人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-061-09

深層ブラインド音源分離と転移学習に基づく遠隔音声認識の評価

合澤　隆拓[1,2*]　坂東 宜昭[2]　糸山 克寿[1]　西田 健次[1]　中臺 一博[1]

[1] 東京工業大学　[2] 産業技術総合研究所

概要: 本稿では，深層ブラインド音源分離と転移学習に基づく遠隔音声認識について述べる．複数話者による同時発話の遠隔音声認識には，混合音から各話者の音声を抽出する音源分離が不可欠である．複数話者の遠隔音声認識では，発話区間情報と深層生成モデルを用いた弱教師あり深層ブラインド音源分離が高い性能を発揮することが知られている．しかしこの手法は，混合音を時間周波数クラスタリングした結果を補助特徴量として用いるため，特徴量の計算に時間を要する課題があった．本研究では，より簡便な特徴量で頑健に動作する枠組みの確立を目指す．具体的には，学習データに対するクラスタリング結果を疑似教師として深層生成モデルを事前学習し，学習済みモデルを初期値として深層ブラインド音源分離に転移する．ディナーパーティの遠隔音声認識 (CHiME-6 Challenge) の単語誤り率を評価することで，提案法の有効性を検証した．

1　はじめに

音源分離は，観測信号から個々の音源を抽出する技術で，遠隔音声認識のフロントエンドとして不可欠である [1]．スマートスピーカに代表されるように，単一話者の遠隔音声認識は高い性能を達成しているが，複数話者による会話の遠隔音声認識は，オーバーラップや話者数がフレームごとに変動するといった多くの課題がある [2]．未知環境でも頑健に動作する遠隔音声認識のフロントエンドには，表現力が高く未知環境でも頑健な手法が求められる．

音源やマイクアレイに対する事前の情報を殆ど用いないブラインド音源分離 (BSS) は，遠隔音声認識のフロントエンドとして広く活用されている [1,3,4]．例えば，混合複素角度中心ガウスモデル (cACGMM) [3,5] は，少ない計算量で小さな音源の移動や残響にも耐える手法として広く研究されている．しかし，本手法は線形の生成モデルに基づくため，表現力に限界があった．そこで非線形生成モデルに基づく深層フルランク空間相関分析 (Neural FCA) [6] が提案されている．このモデルは，VAE [7] の枠組みに基づき，混合音から各音源の潜在変数を推定する推論モデルと潜在変数から分離音を再構成する音源生成モデルからなる．Neural FCA は，数値混合音を用いた音声分離 [6] や，遠隔音声認識のフロントエンド [8] として，cACGMM を含む従来の BSS を上回る性能が報告されている．

Neural FCA を遠隔音声認識のフロントエンドに適用した従来研究 [8] では，音源数が既知であると仮定する従来の Neural FCA に対して，各音源の発話区間情報を生成モデルに導入し，音源数が変動する日常会話の認識を行った．本手法は発話区間情報を用いるため弱教師あり Neural FCA と呼ばれ，cACGMM の分離結果を補助特徴量として推論モデルに入力する．本枠組みは，推論時に cACGMM と Neural FCA の 2 つの BSS を実行する必要があるため，比較的大きな計算時間を要すが，cACGMM の分離結果を入力しない場合は大幅に性能が劣化する．これは，混合音のみの特徴量から教師なしで音源分離を学習する問題が難しく，性質の悪い局所解に陥ってしまうためと考えられる．

統計的モデルを用いた教師なし音源分離により推定された信号を疑似教師データとして教師あり学習する手法が提案されている．戸上ら [9] は，混合音から局所ガウスモデル (LGM) に基づく音源分離手法で分離された信号を疑似教師とし，疑似教師信号と分離信号の差分を Kullback-Leibler ダイバージェンスで最小化する学習を提案した．疑似教師あり学習したモデルは，学習データ全体から普遍的な知識を獲得するため，疑似教師そのものよりも高い分離性能を達成している．

