t)$, $sin\tilde{\theta}(\omega', t)$

表 1　単独特徴パラメータによる話者識別結果 (%)

速度	特徴量	認識率 (%)	
		32 混合	64 混合
普通の発声速度	MFCC	98.0	97.7
	$\tilde{\theta}$ (60-700Hz)	59.6	52.6
	$\tilde{\theta}$ (300-1000Hz)	51.1	61.0
	$\tilde{\theta}$ (600-1300Hz)	38.1	36.1
	$\cos\tilde{\theta}$, $\sin\tilde{\theta}$(60-700Hz)	72.3	73.4
	$\cos\tilde{\theta}$, $\sin\tilde{\theta}$(300-1000Hz)	59.7	60.4
	$\cos\tilde{\theta}$, $\sin\tilde{\theta}$(600-1300Hz)	32.7	37.4
	spectrum(60-700Hz)	76.4	78.0
	spectrum(300-1000Hz)	67.1	67.4
	spectrum(600-1300Hz)	64.3	67.9
	spectrum(1060-1760Hz)	59.9	57.1
	spectrum(1480-2180Hz)	47.1	50.0
遅い発声速度	MFCC	96.4	94.9
	$\tilde{\theta}$ (60-700Hz)	56.0	51.7
	$\tilde{\theta}$ (300-1000Hz)	47.4	56.6
	$\tilde{\theta}$ (600-1300Hz)	34.4	34.7
	$\cos\tilde{\theta}$, $\sin\tilde{\theta}$(60-700Hz)	71.3	70.4
	$\cos\tilde{\theta}$, $\sin\tilde{\theta}$(300-1000Hz)	58.4	57.6
	$\cos\tilde{\theta}$, $\sin\tilde{\theta}$(600-1300Hz)	33.7	39.0
	spectrum(60-700Hz)	73.9	75.1
	spectrum(300-1000Hz)	65.9	67.0
	spectrum(600-1300Hz)	61.4	65.1
	spectrum(1060-1760Hz)	56.0	55.9
	spectrum(1480-2180Hz)	45.3	47.0
速い発声速度	MFCC	95.1	95.1
	$\tilde{\theta}$ (60-700Hz)	59.1	51.6
	$\tilde{\theta}$ (300-1000Hz)	48.6	57.6
	$\tilde{\theta}$ (600-1300Hz)	30.9	31.7
	$\cos\tilde{\theta}$, $\sin\tilde{\theta}$(60-700Hz)	67.7	72.0
	$\cos\tilde{\theta}$, $\sin\tilde{\theta}$(300-1000Hz)	56.1	55.6
	$\cos\tilde{\theta}$, $\sin\tilde{\theta}$(600-1300Hz)	28.1	37.1
	spectrum(60-700Hz)	74.9	73.7
	spectrum(300-1000Hz)	68.4	68.3
	spectrum(600-1300Hz)	60.3	65.1
	spectrum(1060-1760Hz)	52.4	51.0
	spectrum(1480-2180Hz)	41.6	47.0

3. 組合せ手法と話者識別・照合の決定方法

本稿では、MFCC に基づく GMM ／ HMM を位相に基づく GMM と組み合わせる。二つの方法を使って話者識別・照合をする時に、*MODEL 1* の尤度と *MODEL 2* の尤度を線形的に結合し、新しいスコア L_2^n は以下の式になる。

$$L_2^n = (1-\alpha)L_{MODEL1}^n + \alpha L_{MODEL2}^n \qquad (6)$$

三つのモデルを組み合わせる時に, 新しいスコア L_3^n は以下の式になる。

$$L_3^n = (1-\beta)L_{MODEL1}^n + \beta\{(1-\alpha)L_{MODEL2}^n + \alpha L_{MODEL3}^n\} \qquad (7)$$

ここで、L_{MODEL}^n は *MODEL* の n 番目の話者の尤度で、$n = 1$、2、$\cdots$、N、N は登録話者の数である。α と β はそれぞれお重み付き係数である。

話者識別では、最大尤度をもつ話者が目標話者として決定される。話者照合 [24]〜[27] では、入力音声に対する本人の音響モデルの照合スコアと事前に設定する閾値との大小関係により、受理／棄却の判断を行う。尤度の影響を低減するために、これまでにいくつかの正規化手法が研究されてきた。Rosenberg らは、背景話者として、各本人話者に対して設定するコホート話者を用いて尤度の正規化方法を提案している [26]。コホート話者モデル方法は、簡単で良い性能が得られるので、本研究でも用いた。本稿では、コホートの話者の数は 3 とした。

