スペクトル推定を用いたマイク数以上の同時発話に対する音声認識

平澤 恭治 † 高橋 徹 ‡ 駒谷 和範 ‡ 尾形 哲也 ‡ 奥乃 博 ‡
† 京都大学 工学部情報学科 ‡ 京都大学大学院 情報学研究科 知能情報学専攻

1. はじめに

私たちは日常生活で生じる多数の音を2つの耳で聞き分けることができる. これまでコンピュータにも音を聞き分けさせるために様々な音源分離手法が提案されてきたが, 人の聴覚のように音源数 N がマイク数 M を上回る劣決定 (underdetermined) 状況に対応しているものはごく少数であった. 我々は図1に示した劣決定状況下での混合音声認識システムの開発を進めており, 本稿ではその音源分離部分について報告する.

劣決定状況下の音源分離には大きく分けて2種類の手法が存在する. 1つは時間周波数マスクを用いる手法 [1] であり, もう1つは混合行列の逆行列を用いる手法 [2] である. これらの手法では各時間周波数でパワーの強い音源 (支配的音源) を推定する必要があるが, 支配的音源の推定に失敗すると分離歪みが大きくなり, 音声認識精度が大きく低下する. したがってシステム全体の性能を高めるためには, 支配的音源の推定が重要な鍵となる.

上記2つの手法の仮定を考えると, 前者は各時間周波数に支配的音源が高々1つと仮定しているのに対し, 後者は高々マイク数 M と仮定している. そのため本稿では仮定の弱い後者の手法 (以下ベース手法と呼ぶ) を基本に, その分離結果を修正することを考える. この際調波構造はその周辺に比べてパワーが強いため優先的に支配的音源に含めることで, 支配的音源の推定精度を向上させる. 最後に3話者2マイクを用いた混合音声の分離認識実験を行い, 本手法の有効性を確認する.

2. ベース手法による劣決定音源分離

2.1 周波数領域での音声の混合

時間遅れのある混合を単純な乗算で表すため, 本稿では入力信号を短時間フーリエ変換 (STFT) して周波数領域で分離を行う. この時音声の混合は以下のように表せる.

$$\mathbf{x}(f,t) = \sum_{j=1}^{N} \mathbf{h}_j(f) s_j(f,t) = \mathbf{H}(f)\mathbf{s}(f,t) \qquad (1)$$

ここで t は時刻フレーム番号, f は周波数ビン番号, $\mathbf{x} = [x_1, x_2, ..., x_M]^T$ は各マイクでの観測音, $\mathbf{s} = [s_1, s_2, ..., s_N]^T$ は各音源の元信号, $\mathbf{h}_j = [h_{1j}, h_{2j}, ..., h_{Mj}]^T$ は音源 j から各マイクへの伝達関数, $\mathbf{H} = \{h_{ij}\}$ は伝達関数の行列 (混合行列) である. なお, 本稿では支配的音源の推定に焦点を当てるため, 混合行列は既知とする.

2.2 混合行列の逆行列を用いた音源分離

ベース手法では時間周波数ごとに独立に分離を行うため, 以下パラメータ f と t を省略する.

まず支配的音源の集合を $K = \{k_1, k_2, ..., k_M\}$ としたときの分離結果を考える. K に含まれない音源を無視すれば, 観測音は以下のように近似できる.

$$\mathbf{x} \approx \sum_{s=1}^{M} \mathbf{h}_{k_s} s_{k_s} = \mathbf{H}_K \mathbf{s}_K \qquad (2)$$

Simultaneous Speech Recognition of More Sources than Sensors using Spectrum Estimation: *Yasuharu Hirasawa, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

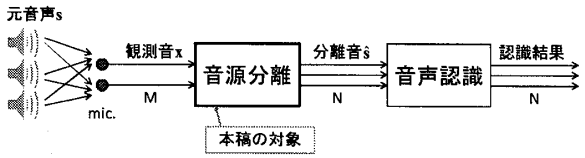

図1: 劣決定混合音声認識システム ($M < N$)

