6J-8

実環境音声認識のためのロボット聴覚システム開発とパラメータチューニング

高橋 徹 † 中臺 一博 ‡ 駒谷 和範 † 尾形 哲也 † 奥乃 博 †

† 京都大学大学院 情報学研究科 ‡（株）ホンダ・リサーチ・インスティチュート・ジャパン

1. はじめに

実環境下音声認識では, 複数同時に呈示される音声の認識が重要である. それを実現し, ロボットに搭載することで, 我々は, ロボットと自然な音声コミュニケーションを達成できる. 著者らは, ロボットが実環境下で音声を理解する枠組をロボット聴覚と呼び, ロボット聴覚システム HARK (HRI-JP Audition for Robots with Kyoto University) [5] を開発している. このシステムをヒューマノイドロボット HRP-2 に実装し 3 話者同時発話認識を可能にした [1]. ロボット聴覚システムには, 多くのチューニング可能なパラメータが存在し, 最適化が困難な問題がある. この問題に対し, 山本ら [4] は, GA を用いた最適化手法を開発した. 分離音声のための音響モデルの適応法は, 高橋ら [2] によって開発されている. しかし, パラメータ最適化と, 分離歪を含む音声への音響モデルの適用処理の両方を, 同時に最適化する手法は, 開発されていない. これらを現実的な時間で同時に最適化する手法を開発する. 開発手法の利点は, 各最適化ステップで単語正解精度がクローズドデータに対して単調増加する点, 最適化ステップで計算コストの主要部分を占める音響モデルの再構築の回数を削減できる点である. 音源分離パラメータと音響モデルの分離歪への適用を同時に行い, 音源分離のパラメータのみ最適化した場合から 8 % 単語正解精度を改善できた.

2. ロボット聴覚システム

我々は, 情報の爆発的増大を info-plosion [3] と呼ぶことに倣い, 音声情報の同時多発的発生を info-plosion sound と名付ける. info-plosion sound 時代の音声認識には同時に呈示される音声を認識する必要があり, ロボット聴覚システムは, それを可能にする. 我々は, をロボット聴覚研究プラットフォーム HARK を OSS として公開している. HARK は, 音源定位・音源分離・分離音認識をはじめ, 多くの処理モジュールを含んでいる.

info-plosion sound の問題は, 混合音声を定位, 分離し, 認識する問題である. ロボット聴覚システムのパラメータを同時に呈示される音声の認識精度の最大化するために最適化しなければならない. 最適化には, 混合音声データセット $X(n), n = 1, \ldots, N$ に対して分離処理を行い, 分離音声の認識精度を最大化する. 音源数 M が, 既知としてシステムを構築する. 音源定位部・音源分離部・分離音認識部分のハイパーパラメータを Π, Ψ, Φ とする. パラメータ Π, Ψ で分離した音声に適応した音響モデルを $M(\Pi, \Psi)$ とする. 時刻 n における C チャネル, L サンプルからなる音声フレームを $C \times L$ 行列 $X(n)$ とする.

音源定位処理, 音源分離処理, 分離音声の認識精度は,

$$\theta(n) = L(X(n); \Pi) \quad (1)$$

$$Y(n) = S(X(n), \theta(n); \Psi) \quad (2)$$

$$A = R([Y(n+1), \ldots, Y(n+K)]; M(\Pi, \Psi), \Phi) \quad (3)$$

と表される. 定位処理結果 $\theta(n)$ は, ベクトルで, 音源数分 M 個の定位角度を要素に持つ. 分離処理結果 $Y(n)$ は, $M \times L$ 行列である. K は, 音声認識する (単語, 文などの) 分析単位を含む十分な長さを表し, A は, 音源数 M 個の要素を持つベクトルで, 各要素に音源毎の認識精度である.

直接的なパラメータチューニングは,

$$\arg\max_{\Pi, \Psi, \Phi, M(\Pi, \Psi)} \mathrm{ave}(A) \quad (4)$$

となる. ただし, ave() は, ベクトル要素の平均を求める関数とする. しかし, 最適化すべきパラメータ数が多いため A の直接最大化は困難である. 何故ならば, Π, Ψ を変更する度に, 音響モデル M の再構築が必要で, 現実的な処理時間で処理できない. 部分的にチューニングすることで逐次最大化を行う. この手法は, 局所最大値に陥る可能性があるが, 音響モデルの再構築を辟けることで, 現実的な時間で準最適化が可能である.

