E-019

環境に依存しない音声認識のための音源分離法
Semi-Source Separation for Environment-independent Speech Recognition

迫川 海翔† 崔 龍雲†
Kaito Sakogawa Yongwoon Choi
†創価大学大学院 理工学研究科
E-mail：e20m5308@soka-u.jp

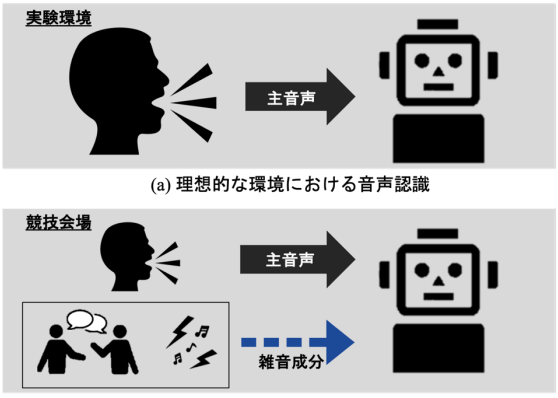
1. はじめに

音声認識技術の発展により，ロボットやデバイスが，人からの音声による指示や要望を，正確に認識することができるようになった．このように，音声を主体とするコミュニケーション手段の利用やサービス環境が発展している．中でも，家庭空間において，Apple 社の Siri や Amazon 社の Alexa，Google 社の Google アシスタントといった音声アシスタントデバイスが普及している．こういった背景から，人から人への指示と同様に，人が発した音声から指示を推測可能になったことが大きな要因である．現状の音声認識技術は，図 1 (a)のように主音声のみが容易に取得できる理想的な環境に近い場合，人の音声による指示・命令の音声認識率は 99%を超えている[1]．しかしながら，図 1 (b)のように主音声以外の音も聞こえ，尚且つ事前情報のない家庭環境等においての音声認識率はその限りではない．このため，複数の音が混ざり合った音声信号から，主音声を抽出する需要が高まっている．図 1(b)のような環境は，家庭環境に多く見られる．

実際に複数の音が混ざり合う環境での音声認識を想定した実例として，本研究室が出場している RoboCup@home[2]という大会が存在する．この大会では，家庭環境を模した部屋が競技会場として用いられ，「音声での指示を正しく実行できるか」というタスクをロボット判断して行われる．2018 年の本大会において，研究室でのロボットの動作確認の際には認識することのできた命令文が，実際の大会会場では認識できなかったという事例が発生した．これは，音声認識ソフトの問題ではなく，主音声と雑音成分が混ざり合ってしまったことが原因である．この事例のような場合において，音源分離という手法が用いられる．

図 1(b) のような環境に対する音源分離手法として，広く用いられているのが独立成分分析（Independent Component Analysis, ICA）[3] である．これは，図 2(a)のように，音源数 N＜マイク数 M の場合，高い分離性能が示されている．しかし図 2(b)のような，N ＞ M の場合においては ICA の適用が困難になる．これに対し，音源数に依存しない音源分離手法として非負値行列因数分解（Non-negative Matrix Factorization, NMF）[4] がある．NMF は非負の値の特徴を持つ集合に分化させることができる手法である．この手法は単純な音の集合を容易に分離することができるが，人の声が混ざり合ったデータに対しての分離能力は不十分である．分離性能を向上させるためには，NMF で分離した音がどの人の声に該当するのかを分類する必要がある．

そこで本研究では，単一マイクで収音した複数の人の声が混ざり合ったデータの中から，主音声を抽出するために，



(a) 理想的な環境における音声認識
(b) 実環境における音声認識

図 1： 理想環境と実環境の違い

(a) ICAの音源分離が可能な場合 (b) ICAの音源分離が不可の場合

図 2: マイクの個数と音源数による ICA の可否

NMF とクラスタリング手法を用い音源分離の実験を行い，雑音環境下での音声認識率向上を目的とする．本稿では，入力データから主音声を分離、推定するシステム概要とその手法について述べ，検証実験の評価を行う．