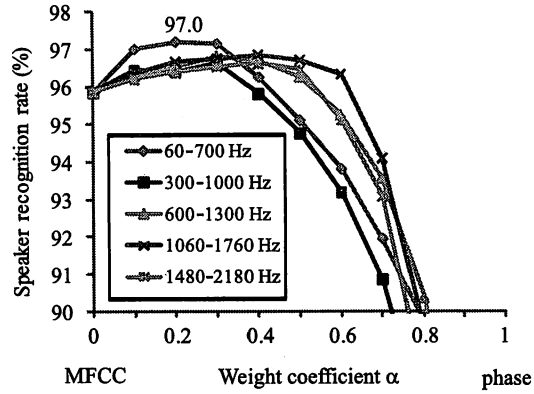

図 2　MFCC と対数パワースペクトルの組み合わせ実験の話者認識率 (混合数 64、発声速度=3 話速の平均)

4. 認識実験

4.1 実験データ

本実験で使用したデータは、男性 22 名、女性 13 名が約 10ヶ月にわたる 5 時期に、普通、遅い、早いスピードで発声した NTT データベースである [22]。

特徴パラメータは、標本化周波数 16kHz、フレーム長 25ms、フレーム周期 10ms の分析条件で、12 次元の MFCC に基づく GMM と 25 次元の MFCC に基づく HMM のほか、2 節で述べた方法 (フレーム長 12.5ms、フレーム周期 5ms の位相) を用いて抽出する。学習には最初の 1 時期の普通の速度で発声された 5 文を用い、テストでは、その他の 4 時期の各 5 文の 1 文ずつを用いた。平均発話時間は約 4 秒である。テストで用いた文は、学習に用いた文とは異なるが、全話者、全時期で同じである。実験に用いたテスト音声の総数は 35(名) × 4(時期) × 5(文) × 3(速度)= 2100(文) である。

4.2 クリーン音声による話者認識実験

4.2.1 話者識別実験

i) 単独特徴パラメータでの実験

特徴パラメータ、各話者ごとに GMM(混合数 32,64) を作成し、認識実験を行った結果を表 1 に示す。従来の位相情報は括弧内に示す帯域に対応する低域の 12 個の周波数における位相 $\{\tilde{\theta}\}$ の 12 次元、今回の位相情報は $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ の 24 次元である。また spectrum は対数パワースペクトルの低周波数帯域での単独認識結果を示す。位相 $\{\tilde{\theta}\}$、提案手法 $\{\cos\tilde{\theta}$、$\sin\tilde{\theta}\}$、対数パワースペクトルのいずれも、低い周波数帯域で有効であった。それぞれの混合数において MFCC の認識率が最もよかったが、位相情報にも話者識別の能力をもつことが分かった。また、今回提案した位相情報 $\{\cos\tilde{\theta}$、$\sin\tilde{\theta}\}$ が従来の位相情報 $\{\tilde{\theta}\}$ よりも有効であることがわかった。

ii) 2つの特徴パラメータの組み合わせ実験

位相単独での実験の結果を受けて、MFCC と位相情報や対数パワースペクトルの GMM を組み合わせて実験を行った。MFCC と対数パワースペクトルとの組み合わせによる実験を行った結果を図 2 に示す。この 2 つの組み合わせでは、普通の発声速度、速い速度、遅い速度の 3 話速の平均の結果である。周波数帯域 (60-700Hz) での対数パワースペクトルを用いた場合に最も良い重みで 97.0% の認識率を得た。また MFCC と位相との組み合わせによる実験を行った結果を図 3、4 に示

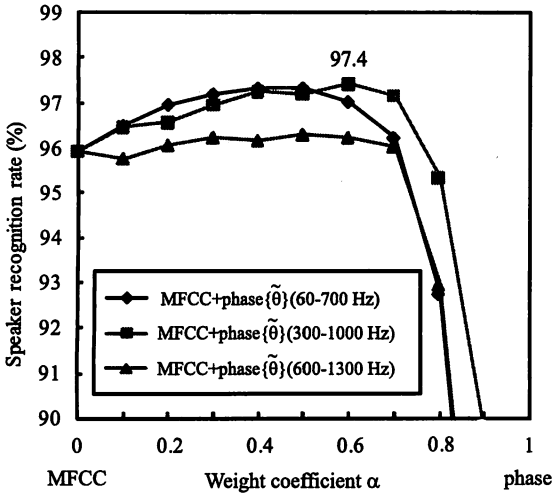

図 3 MFCC と位相 $\{\tilde{\theta}\}$ の組み合わせ実験の話者認識率 (混合数 64、発声速度=3 話速の平均)

