4B-3

GAによる話者位置への同時発話認識システムの最適化

山本 俊一 * 中臺 一博 † 中野 幹生 † 辻野 広司 † Jean-Marc Valin‡
武田 龍 § 駒谷 和範 * 尾形 哲也 * 奥乃 博 *
* 京都大学情報学研究科知能情報学専攻 †(株) ホンダ・リサーチ・インスティチュート・ジャパン
‡CSIRO ICT Center § 京都大学工学部情報学科

1. はじめに

人と同じ生活環境で活動するロボットが人とコミュニケーションするためには，人と同じように様々な状況下で多様な音を聞き分ける機能が必要である．田崎らは近接学に基づいて，距離によってコミュニケーションする相手を適切に選択する手法を開発している [1]．同様に，ロボットの音声インタフェースも，相手の位置によって最適な音声認識パラメータを選択できれば，混合音の認識精度の向上が期待できる．

本研究の目的は，同時発話認識システム [2] の音声認識の精度を向上させるために音源分離とミッシングフィーチャマスク（*Missing Feature Mask*, MFM）自動生成のパラメータを最適化し，話者の方向と距離によって最適なパラメータを利用できるようにすることである．パラメータ同士の依存関係を考慮して手動で調整することは困難であるので，本稿では，遺伝的アルゴリズム（*Genetic Algorithm*, GA）[3] によってパラメータを最適化し，位置による最適パラメータの選択について検討する．

2. 同時発話認識システム

本節では，同時発話認識システム（図 1）を最適化すべきパラメータの観点から説明する．

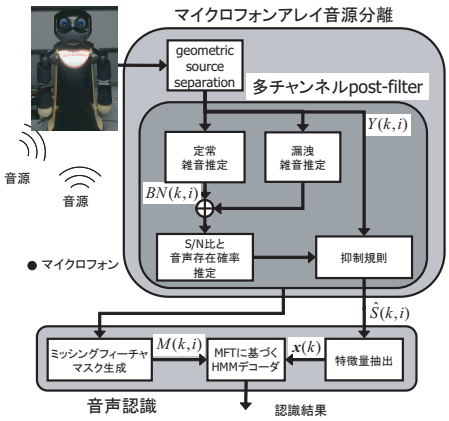


図 1: 同時発話認識システムの概要

2.1 マイクロフォンアレイ音源分離

マイクロフォンアレイ音源分離 [4] は，Geometric Source Separation (GSS) と多チャンネル post-filter から構成される．詳細は [4] を参照されたい．GSS はマイクロフォンの幾何的な制約による適応ビームフォーマによる音源分離手法である．多チャンネル post-filter は，GSS によって分離された音から個々に推定した背景雑音（定常性雑音）と他音源からの漏洩（非定常性雑音）を統合して雑音推定を行い，分離音から雑音除去処理を行う．

マイクロフォンアレイ音源分離システムで利用されるパラメータのうち，音声認識の精度に関係しそうなものは 12 個である．これらを大きく分けると以下の 3 種類に分けられる．

(1) 他音源からの漏洩推定のための係数に関するもの
(2) 事前 S/N 比推定のための係数に関するもの
(3) 残響に関するもの

2.2 MFM 自動生成

MFM を自動生成するには，分離音声のスペクトルのうち，どの周波数帯域が歪んでいるかという情報が必要である．本手法では，多チャンネル post-filter の入力，分離音，推定された背景雑音のスペクトルを利用して，MFM を生成する．

まず，多チャンネル post-filter から得られる入力，出力，背景雑音のスペクトルをスペクトル特徴量の抽出と同様にメル周波数スペクトルに変換する．フレーム k，サブバンド i のときの多チャンネル post-filter への入力を $Y(k,i)$，多チャンネル post-filter からの出力を $\hat{S}(k,i)$，多チャンネル post-filter で推定された背景雑音を $BN(k,i)$ とする（図 1）．

