調波構造抽出と高信頼度フレーム選択を用いた雑音下での話者識別*

◎藤原弘将, 北原鉄朗 (京大), 後藤真孝 (産総研), 駒谷和範, 尾形哲也, 奥乃博 (京大)

1 はじめに

話者識別技術を実環境下で適用するためには，雑音に対する耐性が大きな課題となる．話者識別の雑音に対する耐性を高めるため，様々な研究が行われてきた．代表的手法として，スペクトルサブトラクション (SS) 法 [1]，隠れマルコフモデル (HMM) 合成法 [2, 3] や，それらの改良手法が知られている．SS 法は，雑音のパワースペクトルを，雑音が重畳した音声信号のパワースペクトルから減算することで，雑音を低減させる．HMM 合成法は，ケプストラム係数を特徴パラメータとする無雑音音声の HMM と雑音の HMM から，目的の雑音環境の音声 HMM を合成する．いずれの手法も，雑音のパワースペクトルが既知という前提を置いたり，突発性雑音や非定常雑音に弱い，という問題点があった．

本研究では，これら問題点を解決するため，**調波構造抽出**と**高信頼度フレーム選択**に基づく話者識別手法を提案する．調波構造抽出では，音声の調波構造を抽出し再合成する．これにより，雑音の影響が低減された音響信号を得ることを狙う．本手法は，雑音の性質を仮定していないため，未知の雑音環境に対しても頑健に機能する．高信頼度フレーム選択では，各フレームに対して音声としての信頼度を計算し，識別の際に，各特徴ベクトルの対数尤度を信頼度で重み付けする．これにより，信頼度の高い，すなわち雑音の影響が小さいフレームに高い重みが付与され，突発性雑音や非定常雑音に対して頑健となることを狙う．

2 話者識別手法

本研究では，各登録話者ごとの特徴ベクトルのデータベースに基づいて，提示された雑音環境の音声の発話者を同定する．本稿では，著者らが提案した伴奏を含む音楽音響信号中の歌手名の同定手法 [4] のアプローチを元に，雑音下話者識別に応用する．処理の流れを，図 1 に示す．

2.1 調波構造抽出

雑音の影響を軽減するため，音声の調波構造を抽出し，再合成することで，音声が分離された音響信号を得る．これらの処理は，以下のように行う．

2.1.1 基本周波数 (F0) 推定

後藤の PreFEst[5] を用いて，音声の F0 を推定する．PreFEst は，制限された周波数帯域において最も優勢な調波構造の F0 を EM 法を用いて推定する手法である．まず，観測スペクトルが，全ての可能な F0 に対応する音モデルの重みつき和からからなる確率モデル，

$$p(x|\theta^{(t)}) = \int_{\mathrm{Fl}}^{\mathrm{Fh}} w^{(t)}(F)p(x|F)dF, \qquad (1)$$