2. 関連研究と研究目的

2.1 関連研究

音源分離の中でも，ブラインド音源分離[5]は，事前情報を必要としない音源分離できるため需要が高く，教師なしの音源分離手法では 1 章で述べた ICA が代表的である．

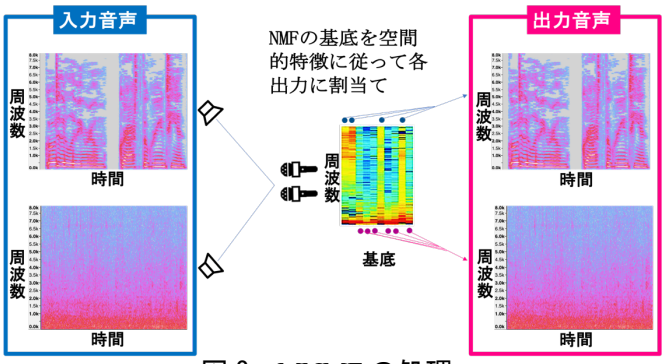


図 3：MNMF の処理

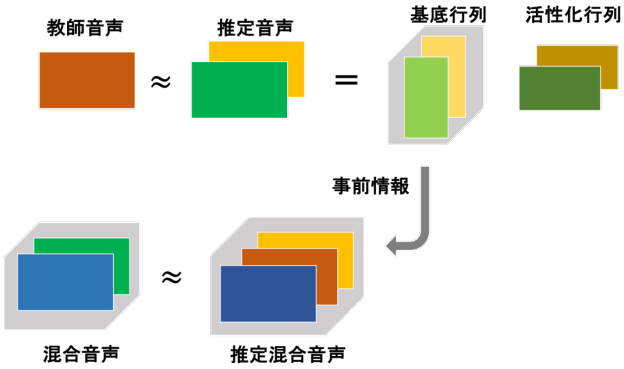


図 4：SSNMF の処理

また，澤田ら[6]が提案した多チャネル NMF（Multi_channel NMF, MNMF）手法では，図 3 のような流れで処理を行い，ICA よりも少ないマイクでの音源分離を実現し，NMF よりも高精度な音源分離を実現している．しかしながら，これらの手法は，複数のマイクが必要となることや，計算コストが高いという課題も存在している．さらに，NMF を拡張した手法のなかで，関ら[7] が提案した半教師あり NMF（Semi-Supervised NMF, SSNMF）では，図 4 のような流れで処理を行い，単一マイクの情報不足に対し，種音源を事前録音し，その基底部分を教師として分離を行っている．この手法では，複数個のマイクを用いる手法と比べ，分離性能はやや劣るが，NMF のみでの分離性能をはるかに上回る結果となった．しかし，適応分野が楽器の音にとどまっており，人間の音声のような複雑な音への適用の場合の精度が不明瞭である．

2.2 研究課題

本研究では，単一マイクから得られる情報のみを用いつつ，情報不足を解決し，複数のマイクを用いる手法と同等程度の分離精度を実現することが最も重要な課題となる．関連研究[6]では，複数のマイクを用いた手法であったが，計算コストが高くなるという課題があり，リアルタイムなコミュニケーションが難しくなってしまう．

そこで，話者の音声をあらかじめ録音し，そのデータを元に NMF を行う関連研究[7]の SSNMF を人間の音声に応用行う．しかし，現状の SSNMF の研究では，複数の楽器の混ざり合った音源を，楽器ごとに分離していく手法であり，人の音声のように複雑な音源への適用は未実装である．楽器と人の音声では分離する難易度が異なる．人間の音声は楽器の音とは違い，話す人によって音声の周波数の変化幅が異なり，一音の発音であっても，単純な 1 つの周波数ではなく，複数の周波数を含有している複雑な音である．そのため，どの音が主音声の音声要素なのかが不明であるという問題がある．

