調波構造を用いたL1ノルム最小化に基づく劣決定音源分離手法の性能評価

○平澤恭治　高橋徹　尾形哲也　奥乃博
京都大学大学院 情報学研究科

Evaluation of methods for under-determined speech separation based on L1-norm minimization using harmonic structure

*Yasuharu HIRASAWA　Toru TAKAHASHI　Tetsuya OGATA　Hiroshi G. OKUNO
Graduate School of Informatics, Kyoto University

Abstract— In this paper, we propose a method for under-determined speech separation which uses new constraints exploiting harmonic structure. The conventional L1-norm minimization methods have an ability to handle many sound sources, however, acoustic feature values of their separation results are much distorted. Since harmonic structure has high power, which means it affects acoustic feature values very much, our method focuses on the harmonic structure and adds new constraints on it to preserve its structure. We carried out an experiment that simulates three to six simultaneous utterances using impulse responses recorded by two to four microphones in an anechoic chamber. The experiment reveals that our method can increase the speech recognition ratio by up to six point.

Key Words: sound source separation, under-determined, l1-norm minimization, harmonic structure

1. はじめに

近年多数の非産業用ロボットが開発され, その外見や動作など様々な面から注目を受けている [1]. そのような非産業用ロボットを単なる娯楽としてでなく, 実際に使用される実用的なロボットとするためには, 人間と意志疎通を行うための音声インタラクション機能が重要である. しかし現状のロボットは, 喋る機能 (音声合成) に比べて聴く機能 (音声認識) が大きく劣っている. これは喋ること (出力) に対して聴くこと (入力) は外界の影響を受けやすく, 実世界においては様々なノイズを含んだ混合音を観測し, その混合音から認識することが要求されるためである.

実環境には無数の音源が存在することから, 音源数がマイク数を上回る状況, つまり “劣決定状況” が頻繁に生じる. しかし従来の音源分離の研究は, 非劣決定状況を対象にしたものが多く, 多数の音が存在する環境には適用できなかった. そこで我々はそのような状況においても頑健な音声認識を実現するべく, 劣決定同時発話の分離・認識に関する研究を行なっている. 劣決定状況における音源分離手法は大きく分けて 2 種類あり, 1 つは各時間周波数領域にマスクをかける時間周波数マスク法 [2] で, もう 1 つは推定した混合行列により各時間周波数領域内の混合音を分離する L1 ノルム法 [3] である. ここで後者は音源数に対する仮定が弱く, 多数の音源を扱える一方, 分離結果に音声特徴歪みが生じやすいという問題点がある.

我々は音源数に対する仮定の弱い L1 ノルム法に着目し, さらに調波構造を用いた制約を加えることを提案する. これにより L1 ノルム法の問題である音声特徴の歪みが削減され, 音声認識しやすい分離音声を出力できる. 実験では 3 人から 6 人の話者の同時発話をシミュレートし, 話者数より少ない数の混合音を用いて分離を行なう. 本手法により MFCC 距離・音声認識率ともに改善することを示す.

2. 劣決定状況における音源分離手法

まず本稿で用いる変数の定義を行う. 以降ではマイク数を M, 話者数を N と書き, マイク番号を i, 話者番号を j とする. マイク i の観測音を $x_i(f,t)$, 話者 j の発話音声を $s_j(f,t)$ と表し, これらのベクトルをそれぞれ $\mathbf{x}(f,t)$, $\mathbf{s}(f,t)$ と書く. また, 話者 j からマイク i への時不変な伝達関数を $h_{ij}(f)$ とし, $h_{ij}(f)$ からなる混合行列を $\mathbf{H}(f)$ とする.

2·1 劣決定音源分離問題の定式化

本節では問題を正確に表現するために, 音声の混合過程を定式化する. なお, 本稿では音声のスパース性を向上させるために, 入力音声を短時間フーリエ変換 (STFT) を用いて時間周波数表現に変更し, 時間周波数領域上で音源分離を行う.

本稿では音声の混合は線形時不変な混合モデルであると仮定するので, 以下のように表現できる.

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

この式を用いて, 本稿が対象とする問題設定は以下のように表現される.

入力 M マイクでの観測音 $\mathbf{x}(t,f)$
出力 N 話者の分離音 $\hat{\mathbf{s}}(t,f)$
仮定 式 (1) による線形時不変な混合過程

なお以下では基本的に時間周波数領域ごとに独立に処理が行うため, 数式中の f や t を省略する.

2·2 音源分離手法への要求条件

本稿が対象とする音源分離手法に対しては, 以下の 2 つの要求が存在する.