本研究では，従来の BSS の分離結果を用いて疑似教師あり学習させた分離モデルから，弱教師あり Neural FCA を転移学習することで，より高い認識性能の実現を目指す．具体的には，混合音に従来の線形 BSS 法である cACGMM [3] を適用し，その分離結果を疑似教師として音源生成モデルとその推論モデルを学習する．学習されたモデルを初期値として弱教師あり Neural FCA を学習することで，混合音のみの特徴量では破綻する課題を解決する．cACGMM の分離結果を従来手法 [8] では推論モデルの入力に補助情報として使っていたが，本研究では疑似教師として用いるため，推論時は cACGMM の計算を削減できる．提案法は，ホームパーティでの会話を収録した CHiME-6 データセット [2] を用いて評価した．

*連絡先：東京工業大学
〒 152-8552 東京都目黒区大岡山 2-12-1
E-mail: aizawa@ra.sc.e.titech.ac.jp

2 深層 BSS に基づく音声分離

本研究の基盤となる弱教師あり Neural FCA [8] について説明する．

2.1 問題設定

本手法では，ホームパーティのような N 人が会話している状況で，以下の問題設定により音源分離を行う．

入力: M チャネル混合音 $\mathbf{x}_{ft} \in \mathbb{C}^M$ と話者 $n = 1, \ldots, N$ の時間フレーム t での発話有無 $u_{nt} \in \{0, 1\}$
出力: 話者 n の分離音 $\hat{s}_{nft} \in \mathbb{C}$

ここで，$f = 1, \ldots, F$ および $t = 1, \ldots, T$ はそれぞれ，周波数および時間インデックスを表す．

2.2 生成モデル

弱教師あり Neural FCA では，従来の深層 BSS (Neural FCA) に発話区間変数を導入した生成モデルを定義する．観測混合音 $\mathbf{x}_{ft}$ を以下のように N_{spk} 個の音源信号と N_{noi} 個の雑音信号の和 $s_{nft} \in \mathbb{C}$ $(n = N_{\mathrm{spk}} + 1, ..., N_{\mathrm{spk}} + N_{\mathrm{noi}})$ で表す．

$$\mathbf{x}_{ft} = \sum_{n \in \mathfrak{N}_t} \mathbf{a}_{nf} s_{nft} \tag{1}$$

ただし，$\mathfrak{N}_t = \{n | u_{nt} = 1\} \cup \{N_{\mathrm{spk}} + 1, ..., N_{\mathrm{spk}} + N_{\mathrm{noi}}\}$ は時間 t に存在する音源の集合，$\mathbf{a}_{nf} \in \mathbb{C}^M$ は音源 n のステアリングベクトルである．音源信号のパワースペクトル密度 (PSD) は，ピッチや包絡といった音源の特徴を表す低次元の潜在ベクトル $\mathbf{z}_{nt} \in \mathbb{R}^D$ を用いて以下のような零平均複素ガウス分布で表現する．

$$s_{nft} \sim \mathcal{N}_{\mathbb{C}}\left(0, g_{\theta,nf}(\mathbf{z}_{nt})\right), \tag{2}$$

$$z_{ntd} \sim \mathcal{N}\left(0, 1\right) \tag{3}$$

ここで，$g_{\theta,nf} : \mathbb{R}^D \to \mathbb{R}_+$ は，$\mathbf{z}_{nt}$ から PSD を出力するパラメータ θ を持つ深層ニューラルネットワーク (DNN) である．以上より，観測混合音 $\mathbf{x}_{ft}$ は，以下の多変量複素ガウス分布に従う．

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_{n \in \mathfrak{N}_t} g_{\theta,nf}(\mathbf{z}_{nt}) \mathbf{H}_{nf}\right) \tag{4}$$

ただし，$\mathbf{H}_{nf} = \mathbf{a}_{nf}\mathbf{a}_{nf}^{\mathsf{H}} \in \mathbb{S}_+^{M \times M}$ は周波数 f における音源 n の空間相関行列 (SCM) である．本稿では，$\mathbf{a}_{nf}$ の比較的小さな変動を許容すため，SCM をフルランクに緩和する．また DNN $g_{\theta,nf}$ は，次節で説明するように，事前に収集した多チャネル混合音と発話区間情報のみから教師なし学習する．

