5P-02

混合音に対する音源分離の不確実性を考慮した同時発話音声認識

板倉 光佑[1]　西牟田 勇哉[2]　坂東 宜昭[2]　糸山 克寿[2]　吉井 和佳[2]
[1]京都大学 工学部 情報学科　[2]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

実環境下での音声認識は，雑音や残響を含んだ混合音を処理する必要があるため，そのような混合音に対して音源分離などをすることにより，もとの音声を復元するアプローチが主流である [1]．しかし，一意に音声信号を復元し，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，復元した音声が音声認識にとって最適である保証はなかった．さらに，従来の音声認識の研究は単独発話を想定したものが多く，複数話者の同時発話を含む音声信号はそのままでは扱えなかった．

そこで本稿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する (図 1)．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．ただし，そのためには音源分離を確率的に取り扱う手法も必要となるため本手法では大塚らによるノンパラメトリックベイズを用いた音源分離手法 [2] を使用する．本手法により，複数話者による同時発話音声に対する単語正解精度が向上することを確認した．

2. 音源分離・音声認識の統合モデル

本章では，音源分離・音声認識を確率的に統合するためのアプローチについて述べる．2.1 節で音源分離と音声認識の確率的統合のための理想的モデル，2.2 節でそのモデルを解くための本手法のアプローチについて述べる．

2.1 確率モデルによる統合

まず本手法の基となるモデルについて考える．従来の音声認識では，式 (1)(2) に従い，音源分離により混合音 X から分離音声 S^* を，音声認識により分離音声 S^* から認識結果 Z^* をそれぞれ点推定していた．

$$S^* = \underset{S}{\operatorname{argmax}}\, p(S|X) \tag{1}$$

$$Z^* = \underset{Z}{\operatorname{argmax}}\, p(Z|S^*) \tag{2}$$

しかし，このような手法では分離音声 S^* を一意に定める必要があり，音源分離と音声認識は独立した処理として扱われる．そこで本手法では，従来手法のように点推定により S^* を一意に定め，認識結果 Z^* を求めるのではなく，ベイズ推定により S について周辺化することで Z の予測分布を求め，Z^* を推定する．ベイズ推定では式 (3)(4) に従い認識結果 Z^* を推定することができる．

$$p(Z|X) = \int p(Z|S)p(S|X)dS \tag{3}$$

$$Z^* = \underset{Z}{\operatorname{argmax}}\, p(Z|X) \tag{4}$$

式 (3) の積分計算を行うことは難しいが，Markov chain Monte Carlo methods (MCMC) を用いることで，より取り扱いやすい式への近似が可能である．MCMC により式 (3) を近似し，式 (4) と合わせると式 (3)(4) は式 (5) のように近似することができる．

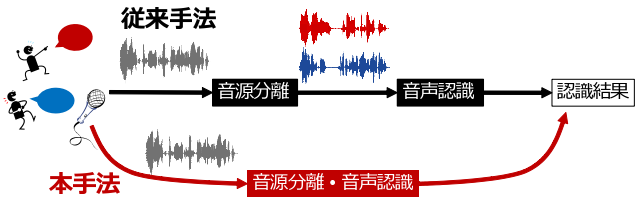

図 1: システム概要

$$Z^* \approx \underset{Z}{\operatorname{argmax}}\, \frac{1}{L}\sum_{l=1}^{L} p(Z|S_l) \tag{5}$$

$$S_l \sim p(S_l|X) \tag{6}$$

ここで，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．

式 (5)(6) を解くことで認識結果を求めることが出来るが，式 (5) を解くことも容易ではない．よって本手法では 2 つの仮定をおくことで式 (5) を簡略化し結果を得る．

2.2 モデルへのアプローチ

本節では式 (5)(6) に対するアプローチを述べる．まず式 (6) に従って S_l をサンプリングする．これは MCMC に基づいて分離音を出力する必要がある．ここで，大塚らにより MCMC の 1 つである Collapsed Gibbs Sampler (CGS) に基づいた分離手法が提唱されている [2]．大塚らの手法では，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．大塚らの手法はこの 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ごとに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．またこの大塚らの手法では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．したがって本手法ではこの大塚らの手法により複数の分離音をサンプリングする．

次に，得られた S_l を用いて式 (5) を解く．ただし式 (5) を解くことは困難なので以下の 2 つの仮定をおく．

1. 音声認識結果 Z 中の単語同士には相関がなく，事後分布において独立である．
2. 音声認識結果にはほとんど不確実性がない．

まず，1 の仮定は数式では以下のように表される．

$$p(Z|X) = \prod_{k=1}^{K} p(Z_k|X) \tag{7}$$

ただし，Z_k は認識結果 Z の k 番目の単語を表す．ここで，$p(Z|X)$ の最大化はそれぞれの k について $p(Z_k|X)$ を最大化することと等価となるため，以降はそれぞれの k について最適な単語 Z_k^* を求めることを目的とする．