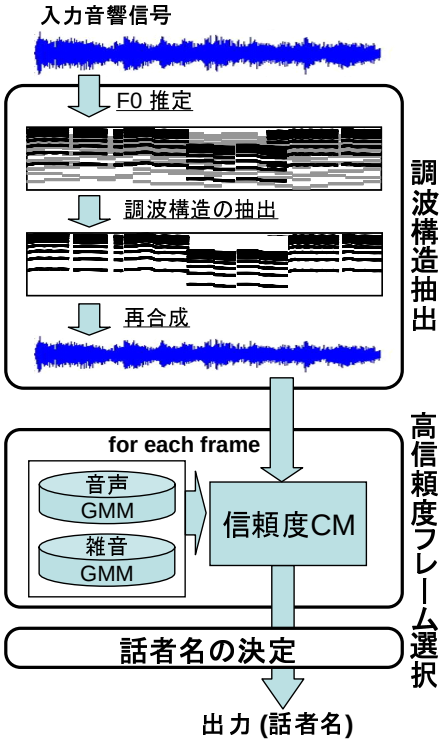

Fig. 1 処理の流れ

から生成されたと考える．ここで，$p(x|F)$ は，各 F0 についての音モデルであり，Fh と Fl は取り得る F0 の上限と下限である．また，$\theta^{(t)} = \{w^{(t)}(F)|\mathrm{Fl} \leq F \leq \mathrm{Fh}\}$ であり，$w^{(t)}(F)$ は音モデルの重みで，$\int_{\mathrm{Fh}_i}^{\mathrm{Fl}_i} w^{(t)}(F)dF = 1$ を満たす．音モデルとは典型的な調波構造を表現した確率分布である．そして，EM アルゴリズムを用いて $w^{(t)}(F)$ を推定し，$w^{(t)}(F)$ を最大にする F を最も優勢な基本周波数として決定する．

PreFEst は，音楽音響信号中のメロディの F0 を推定する手法であり，観測スペクトルが，楽器の伴奏等の調波構造を持つ音の混合から成り立つことを仮定している．そのため，雑音下での話し声の F0 推定に用いると十分な性能を発揮出来ない．しかし，後述する高信頼度フレーム選択を用いることで，F0 推定の誤差に対しても頑健となることを狙う．

2.1.2 調波構造の抽出

推定された F0 に基づき，歌声の調波構造の各倍音成分のパワーを抽出する．それぞれの周波数成分の抽出の際には，前後 20 cent ずつの誤差を許容し，この範囲で最もパワーの大きなピークを抽出する．

2.1.3 再合成

抽出された調波構造を正弦波重畳モデルに基づき再合成する．再合成された音響信号 $s(t)$ は，

*Speaker identification under noisy environments based on harmonic structure extraction and reliable frame selection. by H. Fujihara, T. Kitahara (Kyoto University), M. Goto (AIST), K. Komatani, T. Ogata, and H. G. Okuno (Kyoto University).

Table 1　実験データ・分析条件

音声データ	ASJ-JNAS 音素バランス文 30 話者 (男 15 人，女 15 人) 学習：10 発話，評価：40 発話
雑音データ	効果音大全集 [6] より 学習時：大ホールのロビー 評価時：パーティ会場
分析条件	16 kHz，16 bit，フレーム長 160 ms， フレーム周期 10 ms, ハミング窓
特徴量	12 次元 LPC メルケプストラム [7]

$$s(t) = \sum_{k=1}^{K} A_k \cos(\omega_k t), \qquad (2)$$

と表わされる．ここで，A_k, ω_k はそれぞれ，k 次倍音のパワー，周波数を表わし，t は時間を表わす．

2.2　高信頼度フレーム選択に基づく話者の決定

各フレームの特徴ベクトル $\boldsymbol{x}$の信頼度 CM($\boldsymbol{x}$) を，音声を表す特徴量で学習した音声 GMM (混合ガウス分布) λ_{V}と，ノイズを表す特徴量で学習したノイズ GMM λ_{N}，PreFEst によって得られた F0 の音モデルの重み $w^{(t)}(F)$ を用いて，

$$\mathrm{CM}(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\lambda_{\mathrm{V}})}{p(\boldsymbol{x}|\lambda_{\mathrm{V}}) + p(\boldsymbol{x}|\lambda_{\mathrm{N}})} w^{(t)}(F) \qquad (3)$$

と定義する．本研究では，GMM の混合数として 64 混合を用いた．

話者名の決定には，64 混合 GMM を用いる．それぞれの登録話者について，GMM $\lambda_1, \cdots, \lambda_I$ (添字は話者ラベル) を事前に学習する．識別結果である話者ラベルは，以下の式に基づいて決定される．

$$s = \underset{i}{\mathrm{argmax}} \frac{1}{T} \sum_{t=1}^{T} \mathrm{CM}(\boldsymbol{x}_t) \log p(\boldsymbol{x}_t|\lambda_i) \qquad (4)$$

ただし，$\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}$ を識別対象の特徴ベクトルの時系列とし，$p(\boldsymbol{x}|\lambda_i)$ を，話者 i を表わす GMM の尤度とする．