2.3 償却変分推論を用いた弱教師あり学習

弱教師あり Neural FCA では，多チャネル混合音と発話区間から音源モデル $g_{\theta,nf}$ を弱教師あり学習する．なお，モデルは事前学習した後に本学習に転移させるが，詳細は 3 章で述べる．本学習では対数周辺尤度 $\log p_\theta(\mathbf{X}|\mathbf{H}, \mathbf{U})$ を最大化するような音源モデル $g_{\theta,nf}$ を学習する．ただし，$\mathbf{X} \triangleq \{\mathbf{x}_{ft}\}_{n,t=1}^{N,T}$，$\mathbf{H} \triangleq \{\mathbf{H}_{nf}\}_{n,f=1}^{N,F}$，$\mathbf{U} \triangleq \{u_{nt}\}_{n,t=1}^{N,T}$ である．この対数周辺尤度は直接計算困難なので，以下の推論モデル $q_\phi(\mathbf{Z}|\mathbf{X}, \mathbf{U})$ を導入した変分償却推論 [7] を行う．

$$q_\phi(\mathbf{Z}|\mathbf{X}, \mathbf{U}) = \prod_{n,t,d} \mathcal{N}_{\mathbb{C}}\left(z_{ntd} \mid \mu_{\phi,ntd}(\mathbf{C}), \sigma^2_{\phi,ntd}(\mathbf{C})\right)$$

ただし，$\mathbf{Z} \triangleq \{z_{ntd}\}_{n,t=1}^{N,T}$ は潜在変数の集合を表し，$\mu_{\phi,ntd}(\mathbf{C}) \in \mathbb{R}$ と $\sigma^2_{\phi,ntd}(\mathbf{C}) \in \mathbb{R}_+$ は，特徴量 $\mathbf{C}$ を入力とするパラメータ ϕ を持つ DNN の出力である．特徴量 $\mathbf{C}$ は $\mathbf{X}$ と $\mathbf{U}$ から計算されるが，混合音と cACGMM の分離マスクを入力した場合に，最も良い WER になることが報告されている [8]．

変分償却推論では，学習データに対する以下の変分下限 $\mathcal{L}$ を最大化するように，DNN のパラメータ θ と ϕ，SCM $\mathbf{H}_{nf}$ を同時に最適化する．

$$\mathcal{L} = \mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{X}|\mathbf{Z}, \mathbf{H}, \mathbf{U})] - \mathcal{D}_{\mathrm{KL}}[q_\phi(\mathbf{Z}|\mathbf{X}, \mathbf{U})|p(\mathbf{Z})] \tag{5}$$

第一項は，対数尤度の期待値であり，変分自己符号化器 [7] と同様に以下のように近似される．

$$\mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{X}|\mathbf{Z}, \mathbf{H}, \mathbf{U})] \approx -\sum_{f,t} \log |\mathbf{Y}_{:ft}| - \sum_{f,t} \mathbf{x}_{ft}^{\mathsf{H}} \mathbf{Y}_{:ft}^{-1} \mathbf{x}_{ft} \tag{6}$$

ただし，$\mathbf{Y}_{:ft} = \sum_{n=1}^{N} \mathbf{Y}_{nft} \in \mathbb{S}_+^M$ は，各音源ごとの $\mathbf{Y}_{nft} = g_{\theta,nf}(\mathbf{z}_{nt}^*)\mathbf{H}_{nf} \in \mathbb{S}_+$ の和である． (5) 式の第二項は，潜在変数 z_{ntd} が事前分布から離れないように促す．この式の最大化により，パラメータ θ と SCM $\mathbf{H}_{nf}$ は $\log p_\theta(\mathbf{X}|\mathbf{H}, \mathbf{U})$ を最大化するように，パラメータ ϕ は，$\mathcal{D}_{\mathrm{KL}}[q_\phi(\mathbf{Z}|\mathbf{X}, \mathbf{U})|p_\theta(\mathbf{Z}|\mathbf{X}, \mathbf{H}, \mathbf{U})]$ を最小化するように学習される．各 DNN のパラメータは誤差逆伝播法により最適化する．SCM $\mathbf{H}_{nf}$ は，以下の更新則 [10] を繰り返して最適化する．