この問題に対して，SSNMF の適用後に，どの音がどの人の音声要素なのかを，クラスタリングすることで，複雑な人間の音声にも，SSNMF を適用できることが期待できる．また，SSNMF のみの処理では雑音成分として主音声要素が処理されてしまうが，クラスタリングすることにより，主音声要素が雑音成分として処理されることを防ぐため，音源分離の精度の向上が見込まれる．

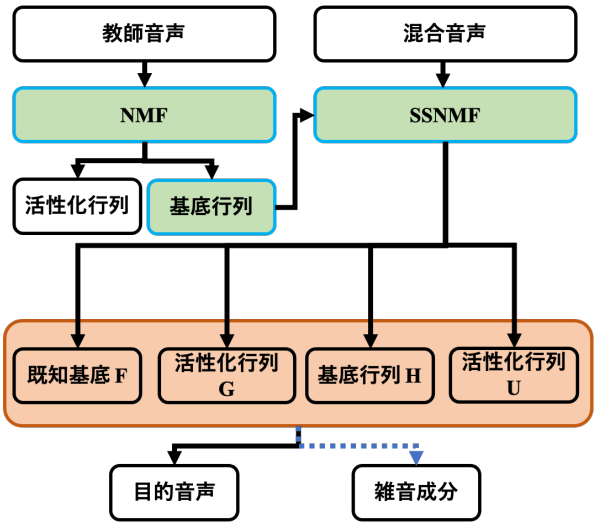


図 5：本手法の処理の流れ

3. 提案手法

本研究では，関連研究[7] の音源分離手法に，k-means 法によるクラスタリングを導入し，環境に依存しない音声認識のための音源分離手法を提案する．本手法の処理の流れは図 5 に示すように，混合音声に対して SSNMF を適用する．その際に使用する教師は，事前に録音した教師音声の基底情報を採用する．その後，クラスタリングにより，分離精度をさらに高める．本手法の処理を以下にそれぞれ説明する．

3.1 教師データの選定

まず，SSNMF に必要な教師データの選定について説明する．SSNMF では精度向上のために話者の音声を事前に録音する．この時の音声情報の過不足により，その後の分離精度が変化する．そのため，話者の音声の特徴をなるべく網羅した音声を録音する必要がある．そこで，今回は教師データを複数個用意し，どの教師データによってどの程度，分離結果に差が生じるのかを確認する．

表 1：命令文の例

番号	命令文
1	Greet Patricia at the bed and ask her to leave.
2	Take the apple from the kitchen to the counter.
3	Bring the cup to the dining table.
4	Follow Skyler from the desk to the kitchen.

表 2：混合する組合わせ

番号	組合せ
1	男性音声+女性音声
2	女性音声+女性音声
3	男性音声+男性音声
4	男性音声+環境ノイズ
5	女性音声+環境ノイズ

表 3：実験結果（一部抜粋）

処理の有無	混合する割合		
	① 1：1	② 3：2	③ 2：1
分離前の認識率	88%	67%	34%
分離後の認識率	88%	77%	56%

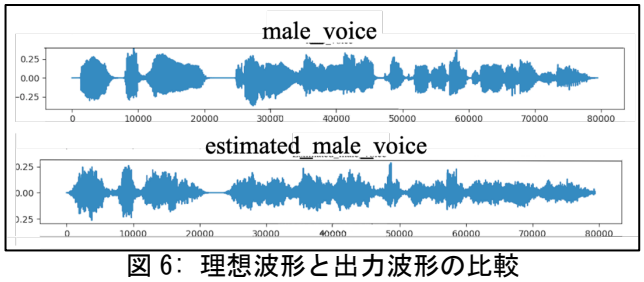


図 6：理想波形と出力波形の比較

3.2 SSNMF を用いた音源分離

本手法の処理の流れを図 5 に示す. 教師となる音声データに対し, NMF を適用する. 本研究では, NMF によって出力された基底行列と活性化行列のうち, 基底行列を教師とする. 一般的な NMF では, 入力となる音声の混ざり合った音声信号 Y, 基底行列 H, 活性化行列 U を用いて(1)式が成り立つ.

$$Y \approx H \cdot U \quad (1)$$

(1) 式に対し, SSNMF では, 音声の混ざり合った音声データの中で, 主音声の基底情報は既知情報として扱う. 既知の主音声の基底情報 F, 未知の主音声の活性行列 G, 雑音成分の基底行列 Hと活性化行列 U を用いて(2)式が成り立つ.

$$Y \approx F \cdot G + H \cdot U \quad (2)$$

(1) 式を用いて求めた, 主音声の基底行列を (2)式の F として処理を進める.

