単 1 のマイクロフォンを利用した同時発話音声認識*

○岡本一輝，村上仁一，池原悟 (鳥取大)

1　はじめに

複数の話者が話したときに，各話者が話した音声の認識を行う場合，複数のマイクロフォンを用いる手法が一般的である [1]．しかし，人間では 1 つの耳だけで複数の音声を聞き分けることが出来る．このような単 1 のマイクロフォンで音声認識を行う研究例は少ない．類似した研究として，複数話者の重畳音声を認識する場合，音声を分離する手法が研究されており，音源モデルを利用した方法が提案されている [2].

本研究では，男女 2 話者が同時に発話した場合に，単 1 のマイクロフォンを使用した状況を想定し，認識率の調査を行う．まず男女個別のモデルを利用して，単純な方法で認識実験を行う．また，雑音環境における認識の手法である，Parallel Model Combination 法に類似した手法を用いて認識実験を行う．

2　単純法

単純法では，男女 2 話者が同時に発話した音声に対し，男性話者，女性話者の HMM をそれぞれ利用して認識を行う．各々の認識結果に対して男性話者と女性話者が同時に認識できた場合の認識率を調査する．

3　Parallel Model Combination 法

Parallel Model Combination 法 [3](以下 PMC 法) は，雑音が重畳した音声を認識する一般的な方法である．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 を合成し，認識を行う．図 1 に PMC 法のモデルを示す.

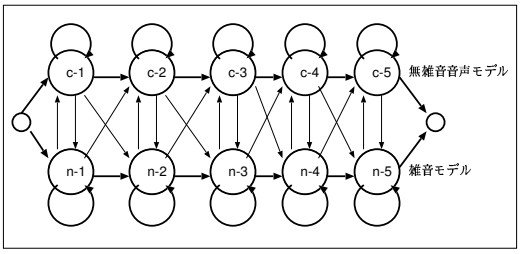

Fig. 1　PMC 法のモデル

4　評価実験

4.1　評価データと学習データ

本研究では，実際に男女 2 話者が同時に発話した音声を使用せず，各々の話者の音声を重畳した音声を利用する．具体的には，ATR 単語発話データベース Aset の男性話者 mau，mms 及び女性話者 ftk，fyn の男女各 2 名を使用する．偶数番号の音声の中から 4 モーラで発話時間がほぼ同じ語を，ランダムに 10 単語ずつ抽出する．それぞれを重畳した音声 (以下クロストーク音声) を作成する．1 セットにつき 100 単語のクロストーク音声を 4 セット作成し，評価データとして利用する．奇数番号の音声は HMM の学習データとして使用する．表 1 に実験に使用した単語を示す．また，図 2 にクロストーク音声認識の手順を示す.

Table 1　実験に使用した単語

	男性話者	女性話者
1	悪質 (akushitsu)	足元 (ashimoto)
2	聞こえる (kikoeru)	可愛い (kawaii)
3	加える (kuwaeru)	勤勉 (kiNbeN)
4	失恋 (shitsureN)	細々 (komagoma)
5	優れる (sugureru)	すまない (sumanai)
6	そのうち (sonouchi)	対策 (taisaku)
7	中毒 (chuudoku)	手拭い (tenugui)
8	内容 (naiyou)	天才 (teNsai)
9	暴力 (bouryoku)	滅ぼす (horobosu)
10	わざわざ (wazawaza)	欲張る (yokubaru)

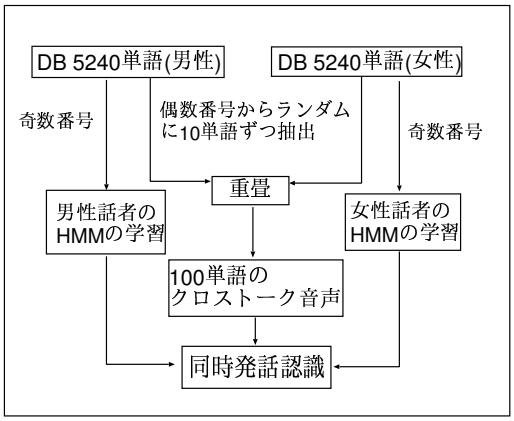

Fig. 2　クロストーク音声認識の手順

4.2　本研究における疑似 PMC 法

本研究では，HTK[4] を使用して実験を行う．しかし HTK で PMC 法を簡単に利用するすることができない．そこで本研究では Parallel Model を構築する際, 音声を音素単位で考え, 各モーラごとに子音と母音に分け，子音は子音同士で母音は母音同士で相互にパスを持った疑似 PMC 法のモデルを 100 個構築する．認識実験においては各クロストーク音声に対し各疑似 PMC 法のモデルの尤度を求め，最も尤度が高かった疑似 PMC 法のモデルを認識結果とする．図 3 に本研究での疑似 PMC 法のモデルを示す.