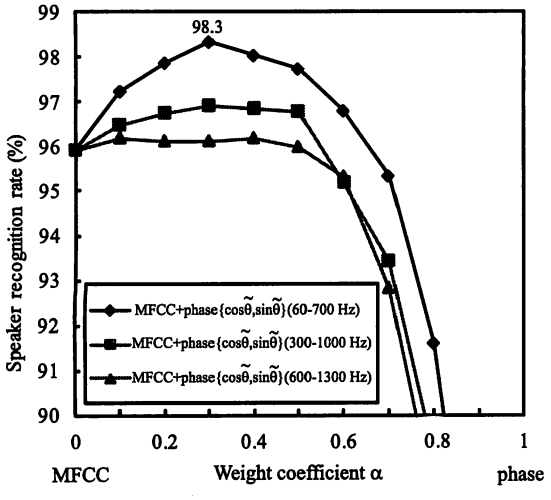

図 4 MFCC と位相 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ の組み合わせ実験の話者認識率 (混合数 64、発声速度=3 話速の平均)

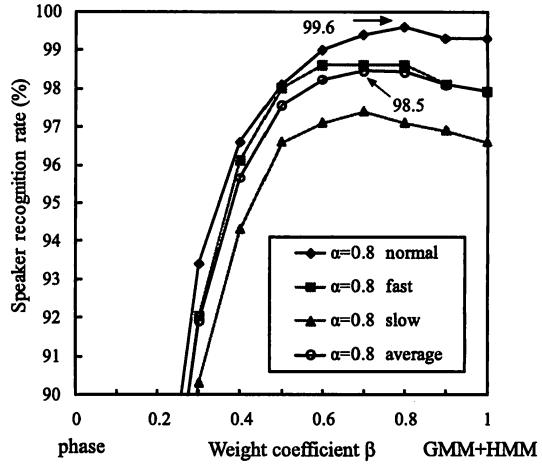

図 5 MFCC に基づく GMM、MFCC に基づく HMM と位相 $\{\tilde{\theta}\}$ の GMM の組み合わせ認識率 (混合数 64、周波数帯域 (60-700Hz))

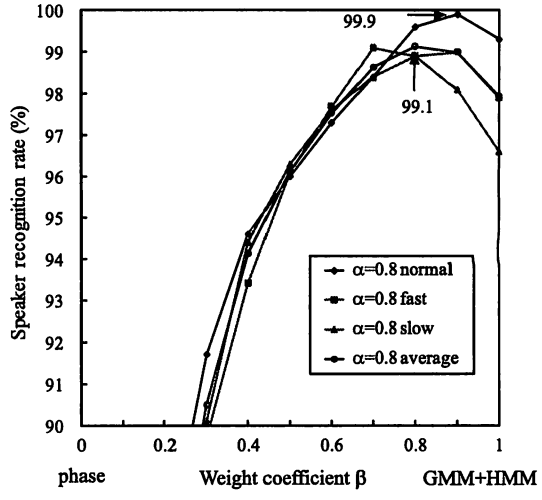

図 6 MFCC に基づく GMM、MFCC に基づく HMM と位相 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ の GMM の組み合わせ認識率 (混合数 64、周波数帯域 (60-700Hz))