3.3 クラスタリング

人の声は, 声を構成する周波数が複数個以上あるため, NMF の適用だけでは適切な分離を行うことができない. したがって, NMF を基調とした音源分離を行う本手法のでは, 「各々の基底ベクトルがどの声に対応するものなのかをラベル付け」する分類問題と同義である. ただし, クラスタリングする音声データは情報量が多く, 冗長過ぎるため, 特徴量を限定して抽出する必要がある. 抽出する特徴量として M.Spirtz ら[8]によって提案されたメル尺度の特徴量を用いる. この尺度は, 人間の音高知覚が考慮された尺度である. 声の混ざり合った混合信号から音声ごとにクラスタリングするために, ソースフィルタ理論[9]を用いる. また, この理論では, 人の声の混ざり合った混合信号に適用すると, この混合信号を NMF で分離して得られる基底行列$\mathbb{H}$, ソース $\mathbb{E}$, フィルタ$\mathbb{F}$ との積で表すことができる. 特徴量をメル周波数に圧縮するためにメル・フィルタバンク行列 $\mathbb{R}$と基底ベクトルとの行列積を計算し, 対数をとったものは以下の式(3)で表される.

$$log(R_{mel,k}H_{k,m}^2) \approx log(R_{mel,k}E_{k,m}^2) + log(R_{mel,k}F_{k,m}^2) \quad (3)$$

式(3)の右辺第一項がソース項であり，右辺第二項がフィルタ項である．式(3)のソース項は人によって変化しないと仮定しているため, この値は右辺第二項に依存したメル尺度の次元数を持つ特徴量空間に人ごとの偏りをもって分布することが期待される. この人ごとの偏りをクラスタリングする手法として, 次元数を更に減らすため式(1)の値をメル周波数ケプストラム周波数(MFCC) に変換して、クラスタリング手法に適用する．具体的には, 離散コサイン変換係数をとって低次元のものだけ抜き出し，k-means 法でラベル付けを行う.

また，分割するクラスタリング数にも工夫が必要である. 本研究では，主音声と雑音成分という 2 種類の混合を想定しているが，クラスタリング数は必ずしも 2 個が最適とは限らない．教師となる基底情報と似た基底を持つ音は，それが雑音成分であったとしても，主音声としてクラスタリングされる．そこで，エルボー法[10]を用いることで，入力となる音声が混ざり合った音声データに対して，最適なクラスタ数を決定する．これにより，その入力音源に適したクラスタ数でクラスタリングすることができ，主音声を高精度に再現することができる.

4. 評価実験

4.1 実験概要

本実験では, SSNMF とクラスタリングを併用した音源分離の有用性を評価する. 実験に使用する命令は, 2019 年の RoboCup@Home で使用された Command Generator を使用し作成する. ランダムで表 1 のような 100 個の命令文を生成し, 100 個の命令文を男女 2 人ずつ計 4 人の話者でそれぞれ

発話し, 400 個のデータとした. 混合する組み合わせは表 2 の通りである. それぞれの組み合わせでの, 2 つの音声の音量比率は 2:1, 3:2, 1:1 とした.

また, 評価手法としては, 認識対象の命令文の正解を用意し, 出力文を音声認識にかけ, その結果を比較し認識率を算出する. 音声認識は GoogleAPI を用いる. また, 合わせて波形の比較も行う.

