Vol.2019-SLP-128 No.6
2019/7/19

DPGMMと敵対的学習に基づく話者の違いに頑健な特徴抽出とゼロリソース音声認識での評価

樋口 陽祐[1]　俵 直弘[2]　小林 哲則[1]　小川 哲司[1]

概要：ディリクレ過程ガウス混合モデル (Dirichlet process Gaussian mixture model; DPGMM) により教師なしの枠組みで音韻に関する情報を取得し，敵対的マルチタスク学習により話者補正を行うことで，ゼロリソース音声認識に適した特徴表現の獲得を試みる．ゼロリソース言語は音素ラベルが付与されていないため，DPGMM の各クラスタが音素に対応すると期待する．しかし，同一の音韻であっても発話者の違いによりばらつきが生じるため，DPGMM のように音声信号のみからデータドリブンで生成されるクラスタは必ずしも音素に対応するとは限らず，話者によるクラスタが形成されている可能性もある．そこで，DPGMM のクラスタの事後確率分布を音素に関する教師としながら，話者に関する敵対的損失を出力層に導入することで，発話者の違いの影響が抑圧された音素に関する事後分布を生成するニューラルネットワークを構築することを試みる．こうして得たネットワークからフレーム単位で得られる音素事後確率ベクトルを話者の違いに頑健な特徴量として利用したところ，Zero Resource Speech Challenge データにおいて，話者情報を効果的に抑圧し，音素に識別的な特徴抽出が行えることを確認した．

キーワード：音声認識，ゼロリソース言語，ディリクレ過程ガウス混合モデル，敵対的マルチタスク学習

Speaker Adversarial Training of DPGMM-based Feature Extractor for Zero-Resource Languages

Yosuke Higuchi[1]　Naohiro Tawara[2]　Tetsunori Kobayashi[1]　Tetsuji Ogawa[1]

1. はじめに

多くの音声処理技術は，人手でラベル付けされた大量のデータを必要する．それに対し，書き起こし (教師情報) の存在しないゼロリソース言語を対象とした音声処理技術が注目を集めている．例えば，発話中のクエリ検出 [1], [2]，音素のサブワード単位獲得 [3], [4]，トピックセグメンテーション [5]，文書分類 [6] などが，ゼロリソース言語を対象として試みられている．本研究では，ゼロリソース言語を対象とした音声認識に焦点を当てる．

ゼロリソース言語の音声認識システムを構築するためには，音声データのみから事前情報無しに音韻の識別に寄与する情報を獲得する必要がある．音素表現の獲得として，深層学習 (DNN) を用いた学習法が提案されており，多様体学習 [7]，自己符号化器 [8], [9]，多言語によるボトルネック特徴量の学習 [10], [11] などのアプローチが検討されている．その中でも，音響特徴量を用いて生成される DPGMM の事後確率ベクトルを，音素に識別的な特徴量として利用することの有効性が広く知られている [12]．しかし，音響特徴量は発話者や周辺雑音といった変動要因を含み，データドリブンに生成されるクラスタは音韻以外の情報を含む可能性がある．その結果，DPGMM 事後確率ベクトルは音韻情報としての信頼性が低くなり，このままでは高精度な音響モデルを構築するのが困難であるため，音韻の識別を阻害する情報を抑圧する必要がある．

音声信号に含まれる変動要因を取り除く手法として，声

[1] 早稲田大学
Waseda University
[2] 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 Communication Science Laboratries, NTT Corporation

道長正規化 (vocal tract length normalization; VTLN) [13] や特徴空間における最尤線形回帰 (feature-space maximum likelihood linear regression; fMLLR) [14] が広く用いられる．これらの手法では，観測される音素系列に対する音響モデルの尤度が最大となるように特徴量の変換を行う．また，敵対的学習を適用した DNN のマルチタスク学習 [15], [16] を用いて，特定のドメインに不変なボトルネック特徴量を抽出する手法が検討されている．例えば [17] では，音素クラスの識別を行う DNN の中間層に対し，雑音の種類を識別するネットワークを敵対的に導入することで，ボトルネック特徴量に含まれる雑音識別に寄与する情報を抑圧している．同様の枠組みで，話者の違いに頑健な特徴量の学習法も提案されている [18], [19]．これらの手法では，識別対象である音素に加え，抑圧対象である話者情報のラベルが付与されているデータが用いられていることに注意すべきである．