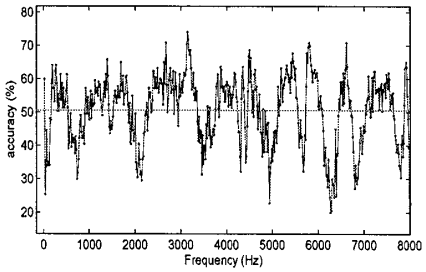

図2: ベース手法による支配的音源の推定精度の一例

ここで $\mathbf{H}_K = [\mathbf{h}_{k_1}, \mathbf{h}_{k_2}, ..., \mathbf{h}_{k_M}]$, $\mathbf{s}_K = [s_{k_1}, s_{k_2}, ..., s_{k_M}]^T$ である. $\mathbf{H}_K$ は正方行列なので, その逆行列を用いて以下のように分離できる.

$$\hat{\mathbf{s}}'_K = H_K^{-1}\mathbf{x} \qquad (3)$$

$$\hat{s}_i = 0 \qquad \forall i \notin K \qquad (4)$$

ここで $\hat{\mathbf{s}}_K := [\hat{s}_{k_1}, \hat{s}_{k_2}, ..., \hat{s}_{k_M}]^T$ であり, この時の分離結果は $\hat{\mathbf{s}}_K = [\hat{s}_1, \hat{s}_2, ..., \hat{s}_N]^T$ となる.

音声がパラメータ λ で表される同一のラプラス分布に従い, かつ互いに独立とすると, 同時対数分布は以下の式で表される. ベース手法では式 (1) の下でこの式を最大化することにより, 支配的音源を推定している.

$$\log p(s_1, s_2, ..., s_N) = -\lambda \sum_{k=1}^{N} |s_k| + C \qquad (5)$$

各要素が実数の場合, 式 (5) を最大化する $\mathbf{s}$ は, 支配的音源の集合 K を適切に選んだときの $\hat{\mathbf{s}}_K$ と等しくなる. K の取り方は ${}_N\mathrm{C}_M$ 通り存在するが, 各 K について $\hat{\mathbf{s}}_K$ を計算し, その中で式 (5) を最大化するものを支配的音源とすればよい.

本稿のように周波数領域での分離を考えると, 各要素が複素数になるため一般には $\hat{\mathbf{s}}_K$ が式 (5) を最大化しない. しかし実数と同様の手法を用いても結果に大きな違いはなく, 計算コストも削減できるという知見 [3] があるため, 本稿では上記のように分離を行った.

2.3 支配的音源の推定精度

ベース手法による支配的音源の推定では, 支配的音源の推定精度が元音源の分布に大きく依存し [2], 周波数ビンによって支配的音源の推定精度が大きく変化する. 図2に周波数と推定精度の関係を示す (支配的音源 M 個を全て正しく推定したときのみ正解とした). 横線は推定精度の平均値を表しているが, 推定精度の高い周波数領域

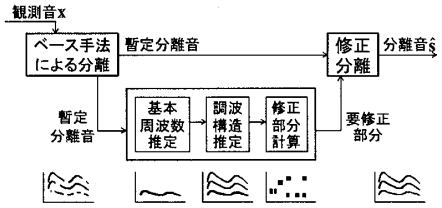

図 3: 本手法の流れ

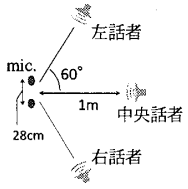

図 4: 話者位置

表 1: 実験条件

話者数 N, マイク数 M	N=3, M=2
サンプリング周波数	16kHz
インパルス応答	無響室にて録音
音源	JNAS 男女 200 文
STFT フレーム長	1024 点 (64ms)
STFT シフト幅	256 点 (16ms)
パラメータ	T=0.03, J=6, F=2
話者位置, 間隔	マイクから 1m, 60 度間隔
言語モデル	統計モデル, 語彙数 21k
音声特徴量	MFCC 25 次元 (12+Δ12+ΔPow)

と推定精度の低い周波数領域が存在することが分かる. これによりベース手法では図 5(上) のように, 特定の周波数付近 (図では 200Hz 付近) で分離に失敗してしまう.

3. 調波構造を用いた制約つき分離

図 3 に本手法の概要を示す. 本手法ではベース手法の出力を用いて調波構造を推定した後, 調波構造が必ず支配的音源に含まれるように要修正部分を再度分離する.

調波構造を用いる理由は (1) 調波構造はパワーの粗密がはっきりしており話者間での重なりが小さいこと, (2) 音声特徴量の Mel-Frequency Cepstrum Coefficient (MFCC) が調波構造のようなパワーの強い部分の分離結果に大きく依存するため高精度に分離する必要があること, (3) 調波構造は周波数方向での倍音構造を持つことを利用して, 推定精度の高い周波数の分離結果で推定精度の低い周波数の分離結果を修正できること, が挙げられる.