4.2 実験結果

認識対象の命令文に対して, 手法適用後の認識率を算出した結果の男性と女性の混合の一部を表 3 に示す. 表 3①では, 本手法適用前での音声認識率は約 34%であったのに対し, 本手法適用後は 56%に向上した. 表 3②では, 処理前が 67%に対し, 処理後が 77%までの向上した. 表 3③では, 処理の前後での認識率がどちらも 88%と変化がなかった. これは, 雑音成分の音量が主音声の半分であるため, 誤認識につながらない程度の無視できる範囲内の雑音であったと考える.

また, 表 3①では, 文章に含まれる (move, bedroom), (count, people)の 2 動作 2 対象のすべての情報が欠落してしまっていたが, 処理後は (count, people)の 1 動作 1 対象の認識に成功した. これにより, 音声認識の後の命令理解の処理につなげていくことができると考えられる. 表 3②の手法適用前での認識結果では, 2 動作 2 対象のうち, (count, people)の 1 動作 1 対象しか認識できていなかったが, 手法適用後では, 命令文の中の 2 動作 2 対象をすべて認識することができていた.

図 6 は表 3①の場合の波形比較である. 元波形である男性音声に含まれていないノイズが, 推定された音声には一様にかけ合わさってしまっているが, 音の出るタイミングの再現はできている. これらの実験結果より, SSNMF とクラスタリングを併用することで, 適用する前と比較すると, 音声認識率は向上した. また, 文章の意味を理解する上で重要な動作とその対象に関しても, 本手法を適用することで, 取得することができた. これらのことから, 本手法は高騒音環境においての音声認識において有用であると考える.

5. まとめ

本稿では, 話者の音声を事前に録音し, 話者の音声特徴を教師として SSNMF を行い, クラスタリング手法を併用することで, 環境に依存しない音声認識が可能になると仮説を立てた. 提案手法では, 関連研究[7]を発展させ, 人間の音声分野にまで拡張を行った. 実験では, 提案手法が関連研究と比較して, マイクの個数が少なくなり, それにより精度はわずかに低下したが, 計算コストを減少させることができた.

今回行った実験に使用した, 入力となる複数の音声の混ざり合った音声データは, 実際の人の音声ではなく, 音声合成した音声を男女 4 人分作成し使用した. そのため, 今回の手法が実環境においても有用かどうか明らかでない. そのため, 実際の人の声を用い, 実験を行なっていき, リアルタイムでの処理の実装を行おうと考えている.

参考文献

[1] Gene Munster. “Annual Smart Speaker IQ Test”. LOUPVENTURES.2018-12-20. https://loupventures.com/annual-smart-speaker-iq-test/, (参照 2021-6-15)
[2] 杉浦孔明,: “ロボカップ@ホーム-人と共存するロボットのベンチマークテスト-”, 人工知能学会誌, Vol.31, No.2, pp.230-236, 2016
[3] Jutten, C. and Herault, J.: Separation of sources, Part I, *Signal Processing,* Vol.24, No.1, pp.1-10, 1991
[4] D.D.Lee and H.S.Seung: Learning the parts of objects with non negative matrix factorization, Nature, 401, pp.788-791, 1999
[5] A. Cichocki and S. Amari: *Adaptive Blind Signal and Image Processing, Learning Algorithm and Applications*, John Wiley & Sons, Ltd, 2002
[6] 澤田宏,: “非負値行列因数分解 NMF の多チャンネル拡張”, 信号処理シンポジウム講演論文集, Vol.33, No.27, pp.396-401, 2012
[7] 関翔悟,: “ケプストラム距離正則化を用いた半教師ありステレオチャンネル楽曲音源分離”, 情報処理学会研究報告, Vol.2017-MUS-115, No.8, pp.1-6, 2017
[8] .Spiertz, V.Gnann,: “Source-filter based clustering for manaural blind source separation”, Conference on digital audio effects, Vol.12, No.9, pp.1-9, 2009
[9] 小林隆夫,: “トピックス 12 ソース・フィルタ理論”, 日本音響学会誌, Vol.57, No.1, p.65-, 2000.
[10] Robert L. Thorndike: Who belongs in the family?; *Psychometrika*, Vol.18, No.4, pp.267-276, 1953