一方，書き起こし文が存在しないゼロリソース言語に対しては，既存の特徴補正方法をそのまま適用できない．それに対し本研究では，DPGMM の事後確率分布を音韻に関する教師情報として用いる．しかし，音声データのみから生成された DPGMM の各クラスタは発話者の違いなどの影響を受け，正確に音素に対応しているとは限らず，発話者の情報でクラスタが形成されている可能性すらある．そこで，話者情報に関する敵対的損失を効果的に導入することで，書き起こし情報無しに，識別的な音響特徴である音素事後確率から雑音である発話者の影響を抑圧するような特徴抽出を行うことを試みる．具体的には，ニューラルネットワークの出力を DPGMM のクラスタの事後分布に近づけることで音韻に関して識別的な特徴を抽出しつつ，敵対的損失を導入することで発話者の影響を抑圧する．

本稿の構成は以下の通りである．まず，2 章で提案する特徴抽出モデルについて説明し，3 章で音素識別実験による提案法より得られる特徴量の評価を行う．最後に，4 章で本稿のまとめと今後の課題について述べる．

2. 特徴抽出

本研究では，敵対的マルチタスク (adversarial multi-task; AMT) 学習により得たニューラルネットワークの中間層，もしくは出力層の情報を，ゼロリソース言語音声認識のための話者の違いに頑健な特徴量とした用いる．提案する特徴抽出用ネットワークを構築する枠組みを図 1 に示す．また，その構築過程は以下の通りである．

(1) 音声データから音響特徴量（例えば，メル周波数ケプストラム係数ベクトルに fMLLR を適用したもの）を抽出する．
(2) 音響特徴量を用いて DPGMM を生成する．
(3) 音響特徴量の各フレームについて，DPGMM を構成する各クラスタに対する事後確率を算出する．
(4) DPGMM の事後確率ベクトルを音韻に関する教師情報とし，別途与えられた話者ラベルも用いて敵対的ネットワークを学習する．

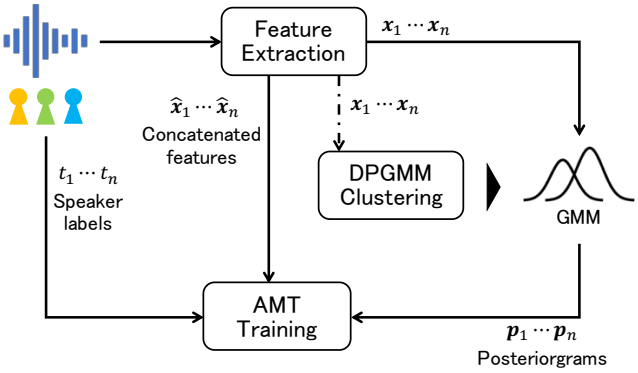

図 1　特徴抽出用ネットワークの構築法．

このようにして得た敵対的ネットワークから抽出される特徴量は，話者の違いに頑健であることが期待できる．次節以降では，DPGMM クラスタリングによる音素情報の取得と，敵対的マルチタスク学習による話者補正について概説する．

2.1 DPGMM を用いた音素情報の獲得

ゼロリソース言語音声から教師なしの枠組みで音響モデルを構築する手法として，DPGMM の有効性が知られている [12]．DPGMM はノンパラメトリックベイズ学習により推定される GMM であり，入力データに対し，パラメータのみならず混合数の最適化も行えることが特徴である．DPGMM の各クラスタがサブワード単位に対応すると仮定すると，各フレームの音響特徴量に対する DPGMM の各クラスタの事後確率は，音韻を識別するのに有効な特徴量として扱うことができる．

音声データの i フレーム目における k 番目のクラスタに対する事後確率は以下のように算出される．

$$p_{i,k} = p(c_k|\boldsymbol{x}_i) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \quad (1)$$

