様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　　８　日現在

機関番号：　１４３０１
研究種目：　挑戦的萌芽研究
研究期間：　2015～2016
課題番号：　１５Ｋ１２０６３

研究課題名（和文）混合音に対する複数同時発話認識のための統一的ベイズアプローチ

研究課題名（英文）A Unified Bayesian Approach to Simultaneous Speech Recognition for Mixture Signals

研究代表者

吉井　和佳（Yoshii, Kazuyoshi）

京都大学・情報学研究科・講師

研究者番号：２０５１０００１

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：本研究では、音源分離を確率的に統合した同時発話音声認識を行う手法を提案した。音源分離により復元される音声信号には不確実性が存在するため、音声信号の事後分布を考慮することで音声認識との統合を行う。これにより、復元すべき音声を一意に定めることなく混合音から直接認識結果を得ることが可能となった。また、音の重畳過程と音源モデルを内包する統合モデルにより、高精度な音源分離を行う手法を考案した。具体的には、重畳過程・音源モデルに対して、混合モデル（LDA）および因子モデル（NMF）のそれぞれのモデル化を行うことで、各モデルの音源分離性能を比較評価した。

研究成果の概要（英文）：We proposed a method that can simultaneously recognize multiple utterances by using a probabilictic model of source separation. Since there is uncertainty about source signals, we combined speech recognition with source separation by considering the posterior distributin of the source signals. This enabled us to obtain recognition results directly from mixture signals without uniquely determining the source signals. In addition, we proposed a source separation method based on an integrated model involving a source model and a superimposition model. Each model is represented as a mixture (LDA) or factor model (NMF) and the performance of each combination was evaluated.

研究分野：　統計的音響信号処理

キーワード：　音源分離　音声認識　確率モデル　ベイズモデル　MCMC

１．研究開始当初の背景

近年, 研究室環境下においては, Deep Neural Network (DNN) の台頭により，共通データセットに対する音声認識精度は飛躍的に向上している．一方で，実環境下における音声認識には多くの課題が残されている．音声認識精度を向上させるには，あらかじめ雑音抑圧や残響除去などの前処理を施すことで，できる限りクリーンな音声信号を復元するアプローチが主流である．

しかし，雑音抑圧や残響除去は数学的に不良設定問題であり，復元すべきクリーンな音声信号には不確実性が存在する．復元すべき音声信号を一意に定め，そのあと音声認識を独立して行う従来のアプローチでは，復元した音声信号が音声認識にとって最適である保証はなかった．また，ほとんど全ての音声認識の研究は入力として単独発話を想定しており，イベント会場などで複数の話者が同時に発話しているような実環境下での音声信号はそのままでは扱えなかった．

従来は，まず混合音 $\boldsymbol{X}$ に対して雑音抑圧・残響除去・音源分離などの信号処理を適用し，尤もらしい単独発話音声 $\boldsymbol{S}^*$ をいったん推定してから，得られた $\boldsymbol{S}^*$ に対して尤もらしい単語列 $\boldsymbol{Z}^*$ を推定していた．すなわち，数学的には，独立した２つの最適化問題

$$\boldsymbol{S}^* = \underset{S}{\operatorname{argmax}}\, p(\boldsymbol{S}|\boldsymbol{X})$$

$$\boldsymbol{Z}^* = \underset{\boldsymbol{Z}}{\operatorname{argmax}}\, p(\boldsymbol{Z}|\boldsymbol{S}^*)$$

を順番に実行していた．このとき，前段は不良設定問題であるにもかかわらず，単独発話音声 $\boldsymbol{S}^*$ を強引に一意に決めてしまうことで，後段の音声認識の精度に原理的な限界が存在するという問題があった．このような状況を受けて，(1) 音響信号処理と音声認識を統合する技術の確立と (2) 音源分離技術自体の改善に合わせて取り組む必要があった．

マイクロホンアレイを用いたマルチチャネル音源分離のためには，これまでに多くの手法が提案されてきた．広く用いられている手法のうちの一つに，独立成分分析 (ICA) がある．ICA は各音源の統計的な独立性を仮定することにより混合行列の逆行列である分離行列を推定し，分離を行う．この ICA をもとに独立ベクトル分析 (IVA) や FastICA などのさまざまな手法が提案されているが，これらの手法は共通してマイク数が音源数より少ない劣決定条件では分離できないという問題点がある．