3.1 調波構造の推定

ベース手法の分離結果を用いて, 音源ごとに調波構造の推定を行う. 本手法ではケプストラム法を用いて, フレームごとに基本周波数の推定を行った. なお, 推定された基本周波数のケプストラム値がしきい値 T 以下であった場合には, そのフレーム内に調波構造はなく, 無音もしくは無声子音であると判断した.

次に推定した基本周波数から調波構造の形を推定する. 今回は基本周波数の第 J 倍音までを考え, そこを中心とした上下の F 周波数ビンに調波構造が存在するとした.

3.2 支配的音源の修正

ここではまず, 時間周波数ごとに修正が必要であるかを考える. これは, その時間周波数に存在する全ての調波構造が支配的音源とみなされているか (集合 P をその時間周波数に調波構造がある音源の集合とした時, K に P が含まれているか) で判断することができる.

次に調波構造が必ず支配的音源に含まれるよう, 必要部分を再度分離する. これは 2.2 節における分離の際に, $K \supset P$ という制約を加えたものであり, それ以外はベース手法と同様に行うことができる.

表 2: 実験結果 : 平均単語認識率 (%)

手法	左話者	中央話者	右話者
(a) ベース手法	64.9	59.7	69.6
(b) 本手法	**69.0**	**63.6**	**71.5**
(c) 最適解	82.3	82.5	85.0

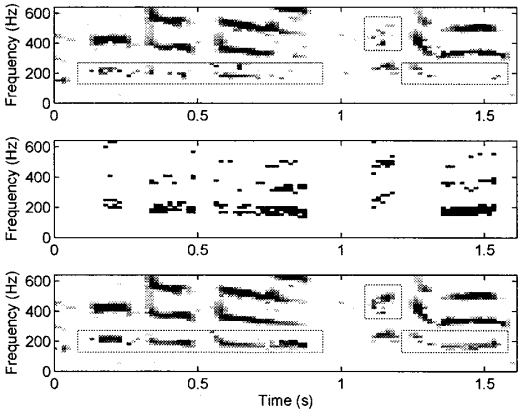

図 5: (上) ベース手法での分離結果 (中) 要修正部分 (下) 本手法での分離結果

4. 評価実験

本手法の効果を確認するために, インパルス応答を重畳した合成音を用いた評価実験を行った. 実験条件を表 1 に, インパルス応答測定時の音源とマイクの位置関係を図 4 に示す.

分離結果の調波構造部分を拡大したものを図 5 に示す. ベース手法 (上) では支配的音源の推定に失敗し, 200Hz 付近の基本周波数部分の多くが欠落していたものが, 要修正部分 (中) に修正を加えることで, 本手法 (下) の通り正しく分離できるようになっている.

また, 音声認識率により評価した結果を表 2 に示す. 参考のために, 支配的音源の正解を与えた場合の認識率を (c) 最適解として載せた. 本手法により音声認識率で 2-4 ポイントの向上が見られた.

5. おわりに

本稿では劣決定状況下における音声認識率向上のための音源分離手法を報告した. 劣決定状況の音源分離では支配的音源の推定が重要であるため, ベース手法の出力に対して調波構造を優先する修正を行うことを提案し, 音声認識率の改善を確認した. 今後, 調波構造を持たない無声子音に対する修正方法の検討や, 今回は既知とした混合行列の推定, 音声認識時の MFT マスクの導入などの発展が考えられる. また, 本研究の一部は, 科研費, GCOE, 日仏研究協力の支援を受けた.

参考文献

[1] S.Araki, H.Sawada, R.Murai, and S.Makino: 'Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors', *Signal Processing*, Vol 87 No.8, pp.1833-1847 (2007).

[2] Y.Li, S.Amari, A.Cichocki, D.W.C.Ho, and S.Xie: 'Underdetermined Blind Source Separation Based on Sparse Representation', *IEEE Transactions on Signal Processing*, Vol.54 No.2, pp.423-437 (2006).

[3] S.Winter, H.Sawada, and S.Makino: 'On real and complex valued L1-norm minimization for overcomplete blind source separation', *Proc. WASPAA 2005*, pp.86–89 (2005).