す。ここでは、128 個の位相要素のうち、低い周波数帯域 (60-700Hz) での 12 個の位相を用いた場合に最も良い重みで MFCC 単独の誤認識率を約 50% ～60% 減少させることができた。対数パワースペクトルとの組み合わせに比べ、誤認識率を減少させることができた。従来の位相情報 $\{\tilde{\theta}\}$ を用いることにより、普通の発声速度で 99.0% [20]、速い速度で 97.0% で、遅い速度で 96.6% の認識率を得た (3 話速の平均で、97.4%)。改善した位相情報 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ を利用することにより、それぞれ 99.3% 98.0% 98.1% の認識率を得た (3 話速の平均で、98.5%)。しかし、位相情報の単独利用における性能の向上と比べて、その向上は少なかった。なおこれに、パワースペクトルの特徴を加えた 3 つの特徴パラメータを用いて認識しても、認識率は向上しなかった。我々が以前に提案した MFCC に基づく GMM と MFCC に基づく HMM の 2 つの組み合わせ実験で重み付け係数が 0.8 のときに最も良い結果が得られた (3 話速の平均で 97.9%) [22]。この併用効果は大きく、最近他の研究でも用いられている [23]。

iii) 3つの手法の組み合わせ実験

これらの結果から、MFCC に基づく GMM、MFCC に基づく HMM、位相の 3 つの尤度を組み合わせた認識実験を行った。結果を図 5、6 に示す。

MFCC に基づく GMM と MFCC に基づく HMM の 2 つの組み合わせ認識率の最も良い結果が重み付け係数 0.8 のときに普通速度で 99.3%、速い速度で 97.9%、遅い速度で 96.6%(平均で 97.9%) [22] であったが、位相情報を組み合わせることで更なる改善が見られた。位相 $\{\tilde{\theta}\}$ を組み合わせたときに、それぞれ 99.6%、98.6%、97.1%(平均で 98.5%) まで向上した。位相 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ を組み合わせたときには、99.6%、98.9%、98.9%(平均で 99.1%) に向上した。普通速度の時で最適な重みで 99.9%が得られた。絶対値の向上はわずかではあるが、誤認識率は半減した。MFCC に基づく GMM、MFCC に基づく HMM と対数パワースペクトルの 3 つの手法で組み合わせも行った。認識結果はそれぞれ 99.4%、99.0%、98.4%(平均で 98.9%) であった。位相 $\{\tilde{\theta}\}$ との組み合わせ結果より向上したが、位相 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ には及ばなかった。この結果から、話者認識において、位相情報は有効な手段であることがわかり、さらに今回行った位相情報の改善によってより頑健な結果が得られた。

表 2　話者照合の等誤り率 (64 混合：%)

speed	normal	fast	slow	Ave.
MFCC_GMM	0.58	1.28	1.38	1.08
MFCC_HMM	0.70	1.00	1.43	1.05
$\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ (60–700 Hz)	5.22	6.15	5.69	5.69
spectrum (60–700 Hz)	6.67	7.94	7.21	7.27
MFCC_GMM+MFCC_HMM	0.35	0.91	1.16	0.81
MFCC_GMM+$\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ (60–700 Hz)	0.38	0.75	0.77	0.63
MFCC_GMM+spectrum (60–700 Hz)	0.42	0.88	1.06	0.79
MFCC_HMM+$\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ (60–700 Hz)	0.28	0.40	0.83	0.50
MFCC_HMM+spectrum (60–700 Hz)	0.39	0.52	0.83	0.58
MFCC_GMM+MFCC_HMM + $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ (60–700 Hz)	0.18	0.37	0.71	0.42
MFCC_GMM+MFCC_HMM + spectrum (60–700 Hz)	0.23	0.49	0.78	0.50

4.2.2　話者照合の実験

話者識別における位相情報の利用の有効性を 4.2.1 節で示した。しかし、本研究に利用されている登録話者の数があまに多くない。そこで、位相情報の頑健性を検証するために、改善した位相情報 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ (周波数帯域 (60-700Hz)) を用いて話者照合の実験も行った(注1)。話者照合でも、話者識別に利用された GMM と HMM に基づく話者モデルを利用する。ここで、音声分析の条件は話者識別の分析条件と同じである。

話者照合の等誤り率 (Equal Error Rate：EER) を表 2 に示す。話者照合の結果と話者識別の結果の傾向は似ている。対数パワースペクトル (周波数帯域 (60-700Hz)) に基づく実験も位相情報を比較するために行った。改善した位相情報 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ は MFCC より劣るが、対数スペクトルより良いし、相対的に高い性能 (約 5%の ERR) が得られた。話者識別と比べると、MFCC に基づく GMM と MFCC に基づく HMM の相補性が比較的小さい。MFCC(声道情報) と位相情報 (音源情報) の相補性が高いので、MFCC と位相情報を組み合わせて話者照合の性能が大きく改善された。MFCC に基づく GMM と位相情報との併用は、MFCC に基づく GMM より 41.7%の相対エラー削減率、MFCC に基づく HMM と位相情報との併用は、MFCC に基づく HMM より 52.4%の相対エラー削減率が達成できた。三つの手法を組み合わせると、MFCC に基づく GMM より 61.1%、MFCC に基づく HMM より 60.0%、MFCC に基づく GMM と MFCC に基づく HMM の組合せより 49.4%の相対エラー率の削減ができた。また、MFCC と位相情報の組合せは、MFCC と対数パワースペクトルの組合せより良い照合性能を達成した。特に、普通速度での位相情報の併用効果が大きい。