1. 元音声に含まれる音声特徴量が劣化しにくいこと
ロボットが人間と音声インタラクションを行う際には, 通常音声認識を用いて音声を文字化する必要がある. 音声認識では入力音声から Mel-Frequency Cepstral Coefficient (MFCC) などの音声特徴量を計算し, その音声特徴量を用いて音声認識が行われる. 音声認識部分の入力には音源分離部分からの出力音声が用いられるため, 音源分離結果が元音声の音声特徴量を保存していることが望ましい.

2. 音源数が増加しても分離精度が低下しにくいこと
我々の最終目的は, 多数の音源がある状況でも複数の音を正しく認識することである. 実世界には多数の音源が存在しており, 音源数を仮定することはできない. 例えばパーティー時の会話のように実際に多数の話者が存在するときや, 高精度な分離結果を求める際には, 多くの音源を考慮して分離を行うことが好ましい. そのような点から, 音源数が増加しても分離精度が低下しないことが必要とされる.

2·3 調波構造による制約つき分離手法

2·3.1 L1 ノルム法

まず我々の提案手法のベースとなる L1 ノルム法 [3] について述べる. なお以下では各時間周波数領域で他の音源に比べてパワーが強い音源のことを “支配的音源” と定義する. L1 ノルム法は “各時間周波数領域に支配的音源は高々マイク数以下である” という仮定を置いた手法で, 他に伝達関数 $\mathbf{H}$ が既知である必要がある. この仮定の下では, 式 (1) を次のように近似することができる.

$$\mathbf{x} \approx \sum_{u=1}^{M} \mathbf{h}_{k_u} s_{k_u} = \mathbf{H}_K \mathbf{s}_K \tag{2}$$

ここで, k_u はその時間周波数内での u 番目の支配的音源のインデックスであり, その集合を K と呼ぶ. また $\mathbf{H}_K$ は元の混合行列 $\mathbf{H}$ の部分行列で $\mathbf{H}_K = [\mathbf{h}_{k_1}, \mathbf{h}_{k_2}, ..., \mathbf{h}_{k_M}]$ と定義し, $\mathbf{s}_K$ も同様に $\mathbf{s}_K = [s_{k_1}, s_{k_2}, ..., s_{k_M}]^T$ と定義する.

この近似を用いると, 式 (2) から次のように分離音を推定することができる.

$$\hat{\mathbf{s}}_K = \mathbf{H}_K^{-1}\mathbf{x}, \tag{3}$$

$$\hat{s}_i = 0 \qquad \forall i \notin K, \tag{4}$$

ここで $\hat{\mathbf{s}}_K = [\hat{s}_{k_1}, \hat{s}_{k_2}, ..., \hat{s}_{k_M}]^T$ であり, 分離結果は

$$\hat{\mathbf{s}}'_K = [\hat{s}_1, \hat{s}_2, ..., \hat{s}_N]^T \tag{5}$$

と表される. なお $\mathbf{H}_K$ が逆行列を持つためには, 各時間周波数領域はマイク数と同じ数の音源を支配的音源として持つ必要がある.

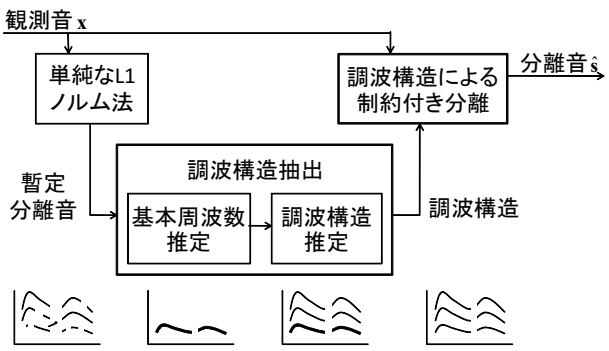

Fig.1 制約付き L1 ノルム法の概要

支配的音源集合 K の推定は通常以下のように行われる. まず音源のパワーに関する分布を独立かつ同一のラプラス分布と仮定する. 次に式 (1) を満たす中で分離結果の尤度がもっとも高くなる K を選ぶ. この時各要素が実数であれば, 分離結果 $\hat{\mathbf{s}}'_K$ の L1 ノルムを最小化する K が, 最尤な支配的音源の集合 K^{opt} であることが示せる.

ただし本稿では時間周波数領域における表現を考えているため, 各要素は複素数となり, 分離結果 $\hat{\mathbf{s}}'_{K^{opt}}$ が最尤解と一致しない. しかしこの場合でも, 上記の方法で求めた解は厳密解に近く, 計算時間が大幅に削減できるという知見 [4] があるため, 本稿ではこの方法により分離したものを L1 ノルム法の結果と考えることにする.