3　評価実験

3.1　実験条件

本手法の有効性を確認するため，テキスト独立話者識別実験を行った．実験データと分析条件を表 1 に示す．実験に用いる雑音重畳音声は，雑音信号と音声信号の S/N 比が 0 dB となるように合成したものである．また，学習時と評価時には異なる種類の雑音を用いている．なお，学習データには，調波構造抽出は行うが，高信頼度フレーム選択は行わず，発話全体を用いて GMM を学習した．

高信頼度フレーム選択の学習データには，ASJ-JNAS データベースの評価に用いていない 274 話者から各 1 発話づつ選択し，学習用雑音を重畳して用いた．これらのデータには，伴奏音抑制を処理した後，音声 GMM と雑音 GMM を学習する．音声 GMM には，正解の F0 を用いて調波構造抽出・再合成された信号を用いた．このとき，正解の F0 は，無雑音音声信号から推定した．雑音 GMM には，PreFEst で推定された F0 密確率度関数の上位のピークから，音声の F0 以外の成分を選び，それらを用いて調波構造抽出・再合成された信号を用いた．

Table 2　実験結果 ただし，extarct.，select. はそれぞれ，調波構造抽出，高信頼度フレーム選択を表す．また，無雑音は，F0 推定を無雑音信号から行ったことを表す．

条件	(i)	(ii)	(iii)	(iv)	(v)
雑音	有り (S/N 0 dB)				無し
F0 推定	PreFEst		-	無雑音	-
extract.	○	○	×	○	×
select.	○	×	×	×	×
誤り率 (%)	10.7	26.5	18.9	5.1	1.5

3.2　結果と考察

表 2 に，実験結果を示す．条件 v(無雑音の場合) では 1.5%と低い誤り率だが，雑音を重畳することで誤り率が 18.9%(条件 iii) と大幅に増加している．それに対し，本手法 (条件 i) を用いることで，誤り率が 18.9%から 10.7%に低下し，誤り率を約 43%削減出来たことがわかる．これらから，本手法によって雑音に対する頑健性が増加したことが確認できる．

条件 ii と iv を比較すると，誤り率が 26.5%と 5.1%と大きな差がある．これは，F0 推定誤りによって，性能が低下したことを意味している．しかし，伴奏音抑制だけでなく，高信頼度フレーム選択を併せて行うことで，誤り率が 26.5%から 10.7%へ低下した．これらから，高信頼度フレーム選択を行うことで，F0 推定を誤ったフレームは低い信頼度となり，識別結果への影響が小さくなったため，誤り率が削減されたことが確認された．

4　おわりに

本稿では，雑音環境化での話者識別手法として，調波構造抽出と高信頼度フレーム選択を提案した．また，評価実験により，誤り率を約 43%削減し，本手法の有効性を確認した．今後は，様々な雑音環境について評価実験を行うと共に，S/N 比が未知の場合を扱えるように手法を拡張する予定である．

謝辞　本研究の一部は，日本学術振興会科学研究費補助金，21 世紀 COE プログラム，CREST の支援をうけた．

参考文献

[1] S. F. Boll, “Suppression of Acoustic Noise in Speech using Spectral Subtraction,” IEEE trans. on ASSP., **27**, 113–120, 1979.

[2] R. C. Rose and E. M. Hofstetter and D. A. Reynolds, “Integrated Models of Signal and Background with Application to Speaker Identification in Noise,” IEEE trans. on SA, **2**, 245–257, 1994.

[3] M. J. F.Gales and S.J.Youn, “Robust Speech Recognition using Paralles Model Combination,” IEEE trans. on SA, **4**, 352–359, 1996.

[4] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, H. G. Okuno, “Singer Identification Based on Accompaniment Sound Reduction and Reliable Frame Selection,” Proc. ISMIR2005, 329–336, 2005.

[5] M. Goto, “A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals,” Speech Com., **43**, 311–329, 2004.

[6] 効果音大全集 45 屋内ノイズ II, King Record.

[7] 今井聖: 音声信号処理 音声の性質と聴覚の特性を考慮した信号処理, 森北出版株式会社, 1996.