$$\mathbf{H}_{nf} \leftarrow \mathbf{B}_{nf}^{-\frac{1}{2}} \left(\mathbf{B}_{nf}^{\frac{1}{2}} \mathbf{A}_{nf} \mathbf{B}_{nf}^{\frac{1}{2}}\right)^{\frac{1}{2}} \mathbf{B}_{nf}^{-\frac{1}{2}} \tag{7}$$

$$\mathbf{A}_{nf} \triangleq \mathbf{H}_{nf} \left(\sum_t g_{\theta,nf}(\mathbf{z}_{nt}^*) \mathbf{Y}_{:ft}^{-1} \mathbf{x}_{ft} \mathbf{x}_{ft}^{\mathsf{H}} \mathbf{Y}_{:ft}^{-1}\right) \mathbf{H}_{nf}$$

$$\mathbf{B}_{nf} \triangleq \sum_t g_{\theta,nf}(\mathbf{z}_{nt}^*) \mathbf{Y}_{:ft}^{-1}$$

ただし，$\mathbf{z}_{nt}^* \sim q_\phi(\mathbf{Z}|\mathbf{X}, \mathbf{U})$ は潜在ベクトルのサンプルである．本研究では，θ と ϕ を 1 回更新するごとに $\mathbf{H}$ を 5 回更新する．

3　転移学習を用いた深層BSSの拡張

提案法では，従来の線形 BSS による分離結果を疑似教師として学習されたモデルを，2 章で述べた弱教師あり深層 BSS 音源分離 [8] に転移する．

3.1　疑似教師あり学習

2 章で導入した推論モデル q_ϕ と音源生成モデル $g_{\theta,nf}$ を事前学習する．具体的に，は混合音 $\mathbf{X}$ と発話区間変数 $\mathbf{U}$ を入力とする cACGMM の一種である補助付き音源分離法 (GSS) [3] による混合音の分離結果 $s_{nft} \in \mathbb{C}$ を疑似教師として，ネットワークが s_{nft} を模倣するように学習する．GSS は，発話区間変数でマスクされた cACGMM の対数周辺尤度を EM アルゴリズムにより最大化することで，時間周波数ビンをクラスタリングする手法である．

本研究の疑似教師あり学習では，以下の変分下限 $\mathcal{L}_s$ を最大化する．

$$\mathcal{L}_s = \mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{S}|\mathbf{Z})] - \mathcal{D}_{\mathrm{KL}}[q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{U})|p(\mathbf{Z})] \quad (8)$$

ここで，$\mathbf{S} \triangleq \{s_{nft}\}_{n,f,t=1}^{N,F,T}$ は分離音の集合である．本変分下限の第一項の最大化は，再構成音が疑似教師音 $\mathbf{S}$ に近くなることを意味し，第二項は潜在変数の正規化項である．パラメータ θ は $\log p_\theta(\mathbf{S})$ を最大化するように，パラメータ ϕ は，$\mathcal{D}_{\mathrm{KL}}[q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{U})|p_\theta(\mathbf{Z}|\mathbf{S})]$ を最小化するように学習される．第一項については，疑似教師音 $\mathbf{S}$ が (3) 式のような零平均複素ガウス分布に従うと仮定し，以下のように近似される．

$$\mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{S}|\mathbf{Z})] \approx -\sum_{n,f,t} \log y_{nft} - \sum_{n,f,t} \frac{|s_{nft}|^2}{y_{nft}} \quad (9)$$

ただし，$y_{nft} = g_{\theta,nf}(\mathbf{z}_{nt}^*)$ はネットワークが出力する PSD である．式 (9) は再構成音と疑似教師音の板倉斎藤距離 [11] と等価である．GSS の出力はフルランク型最小分散無歪 (MVDR) ビームフォーマ [12] を用いて得るため，分離音のスケールは不定である．そこで，最尤推定値となる以下のスケールを推定 PSD y_{nft} に乗じて学習した．

$$a_{nf} = \frac{1}{T}\sum_t \frac{|s_{nft}|^2}{y_{nft}} \quad (10)$$

3.2　転移学習

学習済みの推論モデル q_ϕ と 音源生成モデル $g_{\theta,nf}$ は 2.3 章の弱教師あり学習に転移させる．具体的には，学習済みモデルの重みを初期値として重みを更新する．エンコーダに入力する特徴量 $\mathbf{C}$ は，混合音と基準マイクと各マイクの位相差であり，分離結果で疑似教師あり学習した代わりに cACGMM 分離マスクは入力しない．