この L1 ノルム法の利点は, 各時間周波数に複数の支配的音源が存在することを許容するため, 音源数が増加した場合でも分離精度は大きく低下しない点である. しかし欠点として分離精度が音源のパワー分布や周波数帯域によって大きく変化するという問題があり [5][6], 音声特徴量に歪みが生じやすいという点がある.

2·3.2 調波構造による制約を加えた L1 ノルム法

ここでは我々が提案している, L1 ノルム法に調波構造による制約を加える手法の概略を説明する [5]. なお以下では 2.3.1 で述べた L1 ノルム法を “単純な L1 ノルム法” と呼び, 我々の提案している手法を “制約つき L1 ノルム法” と呼ぶことにする.

この手法の概要を Fig.1 に示す. これは大きく分けて 2 段階に分解でき, 第 1 段階は “単純な L1 ノルム法の分離結果から調波構造を推定する段階” で, 第 2 段階は “推定した調波構造を用いて再度分離を行う段階” である. 2.3.1 でも述べたように単純な L1 ノルム法は音声特徴量に歪みが生じやすい. 調波構造のようなパワーの強い時間周波数は音声特徴量への寄与が大きいため, 我々は調波構造を正しく分類することで音声特徴歪みを削減し, 音声認識率を向上できると考えている. 以下では各段階について詳細に説明する.

第 1 段階では, まず初めに単純な L1 ノルム法を用いて暫定的な分離音を得る. 次に暫定的な分離音に対して Fig.2 に示される調波構造抽出を行う. この調波構造抽出は 2 フェーズに分かれており, 第 1 フェーズでケプストラム法によって各フレームの基本周波数を推定し, 調波構造が存在すると推定されたフレームについては

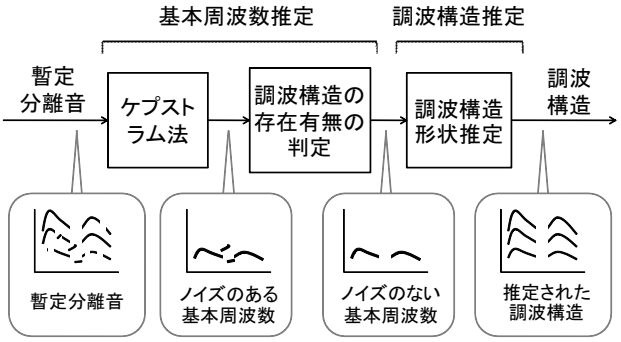

Fig.2 調波構造抽出の概要

第 2 フェーズで調波構造の倍音性を利用して基本周波数から調波構造を推定する．ここで暫定分離音から直接調波構造を抽出するのでなく，一旦基本周波数を介することで，暫定分離音にスペクトルの欠落や漏れノイズが生じていても頑健な調波構造抽出が行える．

続いて第 2 段階では，推定された調波構造を用いた制約条件を追加した上で，再度 L1 ノルム法による分離を行う．この制約条件は“その時間周波数に調波構造を持つ音源は，必ず支配的音源の集合 K に含まれる”といったもので，各時間周波数ごとに異なる制約となる．我々は支配的音源を各時間周波数領域で他の音源に比べてパワーが強い音源のこと，と定義していた．調波構造も同様にパワーの強い部分に関する構造であるため，調波構造を持つ音源が支配的音源の 1 つになるという制約は自然に導かれる．なお 2.3.1 で述べたように L1 ノルム法においては支配的音源集合の要素数はマイク数と等しいので，調波構造を持つ音源の数がマイク数を上回った場合にはこの制約を満たすことができない．その場合には上記の制約ではなく，調波構造を持たない音源は支配的音源集合 K に含まない，という制約を用いる．

3. 実験

本章では，2 章で行った議論の結果を確認するため，無響室で測定したインパルス応答を用いて劣決定状況をシミュレートし，単純な L1 ノルム法と制約つき L1 ノルム法を用いて分離した結果を確認する．

実験では混合行列 $\mathbf{H}$ を既知とし，マイク数 M を 2 から 4 まで，話者数 N を 3 から 6 まで変化させ，そのうち劣決定条件 $(M < N)$ を満たす 9 通りを使用した．詳細な実験条件を Table 1 に，話者とマイクの配置を Fig.3 に示す．

まず，1 つ目の実験として本手法の有効性を確認するために，単純な L1 ノルム法と制約付き L1 ノルム法を用いて分離を行い，MFCC 距離と音声認識率の 2 通りの尺度で評価を行った．なお MFCC 距離は以下の式を用いて計算した．

$$\text{MFCC 距離} = \sum_{t=1}^{T} ||\text{mfcc}(\mathbf{s}_j(t)) - \text{mfcc}(\hat{\mathbf{s}}_j(t))||^2 \quad (6)$$