異なる特徴パラメータの DET(Detection Error Trade-off) 曲線を図 7 で比較する。位相情報を利用すると、誤受理率と誤棄却率のトレードオフは、MFCC のみの方法より格段に優れていることが分かる。

4.3　雑音重畳音声による話者認識実験

クリーン音声に雑音を加えて雑音環境をシミュレートする実験を行った。クリーン音声は話者識別で使った NTT データベースの普通速度で発話した文を用いた。雑音としては、電子協雑音データベースの展示会場のブース内のものと計算機室のワークステーション内のものを用いた [29]。12 次元の MFCC と 12 次元の位相情報 $\{\tilde{\theta}\}$ ／ 24 次元の改善した位相情報 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ (周波数帯域：60-700Hz) を利用し、話者モデルは 32 混合の GMM を用いた。

雑音重畳音声による話者識別と話者照合の結果を表 3 と表 4 に示す。雑音環境下で、位相情報による話者認識率と MFCC による認識率の差はクリーン音声より小さい。特に、話者識別・話者照合に対して、SNR が低い (0 dB) 場合、改善した位相情報 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ 単独の結果が MFCC 単独の結果よりも良かった。結果により、位相が雑音に対して頑健であることがわかる。話者識別と話者照合に対して、全ての SNR 条件において、MFCC と位相情報の組合せ結果は、MFCC 単独の結果より大きな改善ができた。また、改善した位相情報 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ が従来の位相情報 $\{\tilde{\theta}\}$ より有効であることが分かった。

(注1)：話者識別に対しては、話者数が識別の性能に大きく影響する。話者数が無限大の時に、話者識別のエラーの確率が 1 に近づく。しかし、話者照合の性能は話者数に影響されない [28]。

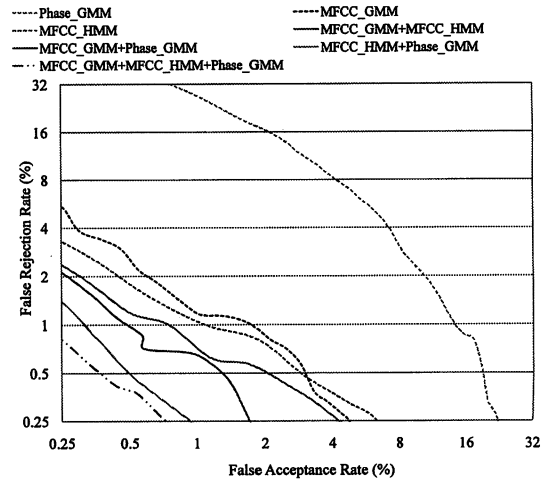

図 7　異なる特徴パラメータによる話者照合の DET 曲線 (64 混合)

5. ま と め

本稿では、切り出し位置によって異なる位相のずれを一致させる位相の利用法とその改善法を提案し、従来の MFCC と併用した。従来の位相情報 $\pi - \tilde{\theta}$ は $-\pi + \tilde{\theta}$ の比較に際して問題があり、これを座標値に変換することにより、認識率を向上させることができた。また、MFCC と改善した位相情報を組み合わせた結果においても従来の位相情報の結果を上回ることができた。この結果を受けて、MFCC に基づく GMM と MFCC に基づく HMM との 2 つの組み合わせに、さらに今回提案した位相情報 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ を組み合わせることで更なる向上が見られた。改善した位相情報 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ に基づく GMM と MFCC に基づく GMM を組み合わせることで、MFCC に基づく GMM より 3 話速の平均で 58.5%の誤認識率が減少した。2 つの組み合わせた GMM とさらに MFCC に基づく HMM の 3 つを組み合わせたときには、MFCC に基づく GMM と MFCC に基づく HMM との 2 つの組み合わせ [22] より 3 話速の平均で 57.1%の誤認識率の減少が見られた。特に、発声速度毎の最適な重みの組み合わせではそれぞれ 99.9%,99.0%,98.9%(平均で 99.3%) の認識率が得られた。また、本研究の結果は、同じデータベースを用いて他の研究の結果 [5], [22], [30], [31] より格段に良かった。