ここで，$X = \{\boldsymbol{x}_i\}_{i=1}^{N}$ は音響特徴量，K は混合数，$\boldsymbol{\pi} = \{\pi_k\}_{k=1}^{K}$ は k 番目の混合要素の混合重み，$\boldsymbol{\mu} = \{\boldsymbol{\mu}_k\}_{k=1}^{K}$ は k 番目の混合要素の平均ベクトル，$\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_k\}_{k=1}^{K}$ は k 番目の混合要素の分散行列を各々表す．式 (1) より，i フレーム目の音響特徴量 $\boldsymbol{x}_i$ に対する DPGMM 事後確率ベクトルは以下のように示される．

$$P(\boldsymbol{x}_i) = \{p_{i,1}, ..., p_{i,K}\} \quad (i = 1, ..., N). \quad (2)$$

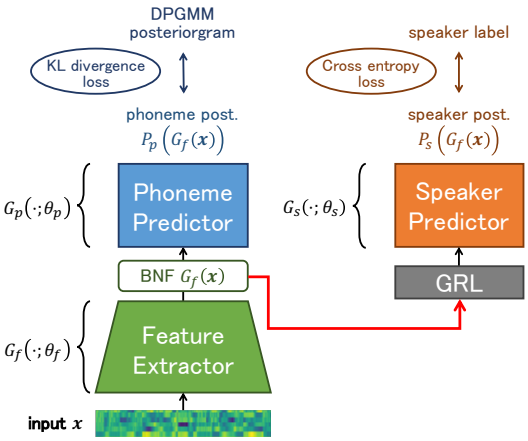

図 2　DPGMM 事後確率ベクトルと話者ラベル情報を用いた敵対的マルチタスク学習によりボトルネック特徴量を話者補正するネットワークの構造（従来方式）

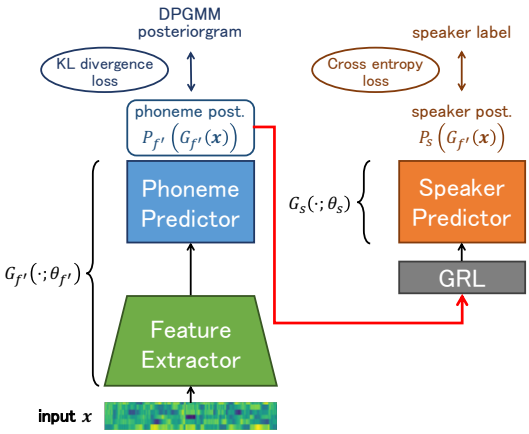

図 3　DPGMM 事後確率ベクトルそのものから話者情報を抑圧するネットワーク構造（提案方式）

2.2　敵対的マルチタスク学習によるボトルネック特徴量の話者補正 (従来方式)

音素を識別する DNN の中間層に対し，話者に関する敵対的損失を導入することで，ボトルネック特徴量に含まれる発話者の情報を抑圧する試みがなされている [18], [19]．このモデルは，特徴抽出器，音素推定器，話者推定器の 3 つのネットワークから構成される．特徴抽出器は入力特徴量をボトルネック特徴量に変換し，音素推定器と話者推定器は，ボトルネック特徴量から各々の目的クラスに対する事後確率の計算を行う．ネットワークの学習において，2 つの推定器はそれぞれの識別損失を最小化するようにパラメータの更新を行う．また，特徴抽出器は，伝搬される 2 つの識別損失より，音素に対する識別精度を最大化しつつ，話者に対する識別精度を最小化するように学習を行う．このような話者に敵対的なマルチタスク学習により得られる特徴抽出器から，音素に関して識別的かつ，話者の違いに頑健な特徴表現を得る．