もし，目的方向の音源，または背景雑音が支配的な場合には，そのメル周波数帯域は信頼できる特徴量であるとする．背景雑音が支配的な場合は，その帯域が無音区間である．このような考察から，MFM を以下のように定義した．ただし，$\{M(k,i)|i=1,\cdots,N\}$ はメル周波数対数スペクトル特徴量に対する MFM を，$\{M(k,i)|i=N+1,\cdots,2N\}$ はスペクトル特徴量の一次微分に対する MFM を表している．

スペクトル特徴量に対する MFM$(i=1,\cdots,N)$ は，

$$M(k,i)=\begin{cases}1, & \frac{S(k,i)+BN(k,i)}{Y(k,i)}>T_{MFM}\\ 0, & otherwise\end{cases}, \qquad (1)$$

スペクトル特徴量の一次微分に対する MFM$(i=N+1,\cdots,2N)$ は，

$$M(k,i)=\prod_{i=k-2,i\neq k}^{k+2} M(i,i-N) \qquad (2)$$

である．

MFM 自動生成で利用されるパラメータは閾値 T_{MFM} で，主に多音源から漏洩したパワーの大きさに依存している．MFT に基づく音声認識システムの原理については，[2] を参照されたい．

3. GA によるパラメータ最適化

2. 章で述べた同時発話認識システムのパラメータを自動的に最適化するために，GA を利用して自動的に最適化を行うようにした．本節では 2. 章で述べた同時発話認識システムのパラメータを GA によって最適化する手法について述べる．

Optimization of Recogniton System of Simultaneous Speech Signals by GA
by Shunichi Yamamoto (Kyoto Univ.), Kazuhiro Nakadai, Mikio Nakano, Hiroshi Tsujino (HRI-JP), Jean-Marc Valin (CSIRO ICT Center), Ryu Takeda, Kazunori Komatani, Tetsuya Ogata and Hiroshi G. Okuno (Kyoto Univ.)

本稿では，各遺伝子が有限の集合の要素のうちのいずれか一つを表すことにする．例えば，遺伝子を以下のように表現する．

遺伝子 1 {0.1, 0.2, 0.3, 0.4, 0.5}

遺伝子 2 {0.9, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, 0.99, 0.1}

このとき，染色体を {0.3, 0.93} のように表す．

GA による最適化の手順は以下の通りである．

(1) 初期集団として，個体を N 個発生させる．各個体の遺伝子はランダムに決定する．
(2) 集団内からランダムに M 個のペアを選択し，一様交叉法で $2M$ 個の子供を生成する．
(3) 各個体の各遺伝子を突然変異率 p_m でランダムに対立遺伝子に置き換える．
(4) 適応度として音声認識の単語正解率を計算する．
(5) 交叉により，個体の数が $N + 2M$ 個に増加しているので，環境に適合しない個体を淘汰し，個体数を N にする．淘汰の方法として，まずエリート選択により適応度の高い順に L 個だけ選択し，残りをルーレット選択する．
(6) 適応度の最大値が数世代変化しなければ収束したとみなし，そうでなければ (2) に戻る．

4. 実験

同時発話認識システムの単語正解率を位置（距離と方向）に関して最適化を行い，最適化したパラメータを評価するための実験を行った．

4.1 データセット

録音には 8 本のマイクロフォンを設置した Robovie-R2 を利用した．録音を行った部屋は 4 m × 5 m の大きさで，残響時間（RT20）が約 0.2 秒であった．ロボットとスピーカの間の距離は 50，100，150，200，250 cm の 5 種類で，スピーカ同士の間隔は 30，60，90 度 に配置した．スピーカはすべて同じ距離に設置した．録音した音声のデータセットは，二話者同時発話 2 セット，三話者同時発話，一話者とポピュラー音楽，一話者とクラシック音楽である．