本稿では、改善した位相情報 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$ を用いて話者照合の実験も行った。MFCC に基づく GMM と位相情報との組み合わせは、MFCC に基づく GMM より 41.7%の相対エラー削減率、MFCC に基づく HMM と位相情報との組み合わせは、MFCC に基づく HMM より 52.4%の相対エラー削減率が達成できた。三つの手法を組み合わせると、MFCC に基づく GMM より 61.1%、MFCC に基づく HMM より 60.0%、MFCC に基づく GMM と MFCC に基づく HMM の組合せより 49.4%の相対エラー率の削減ができた。

最後に、位相情報を用いて、シミュレーション雑音環境下での話者識別・照合を行った。全ての SNR 条件において、MFCC と位相情報の組合せ結果は、MFCC 単独の結果より大きい改善ができた。

文　献

[1] D. A. Reynolds and R. C. Rose, "Robust Text-Independent

表 3　雑音環境下での話者識別率 (32 混合：%)

(a) 展示会場雑音

SNR	0 dB	10 dB	20 dB	Ave.
MFCC	8.9	42.4	76.3	42.5
位相 $\{\tilde{\theta}\}$	4.4	33.0	53.9	30.4
位相 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$	9.9	42.3	63.3	38.5
MFCC+位相 $\{\tilde{\theta}\}$	13.7	55.9	86.8	51.4
MFCC+位相 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$	11.3	66.6	91.3	56.4

(b) 計算機室雑音

SNR	0 dB	10 dB	20 dB	Ave.
MFCC	6.6	36.0	75.6	39.4
位相 $\{\tilde{\theta}\}$	10.7	34.4	55.3	33.5
位相 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$	11.3	39.1	65.7	38.7
MFCC+位相 $\{\tilde{\theta}\}$	9.1	52.3	85.4	49.0
MFCC+位相 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$	8.6	59.3	91.9	53.2

表 4　雑音環境下での話者照合の等誤り率 (32 混合：%)

(a) 展示会場雑音

SNR	0 dB	10 dB	20 dB	Ave.
MFCC	36.1	11.8	4.8	17.6
位相 $\{\tilde{\theta}\}$	37.6	16.3	11.3	21.7
位相 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$	29.9	12.1	6.4	16.1
MFCC+位相 $\{\tilde{\theta}\}$	32.3	8.5	3.6	14.8
MFCC+位相 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$	26.7	7.2	2.0	12.0

(b) 計算機室雑音

SNR	0 dB	10 dB	20 dB	Ave.
MFCC	40.5	13.9	4.6	19.7
位相 $\{\tilde{\theta}\}$	34.4	16.7	11.1	20.7
位相 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$	32.6	12.8	7.1	17.5
MFCC+位相 $\{\tilde{\theta}\}$	31.7	8.9	3.3	14.6
MFCC+位相 $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$	29.4	7.7	1.9	13.0

Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Trans.SAP, Vol. 3, No. 1, pp. 72–83 (1995).

[2] L. P. Heck and M. Weintraub, "Handset-Dependent Background Models for Robust Text-Independent Speaker Recognition", Proc. ICASSP, pp. 1071–1074 (1997).

[3] F. Beaufays and M. Weintraub, "Model Transformation for Robust Speaker Recognition from Telephone Data", Proc. ICASSP, pp. 1063–1066 (1997).

[4] S. V. Vuuren and H. Hermansky, "On the Importance of Components of the Modulation Spectrum for Speaker Verification", Proc. ICSLP, Vol.7, pp. 3205–3208 (1998).

[5] K.P. Markov and S. Nakagawa, "Integrating pitch and LPC-residual information with LPC-cepstrum for text-independent speaker recognition", Jour. ASJ (E), Vol. 20, No. 4, pp. 281–291 (1999).

[6] W. H. Tsai, C. Che and W. W. Chang, "Text-Independent Speaker Identification Using Gaussian Mixture Bigram Models", Proc. ICSLP, Vol. 2, pp. 314–317 (2000).