これに対して，劣決定条件でも分離が可能な手法として時間・周波数クラスタリングに基づく音源分離法が着目されている．このアプローチでは，各音源スペクトログラムが時間・周波数領域でスパースであると仮定することで，混合音スペクトログラムの各時間・周波数ビンにおける観測はそれぞれいずれか一つの音源成分が直接観測されたものであるとみなす．これにより，音の重畳過程が混合モデルにより表される．この混合モデルを推定するため，マイク間の音の位相差とパワー差を特徴量として用いた混合ワトソン分布のクラスタリングや各マイクでの音の位相とパワーを特徴量として用いた混合ガウス分布のクラスタリングによる分離法が提案されている．これらのクラスタリングはそれぞれの時間周波数ビンでそれぞれ独立に行われるためパーミュテーション問題が発生するが，大塚らは潜在的ディリクレ配分法 (LDA) を用いて各時間周波数ビンを各音源に割り当てたあと，それらの音源をさらにいずれかの方向に割り当てることでこのパーミュテーション問題を解決している．

また，単チャネル音源分離のための手法として非負値行列因子分解 (NMF) が広く使われている．単チャネル音源分離ではマイク間の位相差などの空間的な情報を用いることができないため, NMF では音源の低ランク性を仮定することで音源を構成する因子を推定し，音源分離を行う．具体的には，音源が基底とアクティベーションの二つの因子から成るとし，観測のパワースペクトログラムを基底行列とアクティベーション行列の積により近似する．このような因子分解によるモデルを因子モデルと呼ぶ．また，この NMF をマルチチャネル音源分離のために拡張したマルチチャネル NMF (MNMF) も提案されている．MNMF では観測スペクトログラムを基底行列とアクティベーション行列と空間特徴量に分解する．MNMF は空間特徴量も因子分解により推定するため，音源モデル・重畳過程がともに因子モデルによりモデル化された手法である．

２．研究の目的

(1) 音源分離を確率的に統合した同時発話音声認識を行う手法を提案する．音源分離により復元される音声信号には不確実性が存在するため，音声信号の事後分布を考慮することで音声認識との統合を行う．これにより，復元すべき音声を一意に定めることなく混合音から直接認識結果を得ることが可能となる．

(2) 音の重畳過程と音源モデルを統合したモデルにより音源分離を行う．従来法では，重畳過程が混合モデルでも因子モデルでもモデル化されていたことから，その二つのモデル化の関係性に着目し，音源モデルにおいても混合モデルと因子モデルの二つのモデル化を行った．このように，重畳過程・音源モデルに対して，混合モデル・因子モデルのそれぞれのモデル化を行うことで，各モデルの音源分離性能を比較評価する．

３．研究の方法

(1) 潜在変数である単独発話音声 $\boldsymbol{S}$ には不確実性が存在するため，$\boldsymbol{S}$ のあらゆる可能性を織り込みつつ最終的な単語列 $\boldsymbol{Z}^*$ のみを求める統計的手法を確立したい．数学的には，以下の最適化問題

$$p(\boldsymbol{Z}|\boldsymbol{X}) = \int p(\boldsymbol{Z}|\boldsymbol{S})p(\boldsymbol{S}|\boldsymbol{X})d\boldsymbol{S}$$

$$\boldsymbol{Z}^* = \underset{\boldsymbol{Z}}{\operatorname{argmax}}\, p(\boldsymbol{Z}|\boldsymbol{X})$$

を一挙に実行することになる．このとき，確率的な枠組みを用いることで，未知の単独発話音声 $\boldsymbol{S}$ を積分消去することができ，原理上は明示的に考慮する必要がなくなる．現実的には，上記の積分計算を解析的に行うことは難しいので，Markov chain Monte Carlo methods (MCMC) を用いて近似することを考える．

$$\boldsymbol{S}_l \sim p(\boldsymbol{S}_l|\boldsymbol{X})$$

$$\boldsymbol{Z}^* = \underset{\boldsymbol{Z}}{\operatorname{argmax}} \frac{1}{L}\sum_{l=1}^{L} p(\boldsymbol{Z}|\boldsymbol{S}_l)$$