3.3　推論方法

2 回の学習を経て獲得された音源分離 DNN を使い，未知の混合音を分離する．具体的には，混合音 $\mathbf{X}$ と発話区間変数 $\mathbf{U}$ から特徴量 $\mathbf{C}$ を計算し，$\log p_\theta(\mathbf{X}|\mathbf{H},\mathbf{U},\mathbf{Z})$ を最大化するように PSD $g_{\theta,nf}(\boldsymbol{\mu}_{\phi,nt}(\mathbf{C}))$ と SCM $\mathbf{H}_{nf}$ を推定する [6]．分離音 $\hat{s}_{nft}$ は，PSD と SCM から，MVDR ビームフォーマを用いて得る．

4　評価実験

CHiME-6 データセットで提供されている実収録音を用いて提案法を評価した．

4.1　実験設定

CHiME-6 データセットは，複数の家庭で行われたディナーパーティでの音声を記録したもので，各パーティには 4 人の参加者がいる．kitchen，dining，living からなる室内で 5 または 6 台の 4 チャネルマイクアレイ (Microsoft Kinect v2) で収録され，一つのエリアに少なくとも 2 つのマイクアレイが設置されている．train，dev 及び eval セットに分割され，収録時間はそれぞれ 40 時間 33 分，4 時間 27 分及び 5 時間 12 分である．各録音は 16 kHz で収録されている．

提案法のネットワークアーキテクチャは [8] と同一の設計とした．スペクトログラムは短時間フーリエ変換によって求め，窓長 1024，ホップ長 256 とした．音源の数 N_{spk} は 4，潜在変数 D_{spk} は 50 次元とし，雑音源の数 N_{noi} は 2，潜在変数 D_{noi} が 20 次元とした．式 (5) と (8) の KL 項の重みを周期的に変動させる KL アニーリング [6] を行った．学習は，200 エポックとした．計算量を減らすため，混合信号の全 24 チャンネルのうち，最もパワーの大きい 12 チャンネルで学習を行った．推論モデルに入力する特徴量 $\mathbf{C}$ は，提案法では，基準マイクロホンと他のマイクロホン間のチャネル間位相差と，混合音の対数パワースペクトログラムである．

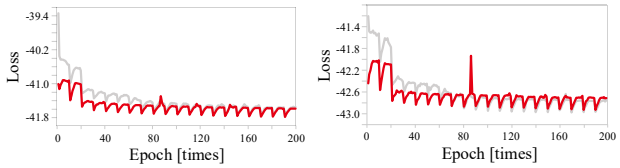
提案法 (WS Neural FCA + 転移学習) は，CHiME-6 Challenge のベースライン音声認識器 [2] を用いて単語誤り率 (WER) で評価した．ベースラインとして，GSS および，GSS を特徴量として用いた場合 (cACGMM → WS Neural FCA) ，事前学習を行わなかった場合 (WS Neural FCA) を評価した．

4.2　実験結果

表 1 に音声認識性能を WER で示す．転移学習した場合，しない場合と比較して 200 エポック目において dev セットで 1.5 pt，eval セットで 0.6 pt 性能が向上

表 1: CHiME-6 データセットの dev set および eval set における WER.