4.2 音声認識

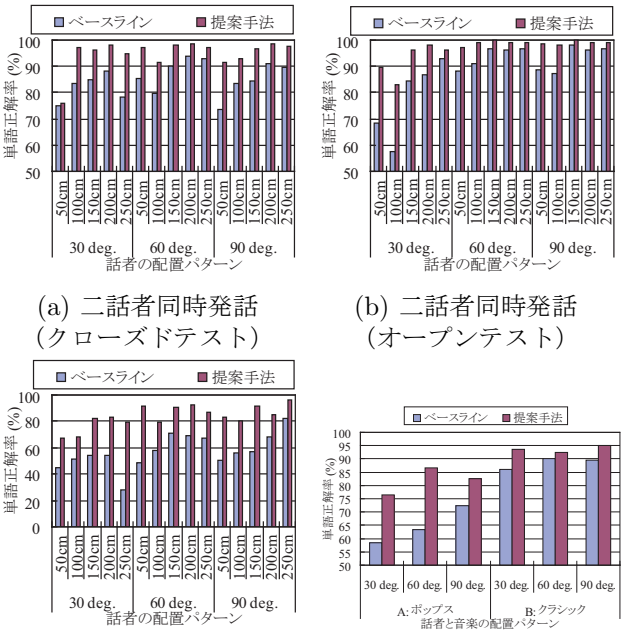
本システムではマルチバンド版 Julian を MFT に基づく音声認識システムとして利用した．音響モデルにはトライフォン（3 状態 4 混合の HMM）を利用し，言語モデルには有限状態文法を利用して 200 語の孤立単語認識を行った．利用したトライフォンは，25 話者分の ATR 音素バランス単語 216 語で学習した．

4.3 GA のパラメータ

GA のための遺伝子と対立遺伝子は，それぞれ同時発話認識システムのパラメータとパラメータの取りうる値の集合に対応している．実験では，集団サイズ $N = 100$，交叉数 $M = 40$，エリート選択数 $L = 10$，突然変異率 $p_m = 0.01$ に設定した．

4.4 最適化したパラメータの評価

最適化したパラメータを評価するために，GA で最適化したパラメータの集団の中から最も適応度の高かったパラメータのを共通の設定として二話者同時発話，三話者同時発話，一話者と音楽の混合音の中央話者の音声認識を行った．最適化したパラメータとベースライン [2] のパラメータによる音声認識結果を図 2 に示す．



(a) 二話者同時発話（クローズドテスト）　(b) 二話者同時発話（オープンテスト）

(c) 三話者同時発話　(d) 一話者と音楽の混合音

図 2: 最適パラメータを利用した同時発話認識結果

平均単語正解率は，二話者同時発話認識のクローズドテスト（最適化と同じデータセットを利用した場合）では 84.9%から 94.7%まで 9.8%，オープンテスト（最適化と異なるデータセットを利用した場合）では 83.3%から 96.7%まで 13.4%，三話者同時発話認識では 57.4%から 83.9%まで 26.5%，一話者と音楽の混合音のうち話者の音声を認識した場合には，76.7%から 87.8%まで 11.1%，それぞれ向上した．

5. おわりに

三話者同時発話認識のようなタスクでは多数のシステムが統合されているので，最適なパラメータで実現していない場合がある．本稿では GA による最適化を試み，得られたパラメータの効果を評価した．その結果，二話者同時発話の中央方向に関して得られた最適化されたパラメータが，話者数の同じ二話者同時発話だけではなく，話者，音源の異なる混合音で中央方向の話者の音声認識にも有効であることが分かった．今後の課題として，音源位置とパラメータの間の関係を明らかにし，自由に配置された音源に対して自動的に最適パラメータを選択する方法の開発が上げられる．

なお，本研究の一部は，科学研究費補助金，21 世紀 COE プログラム，および，電気通信普及財団の支援を受けた．

参考文献

[1] Tasaki, T. *et al.*: Spatially Mapping of Friendliness for Human-Robot Interaction, *IEEE/RSJ IROS 2005*, 521–526.

[2] Yamamoto, S. *et al.*: Enhanced Robot Speech Recognition Based on Microphone Array Source Separation and Missing Feature Theory, *IEEE-RAS ICRA 2005*, 1489–1494.

[3] 樋口哲也 他: 遺伝的アルゴリズムとその応用,『情報処理』, Vol. 34, No. 7, 871–883 (1993).

[4] Valin, J.-M. *et al.*: Enhanced robot audition based on microphone array source separation with post-filter, *IEEE/RSJ IROS 2004*, 2123–2128.