ここで，$\boldsymbol{S}_l$ は MCMC において l 番目にサンプリングされた分離音声を示す．この方式を下記の図 1 に示す．

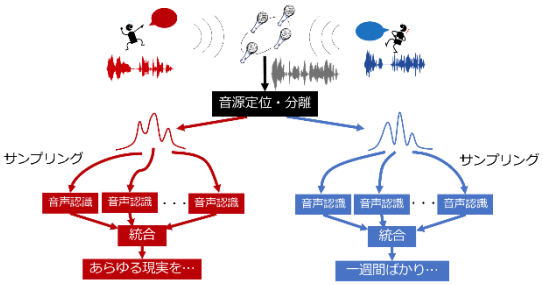

図 1：音源分離と音声認識との統合モデル

まず，分離音声 $\boldsymbol{S}_l$ をサンプリングするため，大塚らによるノンパラメトリックベイズ LDA に基づく音源分離手法を使用する．この手法では，MCMC の一種である Collapsed Gibbs Sampler (CGS) を用いて分離音をサンプルできるため，音声認識との確率的な統合に適している．具体的には，Hierarchical Dirichlet Process Latent Dirichlet Allocation (HDP-LDA) が用いられている．LDA とは主に自然言語処理において単語のトピック分類に用いられる手法であり，HDP-LDA は LDA のパラメータも確率変数として扱うことでトピック数が未知でも分類できるようにしたものである．HDP-LDA を用いることで混合音の各時刻各周波数ビンを音源ごとに複数回クラスタリングした結果から，各時刻各周波数ビンでそれぞれの音源に対する重みを決定し分離音を出力する．このクラスタリング操作を何度も繰り返すことで複数の分離音を出力することができる．同時に音源定位も行うため，複数回サンプリングを行なっても同じ音源から発せられた分離音の同定が容易である．

次に，多数の分離音声の多数のサンプル $\boldsymbol{S}$ を用いて音声認識結果 $\boldsymbol{Z}^*$ を求める．ここで，以下の二つの仮定をおく．

1. 音声認識結果における単語間には相関がなく，事後分布において独立である．
2. 音声認識結果の不確実性は極めて小さい（事後分布がディラックのデルタ関数で近似可能）．

これらの仮定のもとでは，$\boldsymbol{Z}^*$ 中の k 番目の単語 $\boldsymbol{Z}_k^*$ はそれぞれ独立に求まる．

$$\boldsymbol{Z}_k^* = \underset{\boldsymbol{Z}_k}{\operatorname{argmax}} \frac{1}{L}\sum_{l=1}^{L} \delta_{f(\boldsymbol{S}_l)}(\boldsymbol{Z}_k)$$

ここで，$f(\boldsymbol{S}_l)$ はある音声認識器（本研究では Julius を使用）を用いて分離音声 $\boldsymbol{S}_l$ を認識した結果を表す．これは，単語ごとに多数決によって結果を選択することを意味しており，これは ROVER 法に他ならない．ROVER 法の中でも，単語の出現頻度のみを用いる最も単純な方法を意味している．ROVER 法では，単語ごとの信頼度を用いることでさらなる精度の向上が可能である．ROVER 法では，複数の文章に対して単語単位でのアライメントにより単語組を生成し，それぞれの単語組において投票により結果を選択する．投票にあたっては，信頼度を用いる手法では，単語の出現頻度と信頼度平均を重み付けした尺度で投票重みを決定し，各単語組の中で最も得票数の多かったものを結果として出力する．本手法ではこの単語の信頼度を考慮した ROVER 法を用いて認識結果の統合を行う．

(2) 音の重畳過程と音源モデルを統合したモデルにより音源分離を行う（図 2）．

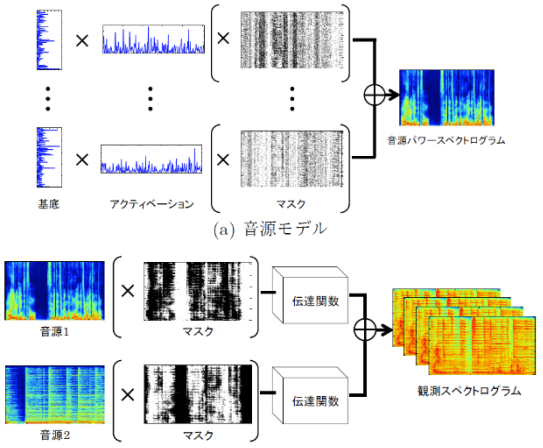

図 2：音源モデルと重畳モデル

音源モデルでは，各音源のパワースペクトログラムは基底・アクティベーション（・マスク）により構成される．重畳過程では，混合音スペクトログラムが音源スペクトログラム・伝達関数（・マスク）により構成される．音源モデル・重畳過程のいずれにおいても，因子モデルあるいは混合モデルを用いた定式化が可能である．

因子モデルでは音源のパワースペクトログラムが基底とアクティベーションの積の和から成り，観測スペクトログラムは音源スペクトログラムと伝達関数から生成されるとする．混合モデルでは音源のパワースペクトログラムがマスクによって選択された基底とアクティベーションから成り，観測スペクトログラムはマスクによって選択された音源スペクトログラムと伝達関数から生成されるとする．提案法では混合モデルには LDA，因子モデルには NMF の枠組みを用いてギブスサンプリングにより音源分離を行う．