このネットワークの学習には，音声データに対して音素と話者に関するラベルが付与されていることを前提としている．しかし，ゼロリソース言語音声については，話者は容易に区別できる一方で，音素ラベルは一切付与されていない．そこで，本研究では，DPGMM を構成する各クラスタの事後確率を，音素推定器が出力すべき確率分布として利用する．図 2 に，DPGMM 事後確率ベクトルと話者ラベル情報を用いた敵対的マルチタスク学習によりボトルネック特徴量を話者補正するネットワークの構造を示す．音素推定器における損失は，音素事後確率分布と DPGMM 事後確率ベクトルの KL ダイバージェンスより算出される．$\boldsymbol{\theta}_f, \boldsymbol{\theta}_p, \boldsymbol{\theta}_s$ をそれぞれ特徴抽出器 G_f，音素推定器 G_p，話者推定器 G_s のパラメータとする．音素推定器の識別損失 $\mathcal{L}_p$ は以下より算出される．

$$\mathcal{L}_p(\boldsymbol{\theta}_f, \boldsymbol{\theta}_p) = D_{KL}\big(P(\boldsymbol{x})||P_p(G_f(\boldsymbol{x};\boldsymbol{\theta}_f);\boldsymbol{\theta}_p)\big). \quad (3)$$

ここで，$P(\cdot)$ は式 (2) から得られる DPGMM 事後確率ベクトルであり，$P_p(\cdot)$ は音素推定器の事後確率分布である．話者推定器の識別損失 $\mathcal{L}_s$ は，クロスエントロピーを用いて以下で定義される．

$$\mathcal{L}_s(\boldsymbol{\theta}_f, \boldsymbol{\theta}_s) = -\mathbb{E}_{t\sim p(t|\boldsymbol{x})}\big[\log P_s(t|G_f(\boldsymbol{x};\boldsymbol{\theta}_f);\boldsymbol{\theta}_s)\big]. \quad (4)$$

ここで，$P_s(\cdot)$ は話者推定器の事後確率分布であり，$t \sim p(t|\boldsymbol{x})$ は，入力特徴量 $\boldsymbol{x}$ に対する話者ラベル t を示す．

ミニバッチ学習を用いた確率的勾配降下法 (SGD) により，ネットワークの重みは以下より更新される．

$$\boldsymbol{\theta}_p \leftarrow \boldsymbol{\theta}_p - \mu\frac{\partial\mathcal{L}_p}{\partial\boldsymbol{\theta}_p}, \quad (5)$$

$$\boldsymbol{\theta}_s \leftarrow \boldsymbol{\theta}_s - \mu\frac{\partial\mathcal{L}_s}{\partial\boldsymbol{\theta}_s}, \quad (6)$$

$$\boldsymbol{\theta}_f \leftarrow \boldsymbol{\theta}_f - \mu\left(\frac{\partial\mathcal{L}_p}{\partial\boldsymbol{\theta}_f} - \lambda\frac{\partial\mathcal{L}_s}{\partial\boldsymbol{\theta}_f}\right). \quad (7)$$

ここで，μ は学習率である．また，λ は Gradient Reversal Layer (GRL) [15], [16] のパラメータであり，話者の識別に対して敵対させる仕組みである．GRL は順伝搬の際には恒等関数であり，逆伝搬の際には損失に対して λ を乗算し，符号を反転させる．

2.3　敵対的マルチタスク学習による DPGMM 事後確率ベクトルの話者補正 (提案方式)

発話者の違いにより同一の音韻であってもばらつきが生じる．そのため，音響特徴量を用いて生成した DPGMM のクラスタは，必ずしも音素もしくはサブワード単位に対応するとは限らず，話者の単位でクラスタが形成される可能性もある．その場合，DPGMM 事後確率ベクトルを音素に関する教師情報として用いるのは妥当でない．この問題を解決するために，生成した DPGMM 事後確率ベクトルそのものから話者情報を抑圧するネットワークを提案する．図 3 に，提案する敵対的ネットワークの構造を示す．特徴抽出器と音素推定器を統合し，入力特徴量から直接，

表 1 訓練データ

Language	#speakers	dur. / speaker	Total
Xitsonga	20	3-25 min	314 min
English	9	165-220 min	1695 min
French	10	110-195 min	1334 min
Mandarin	12	10-25 min	156 min