*Simultaneous speech recognition using single microphone. by OKAMOTO Kazuki, MURAKAMI Jin'ichi and IKEHARA Satoru(Tottori University)

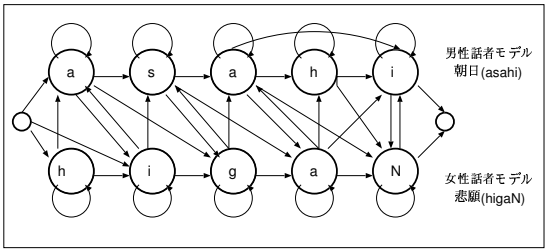

Fig. 3　本研究での疑似 PMC 法のモデル

図 3 は男性話者の音声に「朝日 (asahi)」，女性話者の音声に「悲願 (higaN)」を使用した場合の疑似 PMC 法のモデルである．

4.3　実験条件

実験条件を表 2 にまとめる．本研究では特徴パラメータに MFCC，FBANK 及び MELSPEC を使用する．また音素 HMM の共分散行列には Diagonal-covariance と Full-covariance を使用し，認識率を調査する．

Table 2　実験条件

基本周波数	16kHz
分析窓	Hamming 窓
分析窓長	25ms
フレーム周期	10ms
音響モデル	3 ループ 4 状態・連続混合分布型
stream 数	2(特徴パラメータ + 対数パワー)
混合分布数	母音・撥音・無音 4mixture 子音 2mixture

4.4　実験結果

表 3 及び表 4 に単純法及び疑似 PMC 法の認識率を示す．表中の括弧内の分母は評価データの総数を，分子は認識できたクロストーク音声数を示す．

Table 3　実験結果 (単純法)

話者	mau+ftk	mau+fyn	mms+ftk	mms+fyn	平均
MFCC Diagonal	47% (47/100)	44% (44/100)	41% (41/100)	48% (48/100)	45% (180/400)
MFCC Full	57% (57/100)	62% (62/100)	54% (54/100)	50% (50/100)	56% (223/400)
FBANK Diagonal	27% (27/100)	40% (40/100)	29% (29/100)	38% (38/100)	34% (134/400)
FBANK Full	44% (44/100)	48% (48/100)	50% (50/100)	44% (44/100)	47% (186/400)
MELSPEC Diagonal	8% (8/100)	8% (8/100)	8% (8/100)	17% (17/100)	10% (41/400)

Table 4　実験結果 (疑似 PMC 法)

話者	mau+ftk	mau+fyn	mms+ftk	mms+fyn	平均
MFCC Diagonal	46% (46/100)	55% (55/100)	34% (34/100)	44% (44/100)	45% (179/400)
MFCC Full	52% (52/100)	49% (49/100)	38% (38/100)	42% (42/100)	45% (181/400)
FBANK Diagonal	28% (28/100)	51% (51/100)	30% (30/100)	40% (40/100)	37% (149/400)
FBANK Full	41% (41/100)	40% (40/100)	34% (34/100)	35% (35/100)	38% (150/400)
MELSPEC Diagonal	30% (30/100)	45% (45/100)	24% (24/100)	32% (32/100)	33% (131/400)

実験より以下の結果が得られた．

1. 最も認識精度が高かった実験は，単純法 MFCC Diagonal で，平均 56% の精度が得られた．
2. 音素 HMM の共分散行列に Diagonal を使用した実験において，疑似 PMC 法が単純法より認識精度が高い．
3. 特徴パラメータにおいて，MFCC，FBANK，MELSPEC の順で認識精度が高い．

5　考察

5.1　人間による聴覚実験

実験に使用したクロストーク音声に対して，人間による聴覚実験を行い，認識率を求めた．なお，被験者は男性 3 名，女性 1 名である．表 5 に認識結果を示す．

Table 5　実験結果 (人間による聴覚実験)

話者	mau+ftk	mau+fyn	mms+ftk	mms+fyn	平均
認識率	77% (77/100)	87% (87/100)	74% (74/100)	70% (70/100)	77% (308/400)

表 3，4 と表 5 を比較すると，計算機による誤り率は人間と比べて 2 倍程度である (単純法 MFCC Diagonal)．単純な手法を用いた実験ということを考慮すると，認識率は高いと考えている．また，計算機では話者ごとの認識率のばらつきが大きい．

5.2　本研究における疑似 PMC の有効性

評価実験において疑似 PMC 法で改善が見られたものは，FBANK Diagonal と MELSPEC Diagonal であった．しかし両者共に単純法における認識精度が低い．

以上より，本研究における疑似 PMC 法は単純法での低い認識精度の実験の精度を改善する効果がある．

6　おわりに

男女 2 話者が別々の音声単語を同時に発声した状況を想定し，単 1 のマイクロフォンを利用した音声認識における単純法と疑似 PMC 法の認識率を調査した．その結果，人間による聴覚実験と比較すると誤り率で 2 倍程度の認識率が得られることがわかった．今後は，PMC 法を実現し認識率の調査を行う．

参考文献

[1] Panikos Heracleous, Satoshi Nakamura, Takeshi Yamada, and Kiyohiro Shikano. “ A Microphone Array-Based 3-D N-Best Search Method for Recognizing Multiple Sound Sources. ”IEICE Transactions on Information and Systems Vol.E85-D, No.6, pp.994-1002, 2002.

[2] 武藤，杉山. “ 時間拘束条件下での重畳音声分解法. ” 日本音響学会講演論文集，No. 3-3-21，pp. 135-136, 2001.

[3] M.J. Gales, and S.J. Young. “ Robust Continuous Speech Recognition Using Parallel Model Combination. ” IEEE Transactions on Speech and Audio Processing, Vol.4, No.5, pp.352-359, 1996.

[4] Steve Young, Phil Woodland, and Gunnar Evermann. “ HTK Boo k. ” Cambridge University Engineering Department，2002.