音源モデル ＼ 重畳過程	混合モデル	因子モデル
混合モデル	LDA-LDA	LDA-NMF
因子モデル	NMF-LDA	NMF-NMF

４．研究成果

(1) 本手法と既存手法による二話者同時発話音声に対する単語正解精度の比較を行った．音声認識精度を測る指標として単語正解精度を用いた．評価に使用する音声は ATR503 文 a01~a50 から選択した．音声認識はすべての手法において Julius により行い，言語モデルも同じモデルを使用した．混合前の音声をそのまま認識した結果，Independent Vector Analysis (IVA) により混合音を分離・認識した結果，大塚らの手法により分離・認識した結果，本手法による認識結果を求めた．既存の音源分離手法は多数あるが，IVA は発話音声の混合音に対して高い分離性能を発揮するとされているため，ここでは比較対象として IVA を選択した．混合音声には，50 文の中からランダムに 2 つの文を選び，0◦，60◦，60◦のうちの 2 ヶ所にランダムに配置して得られるシミュレーション混合音を 50 個使用した．結果を表 1 に示す．本手法と従来手法を比較すると，本手法により約 14 pts 単語正解精度が向上することが確認できた．

表 1：二話者同時認識結果

	混合前	IVA	大塚法	**提案法**
単語正解精度	66.8	−17.6	6.17	**25.6**

さらに，話者の増加（三話者の環境下）に対する頑健性を評価した結果を表 2 に示す．この結果より，三話者においても本手法によって IVA，大塚らの手法よりも単語正解精度が向上することが確認できた．ただし，混合前の音声より 41.2 pts 単語正解精度が劣っており，音源分離技術の改善が望まれる．

表 2：三話者同時認識結果

	混合前	IVA	大塚法	**提案法**
単語正解精度	66.8	29.4	29.6	**43.7**

本研究では，ベイズ推定に基づいたモデルを構築することで混合音に対する音源分離の不確実性を考慮した音声認識手法について述べた．本手法により従来手法で音源分離して得られた分離音をそのまま認識するよりも，単語正解精度が向上することを確認した．この方式は，人間は混合音声から分離音を一意に定めることなく発話内容を直接認識できるという知見とよく一致している．

(2) 提案法の分離性能を比較評価するため，シミュレーション混合により合成した音響信号を用いた実験を行った．比較手法として，IVA，マルチチャネル NMF (MNMF)，音源モデルをスパースとし空間モデルに LDA を用いた分離法を用いた．本来，LDA を用いた手法では，音源数の推定を行うことができるが，条件を対等にするため音源数は既知とした．具体的には，残響時間 400 ms のインパルス応答を用いて 3 音源を混合した音声を用いた．マイク数は 4 とした．混合音には，音声のみの混合音と音楽のみの混合音，音声と音楽の混合音をそれぞれ 10 個ずつ使用した．用いる音楽と音声は SiSEC と JNAS の音素バランス文から選択した．サンプリング周波数は 16 kHz とし，STFT では窓幅 512 のハミング窓をシフト幅 256 で使用した．評価尺度として，SDR, SIR, SAR を用いた．

表 3〜5 に実験結果を示す．それぞれの条件において最も数値が大きくなったものを太字で示した．SIR は LDA-LDA が最も高い性能を示したが，SDR と SAR は NMF-NMF が最も高い性能を示した．また，提案法では事前分布のパラメータとして使用するマイクロホンアレイのインパルス応答を必要とするが，無響室で録音したインパルス応答を用いても残響時間 400ms の環境下での混合音の分離ができたことから，未知の環境下においても使用するマイクロホンアレイが同じであれば提案法により音源分離が可能である．

表 3：音楽信号の分離

	SDR	SIR	SAR
IVA	3.4 dB	7.5 dB	7.1 dB
MNMF	4.8 dB	10.0 dB	**7.7** dB
LDA	5.5 dB	15.1 dB	6.3 dB
NMF-LDA	4.2 dB	14.0 dB	5.2 dB
LDA-LDA	5.8 dB	**17.0** dB	6.3 dB
NMF-NMF	**6.0** dB	12.6 dB	7.5 dB

表 4：音声信号の分離

	SDR	SIR	SAR
IVA	0.3 dB	4.9 dB	5.7 dB
MNMF	1.0 dB	6.2 dB	6.7 dB
LDA	0.7 dB	7.4 dB	4.1 dB
NMF-LDA	0.5 dB	**8.7** dB	3.2 dB
LDA-LDA	0.7 dB	**8.7** dB	3.3 dB
NMF-NMF	**3.2** dB	8.1 dB	**7.5** dB