手法	Epoch	Dev set				Eval set			
		Avg.	Dining	Kitchen	Living	Avg.	Dining	Kitchen	Living
GSS (公式実装)	–	51.8	53.8	53.9	48.6	51.3	44.7	61.2	50.3
GSS ($M = 16$) [8]	–	49.8	51.6	52.3	46.4	51.1	45.0	60.8	49.7
GSS → WS Neural FCA [8]	200	48.6	51.2	50.8	45.1	49.0	43.2	56.7	48.9
WS Neural FCA	50	54.8	56.0	58.1	51.0	52.7	45.7	60.2	54.1
WS Neural FCA	100	55.7	56.7	59.6	51.6	52.9	45.7	60.6	54.5
WS Neural FCA	200	55.3	56.2	58.9	51.3	52.9	46.0	60.4	54.1
WS Neural FCA + 転移学習	50	54.1	55.4	57.9	49.9	52.4	45.3	60.3	53.5
WS Neural FCA + 転移学習	100	54.3	55.1	58.1	50.2	52.4	45.2	60.4	53.6
WS Neural FCA + 転移学習	200	53.8	54.8	57.5	49.7	52.3	45.0	60.0	53.8



(a) 学習損失　(b) 検証損失

図 1: 転移学習した場合 (赤) としない場合 (灰) の学習および検証データに対する損失関数.

した．また，図 1 に示す通り，転移学習しない場合は損失関数の収束に 100 エポック程度要しているのに対して，した場合は 50 エポック程度で収束している．

一方，GSS や GSS → WS Neural FCA の結果と比較すると，提案法には改善の余地がある．本稿では，疑似教師の非歪性を重視し，MVDR ビームフォーマの結果を用いたが，ビームフォーマでは妨害音の抑制に限界がある．音源モデルの事前学習においては，MVDR ではなく，時間周波数マスキングの結果を用いた方が性能改善に寄与する可能性がある．また，収束速度は早くなっているが，収束先は転移学習していない場合と大きな差がなく (図 1)，本稿で用いた入力特徴量では効果的な学習が困難な可能性がある．今後は，容易に計算できる学習しやすい特徴量の設計を進める．

5　おわりに

本稿では，従来の BSS の分離結果を用いて疑似教師あり学習したモデルを，弱教師あり Neural FCA に転移学習する枠組みについて述べた．転移学習した場合，しなかった場合と比較して WER がわずかに改善し，損失関数の収束が早いことが示された．WER の大きな改善につながらなかったため，事前学習に用いる疑似教師データや推論モデルへの入力特徴量の改善を進める．

謝辞 本研究の一部は，NEDO および JST ACT-X 数理・情報のフロンティア JPMJAX200N の支援を受けた．

参考文献

[1] K. Shimada *et al.*, “Unsupervised speech enhancement based on multichannel NMF-informed beamforming for noise-robust automatic speech recognition,” *IEEE/ACM TASLP*, vol. 27, no. 5, pp. 960–971, 2019.

[2] S. Watanabe *et al.*, “CHiME-6 Challenge: Tackling multispeaker speech recognition for unsegmented recordings,” in *CHiME 2020 Workshop*, 2020, pp. 1–7.

[3] C. Boeddeker *et al.*, “Front-end processing for the CHiME-5 dinner party scenario,” in *CHiME5 Workshop*, 2018, pp. 1–6.

[4] K. Sekiguchi *et al.*, “Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation,” *IEEE/ACM TASLP*, vol. 28, pp. 2610–2625, 2020.

[5] N. Ito *et al.*, “Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing,” in *2016 24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1153–1157.

[6] Y. Bando *et al.*, “Neural full-rank spatial covariance analysis for blind source separation,” *IEEE SPL*, vol. 28, pp. 1670–1674, 2021.

[7] D. P. Kingma *et al.*, “Auto-encoding variational Bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

[8] Y. Bando *et al.*, “Weakly-Supervised Neural Full-Rank Spatial Covariance Analysis for a Front-End System of Distant Speech Recognition,” in *Proc. Interspeech 2022*, 2022, pp. 3824–3828.

[9] M. Togami *et al.*, “Unsupervised training for deep speech source separation with kullback-leibler divergence based probabilistic loss function,” in *IEEE ICASSP*, 2020, pp. 56–60.

[10] K. Yoshii, “Correlated tensor factorization for audio source separation,” in *IEEe ICASSP 2018*, 2018, pp. 731–735.

[11] F. Itakura, “Analysis synthesis telephony based on the maximum likelihood method,” *ICA*, 1968.

[12] M. Souden *et al.*, “On optimal frequency-domain multichannel linear filtering for noise reduction,” *IEEE TASLP*, vol. 18, no. 2, pp. 260–276, 2010.