音素推定器の出力確率分布を計算することで，DPGMM 事後確率ベクトルの近似を行うネットワークとする．話者推定器は近似された事後確率ベクトルを入力とし，話者クラスに対する事後確率分布を計算する．DPGMM 事後確率ベクトルに含まれる音韻情報を保ちつつ，話者識別結果に敵対するように学習を行うことで，DPGMM 事後確率ベクトルに含まれる話者識別に寄与する情報が抑圧されることを期待する．

$\boldsymbol{\theta}_{f'}, \boldsymbol{\theta}_s$ をそれぞれ変更を加えた特徴抽出器 $G_{f'}$，話者推定器 G_s のパラメータとする．特徴抽出器の DPGMM 事後確率ベクトルに対する近似結果の損失 $\mathcal{L}_{f'}$ は，KL ダイバージェンスを用いて以下で定義される．

$$\mathcal{L}_{f'}(\boldsymbol{\theta}_{f'}) = D_{KL}\big(P(\boldsymbol{x})||P_{f'}(\boldsymbol{x};\boldsymbol{\theta}_{f'})\big). \tag{8}$$

ここで，$P_{f'}(\cdot)$ は特徴抽出器が出力する事後確率分布である．話者推定器の識別損失 $\mathcal{L}_s$ は，クロスエントロピーを用いて以下の通り算出される．

$$\mathcal{L}_s(\boldsymbol{\theta}_{f'}, \boldsymbol{\theta}_s) = -\mathbb{E}_{t\sim p(t|\boldsymbol{x})}\big[\log P_s(t|P_{f'}(\boldsymbol{x};\boldsymbol{\theta}_{f'});\boldsymbol{\theta}_s)\big]. \tag{9}$$

式 (5) ∼ (7) と同様の条件で，ネットワークのパラメータは以下のように更新される．

$$\boldsymbol{\theta}_s \leftarrow \boldsymbol{\theta}_s - \mu\frac{\partial\mathcal{L}_s}{\partial\boldsymbol{\theta}_s}, \tag{10}$$

$$\boldsymbol{\theta}_{f'} \leftarrow \boldsymbol{\theta}_{f'} - \mu\left(\frac{\partial\mathcal{L}_{f'}}{\partial\boldsymbol{\theta}_{f'}} - \lambda\frac{\partial\mathcal{L}_s}{\partial\boldsymbol{\theta}_{f'}}\right). \tag{11}$$

3. 音素織別実験

DPGMM により獲得したゼロリソース言語の音韻情報を用いて，敵対的マルチタスク学習による話者補正を行った．このとき，以下の特徴量について比較を行った．

- **fMLLR**: fMLLR 特徴量
- **DPGMM-post**: DPGMM の事後確率ベクトル [12]
- **AMT-BNF**: DPGMM-post を用いて学習した敵対的ネットワークのボトルネック特徴量 [19]
- **AMT-post**: DPGMM-post を用いて学習した敵対的ネットワークの事後確率ベクトル（提案方式）

評価は，Zero Resource Speech Challenge 2015, 2017 (ZSC2015, 2017) の track1 [20], [21] に基づいた音素識別実験により行った．

3.1 実験データ

実験には，ZSC2015 で指定されるツォンガ語 (Xitsonga)，ZSC2017 で提供される英語 (English)，フランス語 (French)，中国語 (Mandarin) を用いた．各言語において音声データは発話者ごとに用意される．表 1 に各言語における訓練データの詳細を示す．ツォンガ語のテストデータは計 149 分であった．また，英語，フランス語，中国語に対しては各発話が 120 秒のテストデータを利用し，発話長はそれぞれ計 1634 分，1061 分，1522 分であった．

音響特徴量の抽出には，Kaldi 音声認識ツールキット [22] を用いた．フレーム長 25 ms，フレーム周期 10 ms で 13 次元のメルケプストラム係数 (MFCCs) とその動的特徴量 ($\Delta+\Delta\Delta$) を抽出し，ケプストラム平均分散正規化をセグメント単位で適用した．また，前処理における音響特徴量の話者補正として，fMLLR による変換を行い，40 次元の音響特徴量を得た．変換行列の学習には，Kaldi の TIMIT [23] レシピにより学習した英語音声の音響モデルを用いた．