ここで T は時間フレーム数で，$\text{mfcc}(\mathbf{s}_j(t))$ と $\text{mfcc}(\hat{\mathbf{s}}_j(t))$ はそれぞれ，時間フレーム t での音源 $\mathbf{s}_j$ とその推定値 $\hat{\mathbf{s}}_j$ の mfcc 特徴量ベクトルである．

Table 1 実験条件

話者数, マイク数	3-6 話者, 2-4 マイク
sampling 周波数	16 kHz
インパルス応答	無響室で測定
話者間隔・距離	30 度・マイクから 1m
音声データ	JNAS 男女 200 文
STFT フレーム長	1024 点 (64ms)
STFT シフト幅	256 点 (16ms)
音声認識器	julius 3.5 fast
音響モデル	PTM トライフォン 3 状態 HMM
言語モデル	統計モデル/ 2 万単語
使用特徴量	MFCC 12+ΔMFCC 12+ΔPow
分析窓フレーム長	400 点 (25ms)
分析窓シフト幅	160 点 (10ms)

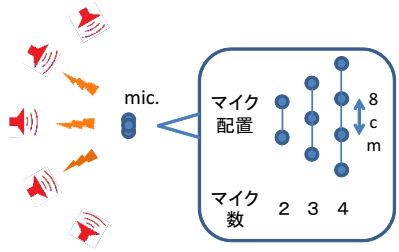

Fig.3 話者とマイクの配置

また，2 つ目の実験として，各時間フレーム内にある時間周波数領域の支配的音源の数を既知として単純な L1 ノルム法を適用した．音源数の推定には混合前の音声データを使用し，様々な支配的音源の組み合わせに対する分離結果のうち，元音源とのスペクトル距離がもっとも短いものを適切な音源数とした．また，推定された支配的音源数がマイク数を下回る場合には，マイクの観測結果の一部のみを用いて分離を行った．

3·1 実験結果

まず評価尺度として式 (6) で表される MFCC 距離を用いた結果を Fig.4 に，音声認識率を用いた結果を Fig.5 に示す．ここで図の横軸はマイク数と話者数の組み合わせを示している．話者数が多いので，手法ごとに MFCC 距離または音声認識率の値が最大の話者と最小の話者の数値のみをプロットした．

図を見ると，まずどの手法でも話者数が増加するにつれて全体的な傾向として MFCC 距離，音声認識率ともに悪化していることが分かる．これは話者数が増加することにより混合音のスパース性が失われるためだと考えられる．

次に同一話者数でマイク数が変化するとどうなるかを確認すると，基本的にマイク数が増加するにつれて MFCC 距離，音声認識率ともに向上している．これはマイク数が増加することにより，入力される情報が増えたためと考えられる．しかし単純な L1 ノルム法と制約

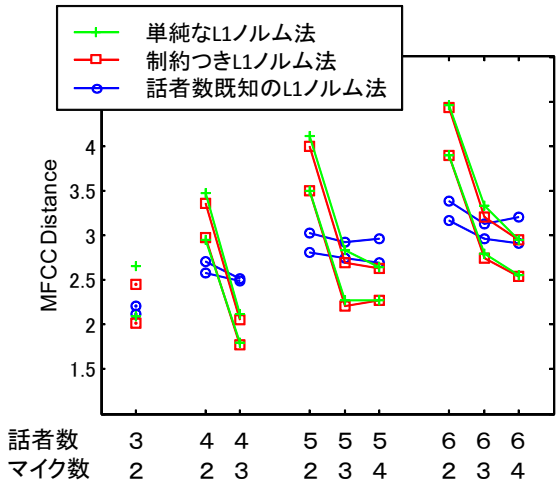

Fig.4 MFCC 距離の変化

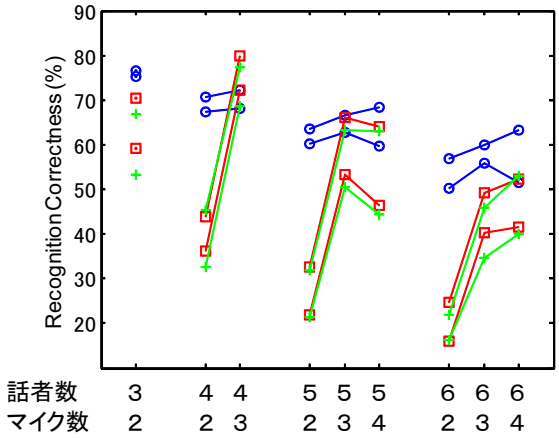

Fig.5 音声認識率の変化 (凡例は Fig.4 と同様)

付き L1 ノルム法による分離結果では, 5 話者 4 マイクの音声認識率が 5 話者 3 マイクの音声認識率よりも低下している. この原因については, 3.2.2 で考察を行う.

