バイノーラル音源分離の音声認識による評価*

○中谷 智広† 南 泰浩†‡
(日本電信電話株式会社 †NTT コミュニケーション科学基礎研究所 ‡NTT サイバースペース研究所)

1 はじめに

実環境音声認識に向けて、混合音声から分離した音声の認識実験を行った。分離による音声の品質改善は、人間の耳で評価すると明らかであるが、音声認識システムで評価した場合には、ある程度の認識率改善にとどまり実際の問題に適用できるレベルには達していなかった [1]。今回、音声認識に悪影響している原因を調査し、認識率向上にフィードバックする目的で新たに実験を行った。認識モデルの学習のために、混合音から分離した音声を用いることで認識率が大幅に改善することを示す。

2 音源分離方法

本稿では、有声音の性質である調波構造と音源位置の情報を組み合わせて用いる音源分離法を適用する [2]。ハンズフリー音声認識の研究では、主に、音源位置に基づいて音声強調を行うマイクロフォンアレイ技術が用いられているが [3]、これと比較して、本方式は、マイク数より多くの音源が存在する場合でも音源分離を行うことが出来るという利点がある。一方、マイク間の位相差・強度差情報を用いる分離法でも、原理的にマイク数以上の音源を分離することが出来る [2, 4]。本稿では、この方法を含めて、バイノーラルマイクロフォン（2 マイク）を用いて 2 音源、および 3 音源の分離、認識を行う。

音源分離法は、調波構造と音源方向情報を用いて有声音を分離する処理（処理 1）と音源方向情報だけから無声音を分離する処理（処理 2）の 2 つからなり、それらの組み合わせ方によって、以下の 3 つの分離法を構成する (図 1 参照)。

- AR (All-residue): 有声音は処理 1、無声音は処理 1 の残差を無声音としてそのまま分離音に付加する。無声音は混ざったまま。
- OR (Own-residue): 有声音は処理 1、無声音は処理 1 の残差に処理 2 を施して分離音に付加。
- DO (Direction-only): 有声音も無声音も処理 2 で一緒に分離。

各分離音を人間の耳で比較すると、OR が一番良くやや劣るが DO はほぼ同等の品質で、無声音がまざったままである AR の音質が一番悪い。一方、スペクトル歪みを LPC ケプストラムを用いて評価すると、処理 2(OR の無声音分離と DO に利用) の歪みが大きいことがわかっており、LPC ケプストラムに基づく音声認識を行った結果は、AR による認識が一番良い [1, 2]。これは、処理 2 の音源方向による分離方式に原因がある。バイノーラルマイクロフォンの頭部伝達関数（HRTF）の特性に基づき、FFT の周波数 bin ごとに音源方向を定め、同一方向の bin だけを使って音を合成するため、周波数 bin に欠落が生じるからである。また一方で、ゼロ点の影響をあまり受けない MFCC に基づく方法では、OR が一番良い認識結果が得られていた。

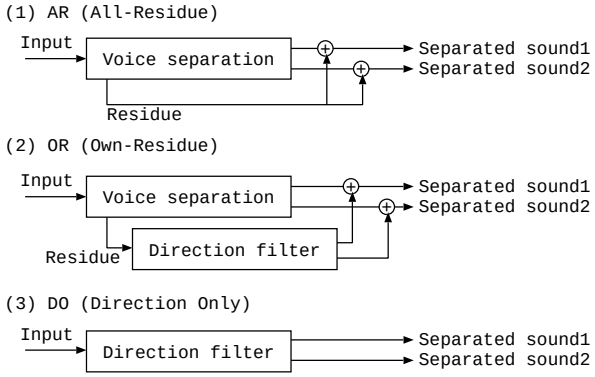

図 1: 3 つの音源分離処理のダイアグラム

なお、本分離方式では、音源数、音源位置の情報は事前情報として与えず、入力音から各方式に基づいて自動抽出する (詳細は [2] 参照)。

3 HMM の学習方法

分離音声の認識実験を行うために、学習モデルとして以下の 3 つのモデルを用意した。

- 単独音モデル：オリジナル音声に HRTF を畳み込んだもので学習
- 分離音モデル：混合音から分離した音声で学習
- 混合音モデル：混合音のままの音声で学習

単独音モデルは、バイノーラルマイクロフォンで音声を録音するときに生じるスペクトル変形を加えた音声で学習を行うので、混合音が完全に構成音に分離されるのであれば、ほぼ正しく認識が行えるはずである。実際に、HRTF を畳み込まない音声で学習したモデルよりも認識率が向上することが確かめられている [1]。分離音モデルは、音源分離の際に生じるスペクトル変形を含んだ音声の分布を学習しているため、クリーンな音声で学習した単独音モデルに比べて、分離音声をある程度ロバストに認識できるようになると期待される。また、混合音モデルは、分離音モデルによる分離音声の認識との比較のために、混合音のままの音声の認識に用いるモデルである。

4 分離実験

ATR 単語データベース (A セット, 5240 単語, 12 kHz 標本化、16 bit 量子化) から 2 話者 (女性 2 人:FKM と FSU)、3 話者 (女性 2 人:FKM と FSU, 男性 1 人:MAU)

*Recognition of mixed speech using binaural sound source separation, by Nakatani, T., Minami, Y. (NTT Corporation)

表 1: 各 5240 単語の分離成功率 (%)

	分離法	FKM	FSU	MAU	平均
2 話者	AR,OR	100	100	-	100
	DO	100	100	-	100
3 話者	AR,OR	99.4	98.3	95.1	97.4
	DO	96.8	95	87.2	93.0