3.2 評価指標

提案手法より得られる特徴量の性能を確認するために，ABX テストの誤り率 [24], [25] に基づいた評価を行った．同じ音素カテゴリに含まれる特徴量系列間の距離が，異なる音素カテゴリの特徴量系列との距離より近くなるかを評価する．任意の特徴系列対 $\boldsymbol{x}, \boldsymbol{y}$ 間の ABX 誤り率は，次式より算出される．

$$\theta(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{m(m-1)n}\sum_{a\in S(\boldsymbol{x})}\sum_{b\in S(\boldsymbol{y})}\sum_{x\in S(\boldsymbol{x})\setminus\{a\}} \left(\mathbb{1}_{d(a,x)<d(b,x)} + \frac{1}{2}\mathbb{1}_{d(a,x)=d(b,x)}\right). \tag{12}$$

ここで，m と n はそれぞれ音素カテゴリ $S(\boldsymbol{x}), S(\boldsymbol{y})$ に属するサンプル数であり，$\mathbb{1}$ は指示関数である．2 つの音声サンプル $\boldsymbol{x}, \boldsymbol{y}$ の距離 $d(\boldsymbol{x}, \boldsymbol{y})$ は，動的時間伸縮法により算出される．距離尺度には fMLLR 特徴量とボトルネック特徴量に対してはコサイン距離，確率分布に対しては KL ダイバージェンスを用いた．本実験では，特徴量の音素識別性能と話者非依存性を評価するために，単一話者の発話のみと，複数話者の発話が混合するデータセットに対して ABX テストを行った．

3.3 実験条件

DPGMM によるクラスタリングには， [9], [10] と同様に，Chen らが公開しているツール[*1]を使用した．事前分布のパラメータは規定のものを利用し，集中度 α は 1 とした．DPGMM の推論は 1500 回行い，対象言語において混合数が収束するのを確認した．ツォンガ語，英語，フランス語，中国語における混合数は，それぞれ 510, 1880, 1623, 510 であった．

敵対的ネットワークの入力は，fMLLR 特徴量の前後 5

[*1] http://people.csail.mit.edu/jchang7/code.php

表 2　単一話者内の ABX 誤り率.

Feature	Xitsonga	English	French	Mandarin
fMLLR	17.42	6.85	8.96	8.74
DPGMM-post	9.19	6.35	9.12	9.75
AMT-BNF	13.98	5.94	8.16	**8.22**
AMT-post	**8.41**	**5.88**	**8.09**	10.06

表 3　複数話者間の ABX 誤り率.

Feature	Xitsonga	English	French	Mandarin
fMLLR	25.70	10.83	14.83	10.35
DPGMM-post	14.00	8.77	12.28	9.46
AMT-BNF	19.68	8.51	12.04	**8.95**
AMT-post	**12.59**	**8.18**	**11.37**	9.55

フレームを連結して得た 440 次元のベクトルとした．従来の敵対的ネットワーク (**AMT-BNF**) を構成する，特徴抽出器，音素推定器，話者推定器のネットワーク構造は各々，{440 - 1024 - 1024 - 1024 - 1024}，{1024 - K}，{512 - C} とした．ここで，K は DPGMM の混合数であり，C は話者クラス数である．提案の敵対的ネットワーク (**AMT-post**) を構成する，特徴抽出器，話者推定器のネットワーク構造は各々，{440 - 1024 - 1024 - 1024 - 1024 - 1024 - K}，{512 - C} とした．両ネットワークの全中間層に活性化関数として ReLU 関数を使用し，過学習を防ぐためにドロップアウトを適用した．ドロップアウト率はすべて 0.2 とした．すべてのモデルは，ミニバッチサイズ 1024，学習率 0.01 で SGD により最適化を行った．GRL のパラメータ λ は，以下の式に基づきスケジューリングを行った．

$$\lambda = \lambda_{\max} \left\{ \frac{2}{1 + \exp(-\gamma p)} - 1 \right\}. \qquad (13)$$