Π を固定し, 音源定位を他のパラメータと独立に最適化する. 最適パラメータを Π^* とし, 定位の正解を $\theta^*(n)$ とおくと, 定位誤差最小化基準で

$$\Pi^* = \arg\min_{\Pi} \sum_n |\theta^*(n) - \theta(n)|^2 \quad (5)$$

最適化可能である. 次に Π^* のもとで, 音源分離部を最適化する. 最初に使用する音響モデル $M^{(0)}$ を用意する. 通常は, クリーンな音声コーパスから学習した音響モデルを用いる. このもとで Ψ を最適化する. 最適な分離パラメータを $\Psi^{(i+1)}, (i = 0)$, とおくと,

$$\Psi^{(i+1)} = \arg\max_{\Psi} R([\hat{Y}(1), \ldots, \hat{Y}(n)]; M^{(i)}, \Psi) \quad (6)$$

$$\hat{Y}(n) = S(X(n), \theta^*(n); \Psi) \quad (7)$$

によって最適化できる. 必要に応じて, $M^{(i+1)} = M(\Pi^*, \Psi^{(i+1)})$ を用い, 反復的に Ψ と, M を最適化できる.

3. パラメータ最適化

音源定位・音源分離・分離音声認識に基くロボット聴覚システムに, 開発したパラメータ最適化手法を適用可能である. HARK を例に具体的最適化手順を説明する.

3.1 音源定位部分の最適化

音源定位部分は, 4 つのモジュール, LocalizeMUSIC, SourceTracer SourceIntervalExtender , SourceSelectorByDirection で構成している. 音源定位は, MUSIC 法に基いている. MUSIC 法では, マイク毎に全方向の伝達関数が必要である. 音源定位部分のチューニングの目的は, 定位誤差を最小化であり, 定位誤差は, 与えるインパルス応答に依存する. 実測のものや合成したものを用いる. インパルス応答を用意後, 最終的にチューニングすべきパラメータは, SourceTracer の音源の存在 VA (Voice Activity) を判定する閾値, SourceIntervalExtender の短い無音区間を無視し連続して音源が存在すると判定する閾値である. 定位部分は, VAD (VA Detection) の精度を最大化する.

Robot audition system development and parameter-turning in real environment : *Toru Takahashi (Kyoto Univ.), Kazuhiro Nakadai (HRI-JP), Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

3.2 音源分離部の最適化

音源分離部は, GSS (Geometoric Source Separation) モジュールと PostFilter モジュールで構成している. GSS では, LocalizeMUSIC モジュール同様に, マイク毎に全方向の伝達関数が必要であり, 用意する伝達関数によって分離性能が異なる. チューニングすべきパラメータは, ポストフィルタの特性を決めるパラメータである.

GSS は, 同時に呈示された音源の定位情報に基いて, 分離行列を適応する. 音源定位部の定位誤差の取扱が最適化に影響する. 式 (6),(7) において, 定位情報に正解 $\theta^*(n)$ を用いる例を示した. 定位誤差に頑健な音響分離パラメータと認識用パラメータを求めるために, 実際の定位結果 $\theta(n)$ を用いる場合もある. 定位誤りは, 分離誤りの原因であるが, 分離誤りを含む分離歪に音響モデルを適用することで, 実際の分離状況に近い音響モデルを構築できる.

3.3 分離音認識部の最適化

分離音認識部は, SpeechRecognitionClinet または, SpeechRecognitionSMNClinet と認識エンジンである拡張版の Julius から構成した. 認識エンジンのパラメータの最適化と音響モデルの分離音への適用が必要である. 認識エンジンのパラメータの最適化は, 紙面の都合上, 割愛する. 実験では, Julius 3.5 のデフォルト設定値を使用した.

複数の音源を混合し, 音響モデル学習用のデータベースを一度分離した分離歪を含む音声データベースを用意する. 分離音声データベースから音響モデルを学習する. 混合音声を作成するにあたり, 音源数とそれぞれの音源の位置は, 実際のロボット聴覚システムを使用状況を想定して決める.

4. 実験・考察

2 話者同時発話システムで, 立単語認識精度を最大化するチューニング実験を行う. ヒューマノイドロボット HRP-2 のマイク配置のみから音源分離・定位・分離を行うシステムを用い, 音源分離部分のパラメータと音響モデルの最適化実験を行った. 比較の単純化のため, 音源分離部分でのチューニング対象を, 音源間の干渉量の事前推定値 (LF : Leakfactor) とする. 発話者の位置はロボットの 0 度方向と 120 度方向に固定して評価した.