を取出して認識評価用の混合音を作成した。混合の前に、各単語の音素ラベルから求まる有声音区間の平均パワーが構成音どうしで同一になるようにパワー正規化を行い、ダミーヘッドマイクロフォンを用いて測定した頭部伝達関数 (HRTF) を畳み込んでバイノーラル音を作成した。2 音の場合は、ダミーヘッドから見て FKM を右 30 度、FSU を左 30 に、3 音の場合は FKM を右 60 度、FSU を左 60 度、MAU を正面に配置した。これらの音を計算機上で足し合わせて混合音とした。オープンテストとするため、認識用と学習用に分けて混合音は作成した。ATR データベース中の 10 個おきの単語 524 個を取出し、促音を除いて共通のラベルとなる単語、および音節数の短い単語を除いた 500 単語を認識用単語とし、残りの 4740 単語を学習用単語とした。各認識用単語 500 組どうし、学習用単語 4740 組どうしの混合音を 2 話者、3 話者についてそれぞれ作成した。

分離実験で正しく分離された単語数の割合を表 1 に示す。分離音は、分離と一緒に抽出される音源方向属性を用いて各話者に対応付けた。話者の配置方向に分離された音声がない場合は、当該話者の分離に失敗したとした。表より、2 話者分離については、AR, OR, DO すべての方法で、全 5240 混合音から全話者が分離された。また、3 話者分離でも、各分離法共に 90%以上の単語の分離を行うことができた。なお、次節の認識実験では、3 話者で分離に失敗した単語のうち分離音モデルの学習対象音は学習から除外し、認識対象音は混合音をそのまま代用した。

5 認識実験

単独音モデル、分離音モデル、および混合音モデルを用いて、500 組の分離音声に対し特定話者単語音声認識を行った。分析条件は、MFCC12 次元、デルタケプストラム 12 次元、HMM 状態数 3、各状態でのガウス混合分布数 5, フレーム長 25ms, フレームシフト 5ms で、モノフォンまたはトライフォンを用いた。認識結果を表 2〜5 に示す。表中の Orig は混合前の音声の認識率、Mix は混合音に対する認識率、AR,OR,DO は各分離法で分離した音声に対する認識率を示す。

表より、すべての条件で分離処理は認識率を向上させている。特に、分離音モデルを用いる効果が大きく、トライフォンによる学習と組み合わせることで大幅な認識率の改善が得られた。また、各分離法を比較すると、単独音モデルでは AR か OR が最も高い認識率を出していたが、分離音モデルでは、DO が一番良い結果を出す場合もあった。これらのことから、以下のことがいえると考える。

- 分離処理は音声認識に必要な特徴量の多くを失わずに、妨害音声の影響を効果的に低減している。
- 分離音モデルを用いることで、各混合条件および分離条件に固有のスペクトル変形を含んだ音声を学習することができる。

表 2: 2 話者分離の単語認識率 (%): モノフォン

モデル	Orig	Mix	AR	OR	DO
単独音	98.3	36.6	62.1	73.3	63.8
混合/分離音	-	71.6	89.7	96.0	95.0

表 3: 2 話者分離の単語認識率 (%): トライフォン

モデル	Orig	Mix	AR	OR	DO
単独音	99.8	44.6	69.5	59.7	47.0
混合/分離音	-	91.1	96.9	96.1	95.0

表 4: 3 話者分離の単語認識率 (%): モノフォン

モデル	Orig	Mix	AR	OR	DO
単独音	98.5	21.0	37.6	48.9	40.7
混合/分離音	-	57.9	66.9	78.2	79.1

表 5: 3 話者分離の単語認識率 (%): トライフォン

モデル	Orig	Mix	AR	OR	DO
単独音	99.7	23.4	38.9	43.2	32.1
混合/分離音	-	80.3	88.9	91.9	94.5

- 分離音モデルによる認識性能は、混合前の音声の特徴に近いものが有利とは限らず、分離音固有の変形の分布に依存すると思われる。

6 まとめ

本稿では、混合音から分離された音声の認識率を悪化させている原因を調査し、認識にフィードバックさせる目的で、調波構造と音源方向を用いた音源分離法による分離・認識実験を行った。実験により、混合音から分離した音声を用いて HMM の学習を行う分離音モデルを用いることで、単語認識率が大幅に改善することを示した。これにより、単語認識に必要な特徴量の多くは分離音声にも含まれていることが予想される。今後の課題としては、混合状態が異なる入力に対する分離音に対しても適用可能な認識法の研究があげられる。現在、ミッシングフィーチャー理論との統合を検討している [5]。また、今回の実験では、実環境音声認識で課題となる残響耐性については調べられていない。分離システムおよび音声認識を残響下で評価することも今後の課題である。

分離音声認識の予備実験を行っていただいた早稲田大学の小川良彦氏、認識実験の技術サポートをしていただいた NTT-AT 社の高橋勝吾氏に感謝する。片桐滋部長をはじめ、日頃より熱心な議論をいただいている NTT CS 研音声オープンラボの方々に感謝する。

参考文献

[1] Okuno, H. G. *et al.*, *Speech Communcation*, Vol. 27, Nos. 3-4, pp. 299–310, Elsevier, 1999.

[2] Nakatani, T. *et al.*, *Speech Communcation*, Vol. 27, Nos. 3-4, pp. 209–222, Elsevier, 1999.

[3] Yamada, T., *et al.*, *Proc ICASSP-98*, pp. 245–248, 1998.

[4] Aoki, M., *et al.*, *J. Acoust. Soc. Jpn. (E)* 20, 2, 1999.

[5] Raj, B., *et al.*, *Proc ICSLP-2000*, 00341, 2000.