[7] C. Miyajima, Y. Hattori and K. Tokuda, "Text-Independent Speaker Identification Using Gaussian Mixture Models Based on Multi-Space Probability Distribution", IEICE Trans. Vol. E84-D, No. 7, pp. 847–855, (2001).

[8] D. Tran and M. Wagner, "A Proposed Likelihood Transformation for Speaker Verification", Proc. ICASSP, Vol. 2, pp. 1069–1072 (2000).

[9] D. E. Sturim et. al., "Speaker Verification using Text-Constrained Gaussian", Proc. ICASSP, Vol. 1, pp. 677–680 (2002).

[10] Alex Park and Timothy J. Hazen, "ASR dependent techniques for speaker identification", Proc. ICSLP, pp. 1337–1340 (2002).

[11] M.D. Plumpe, T.F. Quatieri and D.A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification", IEEE Trans. Speech and Audio Processing, Vol. 7, No. 5, pp. 569–586 (1999).

[12] K.S.R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker verification", IEEE Signal Processing Letters, Vol. 13, No. 1, pp. 52–55 (2006).

[13] N. Zheng, T. Lee and P.C. Ching, "Integration of complementary acoustic features for speaker recognition", IEEE Signal Processing Letters, Vol. 14, No. 3, pp. 181–184 (2007).

[14] J. Gudnason and M. Brookes, "Voice source cepstrum coefficients for speaker identification", Proc. ICASSP, pp. 4821–4824 (2008).

[15] R. Schluter and H. Ney, "Using phase spectrum information for improved speech recognition performance", in proceedings of ICASSP, Vol. 1, pp. 133–136 (2001).

[16] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception", in proceedings of Eurospeech-2003, pp. 2117–2120, 2003.

[17] G.Shi et.al, "On the importance of phase in human speech recognition", IEEE Trans. Audio, Speech and Language Processing, Vol. 14, No. 5, pp. 1867–1874 (2006).

[18] P.Aarabi et.al, "Phase-based speech processing", World Scientific (2005).

[19] 浅川康平, 中川聖一, "MFCC と位相情報を用いた話者認識", 日本音響学会春季発表会, 講演論文集 1-P-17 (2007).

[20] 大塚真司, 王龍標, 中川聖一, "話者認識における位相情報の改善", 日本音響学会秋季発表会, 講演論文集 3-Q-2 (2007).

[21] 大塚真司, 王龍標, 中川聖一, "位相情報と MFCC の併用による話者認識の高精度化", 日本音響学会春季発表会, 講演論文集 3-11-2 (2008).

[22] S. Nakagawa, W. Zhang and M. Takahashi, "Text-independnt/text-prompted speaker recognition by combining speaker-specific GMM with speaker adapted syllable-based HMM", IEICE Trans., Vol. E89-D, No. 3, pp. 1058–1064 (2006).

[23] M-Jin, F.K.soong and C.D.Yoo, "A syllabbe lattice approach to speaker verification", IEEE Trans. Audio, Speech, Language Process., Vol. 15 , No. 8, pp. 2476–2484 (2007).

[24] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", Speech Communication, Vol. 17, No. 1-2, pp. 91–108 (1995).

[25] F. Bimbot et al., "A tutorial on text-independent speaker verification", EURASIP Journal on Applied Signal Processing 2004:4, pp. 430–451 (2004).

[26] A. E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F. K. Soong, "The use of cohort normalized scores for speaker verification", in Proc. ICSLP, Vol. 1, pp. 599–602 (1992).

[27] K. P. Markov and S. Nakagawa, "Text-independent speaker recognition using non-linear frame likelihood transformation", Speech Communication, Vol. 24, No. 3, pp. 193–209 (1998).

[28] A. E. Rosenberg, "Automatic speaker verification: a review", Proc. IEEE, Vol. 64, pp. 475–487 (1976).

[29] http://www.sunrisemusic.co.jp/dataBase/fl/noisedata01_fl.html

[30] Matusi, T., Furui, S., "Concatenated phoneme models for text-variable speaker recognition", in Proc. ICASSP, Vol. 2, pp. 391–394 (1993).

[31] H. Yamamoto, Y. Nankaku, C. Miyajima, K.Tokuda and T. Kitamura, "Parameter sharing in mixture of factor analyzes for speaker identification", IEICE Trans., Vol. E-88D, No. 3, pp. 418–424 (2005).