Simultaneous Speech Recognition Considering the Uncertainty of Separated Sounds: *Kousuke Itakura, Izaya Nishimuta, Yoshiaki Bando, Katsutoshi Itoyama, Kazuyoshi Yoshii (Kyoto Univ.)*

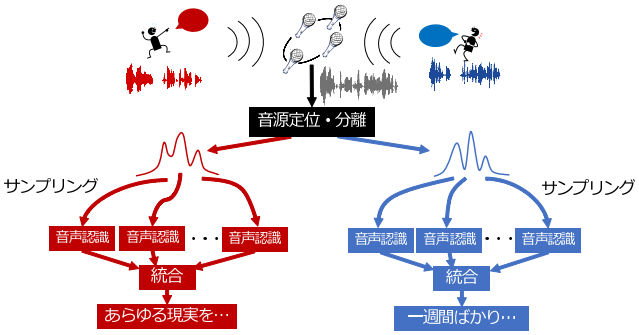

図 2: 内部モデル

また，音声認識にはオープンソースの Julius [3] を使用する．Julius による音声認識は分離音声 S を入力とし，認識結果 Z を出力する関数 $f(S) = Z$ であると考える．すると，2 の仮定より以下の数式が導かれる．

$$p(Z_k|S_l) \quad = \quad 1_{f(S_l)}(Z_k) \tag{8}$$

ここで，$1_{f(S_l)}(Z_k)$ は $f(S_l)$ の k 番目の単語が Z_k のとき 1，そうでないときは 0 となる関数である．

これらの仮定から，式 (5) は式 (9) に変形できる．

$$Z_k^* \quad \approx \quad \underset{Z_k}{\operatorname{argmax}} \frac{1}{L} \sum_{l=1}^{L} 1_{f(S_l)}(Z_k) \tag{9}$$

ここで，式 (9) 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 [4] に他ならない．式 (9) は 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．さらに，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．これらの処理の流れを図 2 に示す．

3. 評価実験

音声認識精度を測る指標として単語正解精度を用いた．単語正解精度は次式のように定められている [5]．

$$\text{単語正解精度} \quad = \quad \frac{C - I}{T} \times 100 \tag{10}$$

C は認識結果において正解した単語の数，I は挿入誤りした単語の数，T は正解の文章に含まれる単語の数である．また，S を挿入誤りした単語の数，D を削除誤りした単語の数とすると T は以下のように求められる．

$$T = C + S + D \tag{11}$$

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．

3.1 二話者同時発話音声に対する認識精度

3.1.1 実験目的・条件

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．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 [6]，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．

表 1: 二話者同時発話音声認識精度

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

表 2: 三話者同時発話音声認識精度

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

3.1.2 実験結果

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．よって本手法は二話者の同時発話音声の認識に対して有効であると言える．

3.2 三話者同時発話音声に対する認識精度

3.2.1 実験目的・条件

本手法の話者の増加に対する頑健性を評価した．IVA，大塚らの手法により分離・認識した結果と提案手法による認識結果を求め比較した．混合音声には，50 文の中からランダムに 3 つの文を選び，$0°, -60°, 60°$ の 3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．

3.2.2 実験結果

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．このことから話者数が増えても本手法が有効であると言える．ただし，混合前の音声より 41.2 pts 単語正解精度が劣っているため，さらなる改善が必要である．

4. おわりに

本稿では，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．

本稿では，1. 音声認識結果において各単語は独立である，2. 音声認識結果には不確実性がない，という 2 つの強い仮定をおいたが，今後はこれらの仮定を緩めることで認識精度の向上を目指す．さらに，残響・雑音除去や音源分離と音声認識の完全なベイズ的統合についても研究を進めていく予定である．

謝辞 本研究の一部は、科研費 24220006 の支援を受けた．

参考文献

[1] 奥乃博 他: “音声ストリーム分離法の提案と複数音声の同時認識の予備実験”, Trans.IPS.Japan-1997.
[2] T. Otsuka *et al.*: “Unified auditory functions based on Bayesian topic model”, *IROS-2012*.
[3] A. Lee *et al.*: “Julius — An Open Source Real-Time Large Vocabulary Recognition Engine”, *EUROSPEECH-2001*.
[4] J. G. Fiscus: “A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)“, *ASRU-1997*.
[5] K. Lee: “*Automatic Speech Recognition: The Development of the SPHINX Recognition System*”, Springer, 1989.
[6] D. Kitamura *et al.*: “Efficient multichannel nonnegative matrix factorization with rank-1spatial model”, *ASJ-2014*.