ここで，p は学習の経過率を表し，エポックが進むに従い 0 から 1 まで線形に増加させた．γ は λ が最大値 $\lambda_{\max}$ に収束するまでの速度を調整するパラメータで，[16] と同様に 10 とした．事前実験の結果，**AMT-BNF** におけるツォンガ語，英語，フランス語，中国語の $\lambda_{\max}$ は，それぞれ 1.0, 9.0, 9.0, 1.0 とした．また，**AMT-post** におけるツォンガ語，英語，フランス語，中国語の $\lambda_{\max}$ は，それぞれ 50.0, 5.0, 7.0, 9.0 とした．

3.4　実験結果

単一話者内と複数話者間における ABX スコアの結果を表 2 および表 3 に示す．**fMLLR** と **DPGMM-post** に比べ，敵対的マルチタスク学習 (**AMT**) では誤り率を削減できていることから，提案の枠組みにより特徴量を抽出することで，話者の違いによる影響の低減が期待できる．2 種の敵対的マルチタスク学習の結果を比較すると，ツォンガ語，英語，フランス語において，**DPGMM-post** は **DPGMM-BNF** よりも良好な性能を与えた．特にツォンガ語に関しては高い削減率で誤りを低減しており，**AMT-post** により対象言語がゼロリソース言語だけでなくリソース豊富な言語に対しても，効果的に話者補正を行えることが確認できた．しかし，中国語に関しては，**DPGMM-BNF** の方が **DPGMM-post** よりも良好なスコアを与えた．中国語の音声には特有の性質があり，音の高さの時間的変化により区別される音素が存在する [26]．本実験では，フレーム単位の音響特徴量により DPGMM の生成を行ったため，DPGMM 事後確率ベクトルにはフレーム前後のコンテキスト情報が考慮されていない．実際に，**fMLLR** に対して **DPGMM** を適用すると，性能が劣化していることが結果から分かる．そのため，中国語に対しては，**DPGMM-post** による話者補正の効果が得られなかったと考えられる．

4.　まとめと今後の課題

ゼロリソース言語音声に対し，話者の違いに頑健な特徴量を抽出するための手法を提案した．DPGMM により音韻に関する情報を獲得し，敵対的マルチタスク学習により話者情報を抑圧することで，話者の違いに非依存な特徴量の学習を行った．発話者の違いにより，音響特徴量から生成される DPGMM には，音素のサブワード単位だけでなく，話者単位に相当するクラスタが含まれる可能性がある．これに対し，従来のボトルネック特徴量を対象とした敵対的ネットワークを改変し，DPGMM 事後確率ベクトルそのものから話者性を抑圧するネットワークを提案した．音素織別実験における ABX テストによる評価の結果，提案の特徴抽出法により話者の違いの影響を低減できることを示した．また，従来法よりも提案の敵対的ネットワークの方が，効果的に話者補正が行えることを確認した．

今後の予定として，ゼロリソース言語に対してより洗練された DPGMM を生成する手法 [27] を，本枠組みに取り入れることを検討している．また，中国語音声に適した特徴抽出や，話者性を効果的に抽出するために，複数フレームの情報を考慮できる構造をモデルに導入することを考えている．

謝辞　本研究は科研費 (課題番号 17K12718) の助成を受けたものである．

参考文献

[1] Zhang, Y. and Glass, J. R.: Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams, *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (2009).

[2] Mantena, G. and Prahallad, K.: Use of articulatory bottle-neck features for query-by-example spoken term detection in low resource scenarios, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2014).

[3] Lee, C.-y. and Glass, J.: A nonparametric Bayesian approach to acoustic model discovery, *Proceedings of the Association for Computational Linguistics (ACL)*