表 5：音楽＋音声信号の分離

	SDR	SIR	SAR
IVA	0.1 dB	5.3 dB	5.3 dB
MNMF	1.8 dB	8.6 dB	6.1 dB
LDA	2.4 dB	11.5 dB	4.5 dB
NMF-LDA	1.1 dB	9.8 dB	4.1 dB
LDA-LDA	2.8 dB	**14.2** dB	3.9 dB
NMF-NMF	**4.9** dB	13.0 dB	**6.6** dB

本研究では，音源モデル・重畳過程のそれぞれに対して，因子モデルと混合モデルを用いたモデル化を行うことで複数の音源分離法を提案した．それぞれのモデル化による音源分離法を比較すると，SDR，SAR の観点では音源モデルと重畳過程が共に因子モデルによりモデル化された手法が最も性能が高く，SIR の観点では音源モデルと重畳過程が共に混合モデルでモデル化された手法が最も性能が高いことを確認した．

今後の方向性として，計算量を削減するため，MCMC の代わりに，変分ベイズ法 (VB) を用いた決定論的な手法を導入することが考えられる．また，音声認識器と統合するような応用を考えた場合には，リアルタイムでオンライン処理を行うための拡張が必要である．これに関しては，ミニバッチ処理やオンライン VB などにより達成可能と考えられる．

５．主な発表論文等
（研究代表者，研究分担者及び連携研究者には下線）

〔学会発表〕（計 7 件）

Kousuke Itakura, Yoshiaki Bando, Eita Nakamura, Katsutoshi Itoyama, Kazuyoshi Yoshii, Tatsuya Kawahara, “Bayesian Multichannel Nonnegative Matrix Factorization for Audio Source Separation and Localization”, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2017.

Kousuke Itakura, Yoshiaki Bando, Eita Nakamura, Katsutoshi Itoyama, Kazuyoshi Yoshii, “A Unified Bayesian Model of Time-Frequency Clustering and Low-Rank Approximation for Multi-Channel Source Separation”, European Signal Processing Conference (EUSIPCO), pp. 2280–2284, 2016.

Kousuke Itakura, Izaya Nishimuta, Yoshiaki Bando, Katsutoshi Itoyama, Kazuyoshi Yoshii, “Bayesian Integration of Sound Source Separation and Speech Recognition: A New Approach to Simultaneous Speech Recognition,” Annual Conference of the International Speech Communication Association (Interspeech), pp. 736–740, 2015.

板倉光佑, 坂東宜昭, 中村栄太, 糸山克寿, 吉井和佳, 河原達也, “マルチチャネル音源分離のための低ランク音源モデルとスパース重畳過程に基づくネスト型ベイズ混合・因子モデル”, 電子情報通信学会 第 19 回情報論的学習理論ワークショップ, IBISML2016-95, Vol. 116, No. 300, pp. 353–359, 2016.

板倉光佑, 坂東宜昭, 中村栄太, 糸山克寿, 吉井和佳, 河原達也, “マルチチャネル音源分離のためのネスト型基底・音源混合モデルに基づく時間周波数クラスタリング”, 電子情報通信学会 音声研究会, SP2016-31, Vol. 116, No. 189, pp. 25–28, 2016.

島田一希, 坂東宜昭, 板倉光佑, 三村正人, 糸山克寿, 吉井和佳, 河原達也, “遠隔音声認識のためのブラインド音源分離に基づくビームフォーマ”, 情報処理学会 第 79 回全国大会, pp. 485–486, 2017.

板倉光佑, 坂東宜昭, 中村栄太, 糸山克寿, 吉井和佳, “音源スペクトログラムの低ランク性とスパース性を考慮した NMF-LDA に基づくマルチチャネル音源定位と音源分離”, 情報処理学会 第 78 回全国大会, pp. 485–486, 2016.

〔その他〕
ホームページ
http://sap.ist.i.kyoto-u.ac.jp/members/yoshii/

６．研究組織
(1)研究代表者
吉井 和佳（YOSHII, Kazuyoshi）
京都大学・大学院情報学研究科・講師
研究者番号：２０５１０００１

(2)研究分担者
糸山 克寿（ITOYAMA, Katsutoshi）
京都大学・大学院情報学研究科・助教
研究者番号：６０６１４４５１

(3)連携研究者
河原 達也（KAWAHARA, Tatsuya）
京都大学・大学院情報学研究科・教授
研究者番号：００２３４１０４

持橋 大地（MOCHIHASHI, Daichi）
統計数理研究所・モデリング研究系・准教授
研究者番号：８０４１８５０８