LF は, 0.25 から 0.70 まで 10 段回 0.05 ステップの範囲でチューニングした. 音響モデルは, Japanese Newspaper Article Sentences (JNAS) 中の音素バランス文から学習した Monophone HMM である. 初期音響モデル M_{Init} は, クリーンなデータベースから学習した. 認識精度の評価には, ATR 音声データベースの音素バランス 216 単語中の 10 単語・男女各 3 名による発話を任意の組み合わせで同時に呈示し, 分離語の単語正解精度を最大にする LF を求める. 評価は, 発話内容と初話者がオープンなデータで音声認識精を求めている. 従って, 必ずしもパラメータ最適化手順において, 精度が単調増加するとは限らない.

各 LF に対する精度を図 1 の実線に示す. 最初に用意するクリーンな音響モデル $M^{(0)}$ において, $LF = 0.55$ が最適であることがわかる. 次に, $LF = 0.55$ で分離した音声データベースを用いて音響モデルを再構築する. このモデル $M^(1)$ で, 各 $LF = 0.55$ に対する認識精度を黒点線で表す. 音響モデルの再構築による認識精度の向上が認められる.

更にチューニングするために, 各 LF に対して単語精度を求めると図 1 の 点線の様になる. 最大値が, 前回と

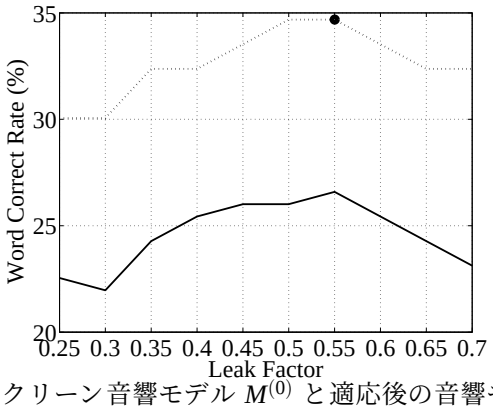

図 1: クリーン音響モデル $M^{(0)}$ と適応後の音響モデル $M^{(1)}$ を用いた分離音声の単語正解精度.

同じ $LF = 0.55$ となり, 再度分離した音声データベースで音響モデルの再構築が不要となり, チューニングを終了する. $LF \neq 0.55$ が最大となる場合には, その LF 値で, 分離した音声データベースで音響モデルを再構築することで, 音響モデルのチューニングを進められる.

実験では, LF を 10 段階でチューニングを行うに当り, 音響モデルの学習は 1 回であった. 分離パラメータと音響モデルを同時に直接最大化するには, それぞれのパラメータで分離し, その分離パラメータで分離した音声データベースで, 音響モデルを再構築する必要がある. この方法では, 10 回の音響モデル構築が必要となる.

一般には, チューニングパラメータは, 1 つ (今回では LF に相当) ではなく, パラメータ数と各パラメータの探索範囲分の音響モデルの構築が必要である. 開発した手法では, 音響モデルの再構築回数が少い.

5. まとめ

実環境音声認識のためのロボット聴覚システムを開発し, パラメータ最適化手法を示した. 開発手法の利点は, 各ステップで単語正解精度がクローズドデータに対して必ず改善すること, チューニングにおいて計算コストの大きい音響モデルの再構築回数を削減できることである. 音源分離パラメータと音響モデルの分離音声への適用を同時に行い, 8 % 改善できた. 認識精度が約 35 % と低かったのは今回用いた音響モデルは Monophone モデルであったためである. Triphone モデルを使用することで, 改善できる. 今後, より実環境に近い条件で, パラメータ最適化を行い, 経験的にチューニングしたシステムと認識精度を比較する必要がある.

本研究の一部は, 科研若手研究 (B), 科研費基盤研究 (S), 科研費特定領域研究 (情報爆発), 日仏研究協力, および京都大学グローバル COE プログラムの支援を受けた.

参考文献

[1] T.Takahashi, et al.,“Soft Missing-Feature Mask Generation for Simultaneous Speech Recognition System in Robots”, *Proc. Interspeech 2008*, pp.992–995, 2008.

[2] T.Takahashi, et al.,“Missing-Feature-Theory-based Robust Simultaneous Speech Recognition System with Non-clean Speech Acoustic Model”, *Proc. IEEE/RSJ IROS 2009*, pp.992–995, 2009.

[3] http://www.infoplosion.nii.ac.jp/info-plosion/index.php

[4] S.Yamamoto, et al., “Genetic Algorithm based Improvement of Robot’s Hearing Capabilities in Separating and Recognizing Simultaneous Speech Signals”, *Proc. IEA/AIE-2006*, LNAI 4031, pp.207-217, Springer-Verlag. Annecy, France, Jun. 2006.

[5] http://winnie.kuis.kyoto-u.ac.jp/HARK/