(2012).
[4] Ondel, L., Godard, P., Besacier, L., Larsen, E., Hasegawa-Johnson, M., Scharenborg, O., Dupoux, E., Burget, L., Yvon, F. and Khudanpur, S.: Bayesian Models for Unit Discovery on a Very Low Resource Language, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2018).
[5] Malioutov, I., Park, A., Barzilay, R. and Glass, J.: Making sense of sound: Unsupervised topic segmentation over acoustic input, *Proceedings of the Association of Computational Linguistics (ACL)* (2007).
[6] Dredze, M., Jansen, A., Coppersmith, G. and Church, K.: NLP on spoken documents without ASR, *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)* (2010).
[7] Thiolliere, R., Dunbar, E., Synnaeve, G., Versteegh, M. and Dupoux, E.: A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling, *Proceedings of the INTERSPEECH* (2015).
[8] Badino, L., Mereta, A. and Rosasco, L.: Discovering discrete subword units with binarized autoencoders and hidden-Markov-model encoders, *Proceedings of the INTERSPEECH* (2015).
[9] Kamper, H., Elsner, M., Jansen, A. and Goldwater, S.: Unsupervised neural network based feature extraction using weak top-down constraints, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2015).
[10] Chen, H., Leung, C.-C., Xie, L., Ma, B. and Li, H.: Multilingual bottle-neck feature learning from untranscribed speech, *Proceedings of the IEEE International Workshop on Automatic Speech Recognition and Understanding (ASRU)* (2017).
[11] Hermann, E. and Goldwater, S.: Multilingual Bottleneck Features for Subword Modeling in Zero-resource Languages, *Proceedings of the INTERSPEECH* (2018).
[12] Chen, H., Leung, C.-C., Xie, L., Ma, B. and Li, H.: Parallel inference of Dirichlet process Gaussian mixture models for unsupervised acoustic modeling: A feasibility study, *Proceedings of the INTERSPEECH* (2015).
[13] Pitz, M. and Ney, H.: Vocal tract normalization as linear transformation of MFCC, *Proceedings of the EUROSPEECH* (2003).
[14] Gales, M. J.: Maximum likelihood linear transformations for HMM-based speech recognition, *Computer speech and language*, Vol. 12, No. 2, pp. 75–98 (1998).
[15] Ganin, Y. and Lempitsky, V.: Unsupervised Domain Adaptation by Backpropagation, *Proceedings of the International Conference on Machine Learning (ICML)* (2015).
[16] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M. and Lempitsky, V.: Domain-adversarial training of neural networks, *Journal of Machine Learning Research*, Vol. 17, No. 59, pp. 1–35 (2016).
[17] Shinohara, Y.: Adversarial Multi-Task Learning of Deep Neural Networks for Robust Speech Recognition., *Proceedings of the INTERSPEECH* (2016).
[18] Meng, Z., Li, J., Chen, Z., Zhao, Y., Mazalov, V., Gang, Y. and Juang, B.-H.: Speaker-invariant training via adversarial learning, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2018).
[19] Tsuchiya, T., Tawara, N., Ogawa, T. and Kobayashi, T.: Speaker Invariant Feature Extraction for Zero-Resource Languages with Adversarial Learning, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2018).
[20] Versteegh, M., Thiolliere, R., Schatz, T., Cao, X. N., Anguera, X., Jansen, A. and Dupoux, E.: The zero resource speech challenge 2015, *Proceedings of the INTERSPEECH* (2015).
[21] Dunbar, E., Cao, X. N., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., Anguera, X. and Dupoux, E.: The zero resource speech challenge 2017, *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (2017).
[22] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P. et al.: The Kaldi speech recognition toolkit, *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (2011).
[23] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G. and Pallett, D. S.: DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1, *NASA STI/Recon Technical Report N*, Vol. 93 (1993).
[24] Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H. and Dupoux, E.: Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline, *Proceedings of the INTERSPEECH* (2013).
[25] Schatz, T., Peddinti, V., Cao, X.-N., Bach, F., Hermansky, H. and Dupoux, E.: Evaluating speech features with the Minimal-Pair ABX task (II): Resistance to noise, *Proceedings of the INTERSPEECH* (2014).
[26] Suen, C. Y.: Computational analysis of Mandarin sounds with reference to the English language, *Proceedings of the Computational Linguistics (COLING)* (1982).
[27] Heck, M., Sakti, S. and Nakamura, S.: Feature optimized DPGMM clustering for unsupervised subword modeling: A contribution to zerospeech 2017, *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (2017).