3·2 考察

3·2.1 単純な L1 ノルム法と制約つき L1 ノルム法

Fig.4 と Fig.5 を見ると, 単純な L1 ノルム法 (緑線:＋) よりも制約つき L1 ノルム法 (赤線:□) の方がほとんどの条件で MFCC 距離, 音声認識率ともに性能が良くなっており, 音声認識率においては最大で 6 ポイント, 平均で 2 ポイントの性能改善がみられた. これは音声の特徴である調波構造を正しく推定することで, 分離結果の音声特徴量を保ち, 音声認識しやすい分離が行われたのだと考えることができる.

3·2.2 5 話者 4 マイクの際に音声認識率が低下した点

実験結果で述べたように, 5 話者 4 マイクの条件においては, マイク数が増えたにも関わらず音声認識率が低下している. この原因を調べるために分離結果のスペクトログラムを確認すると, 支配的音源の推定が正しいにも関わらず他話者の漏れノイズが発生しているのが確認できた. これは, L1 ノルム法では各時間周波数領域内の支配的音源の数がマイク数と等しい必要があるが, 実際には音声のパワー分布はスパースなため, そこまで多くの支配的音源が存在しないことに由来する. また, その場合でも理想環境であれば式 (3) の計算により “パワーがさほど強くない支配的音源” も正しく分離されることが期待されるが, 実際には STFT 時のわずかな近似誤差などの影響で, 他の支配的音源の漏れノイズが生じているのだと考えられる.

そこで 2 つ目の実験として, 支配的音源の数を既知として L1 ノルム法を適用した結果を, Fig.4 と Fig.5 に示す (青線:○). 支配的音源数の決定に一部失敗しているため, 最低認識率は一部の実験条件で低下しているが, 最高認識率の方ではマイク数が増加するにつれて MFCC 距離は低下しており, 音声認識率は向上していくという結果が得られた.

この実験結果から分かるように, L1 ノルム最小化に基づく劣決定音源分離手法を用いて多数のマイクの観測音を分離する際には, 2.3.1 で述べたように時間周波数領域における音源数をマイク数で固定するのでなく, 観測音から推定するなどによって動的に変化させることが必要である.

4. おわりに

本稿では L1 ノルム最小化に基づく音源分離手法に対し, 音声特徴量を保持するために調波構造を用いた制約を加えることを提案した. 実際に制約つき L1 ノルム法と制約なし L1 ノルム法を実装し, 調波構造による制約を加えた手法が様々な話者数とマイク数の環境において音声認識の面で優れていることを確認した. また, 実験 2 では話者数既知として分離を行い, 各時間周波数領域での話者数を推定することで話者数やマイク数が多い場合にも正しく分離できる可能性を示した.

今後の展開としては各時間周波数領域の音源数を推定する手法や, 調波構造以外の制約を追加して分離結果の特徴量歪みを削減する手法の検討がある. また, L1 ノルム法全体の問題として残る, 混合行列の推定手法に関する研究も重要であると考える.

謝辞 本研究の一部は, 科研費基盤研究 (S), 特定領域, JST SICP (日仏研究交流), GCOE の支援を受けた.

参考文献

[1] M. Hirose and K. Ogawa. Honda humanoid robots development. *Phil. Trans. R. Soc. A*, Vol. 365, No. 1850, pp. 11–19, 2007.
[2] O. Yılmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. on Signal Processing*, Vol. 52, No. 7, pp. 1830–1847, 2004.
[3] P. Bofill and M. Zibulevsky. Underdetermined blind source separation using sparse representations. *Signal processing*, Vol. 81, No. 11, pp. 2353–2362, 2001.
[4] S. Winter *et al.* On real and complex valued L1-norm minimization for overcomplete blind source separation. In *Proc. of WASPAA*, pp. 86–89, 2005.
[5] Y. Hirasawa *et al.* Exploiting Harmonic Structures to Improve Separating Simultaneous Speech in Under-Determined Conditions. In *Proc. of IROS*, 2010. *to appear.*
[6] Y. Li, S. Amari, A. Cichocki, D.W.C. Ho, and S. Xie. Underdetermined blind source separation based on sparse representation. *IEEE Trans. on Signal Processing*, Vol. 54, No. 2, pp. 423–437, 2006.