# DNN 事後確率系列の言語モデル化に基づく言語識別

増村 亮 †, Sheri Sever‡, 浅見 太一 †, 政瀧 浩和,† 阪内 澄宇 †
† 日本電信電話株式会社 NTT メディアインテリジェンス研究所
‡ ウォータールー大学 コンピュータサイエンス研究科

†{masumura.ryo, asami.taichi, masataki.hirokazu,
sakauchi.sumitaka}@lab.ntt.co.jp, ‡sheri.sever@gmail.com

## 1 はじめに

近年のグローバル化に伴い，多言語音声認識，多言語音声対話といったアプリケーションが注目されてきている．多言語を扱う音声アプリケーションにおいては，入力音声がどの言語であるかを特定する言語識別技術が重要と言える [1, 2]．入力音声の言語を正確に特定できれば，特定した言語に適した音声認識，および自然言語処理を行うことができる．また特定した言語に適した処理を実施できない場合は，言語識別の時点でタスク外の判定を行うことが可能となる．

これまで音声の言語識別のために様々な手法が提案されているが，本稿では近年非常に高い性能を実現した Deep Neural Network (DNN) に基づく枠組みに着目する [3]．この枠組みでは，短時間の音声フレーム単位で言語を推定する DNN をモデル化することで，言語識別を実現している．DNN に基づく枠組みは，これまでの GMM や i-vector に基づく枠組み [4] と比較して非常に高い性能を示している．

しかしながら従来法は，音声発話全体に対して適切な言語識別を実現できていないという課題が存在する．従来法では，音声発話中の短時間の各音声フレームでそれぞれ独立に求めた事後確率を全体で平均化することで音声発話単位の言語識別を行っているが，それが音声発話全体に対して適した識別を実現できているとは限らない．

この課題を解決するために，我々は従来法で出力される DNN 事後確率の系列としての情報に着眼する．DNN 事後確率の音声発話全体での系列の変動は，言語ごとにパターンがあると考えられる．よって，この変動パターンを直接モデル化することができれば，DNN に基づく枠組みのさらなる高精度化が期待できる．

そこで本稿では，DNN 事後確率系列の言語モデル化に基づく言語識別を提案する．つまり，発話全体の変動パターンを言語モデルでモデル化する．提案法では，DNN 事後確率系列をベクトル量子化に基づき離散系列化し，その離散系列を言語モデルでモデル化することで，言語識別を実現する．これにより，系列の変動まで捉えることができ，従来法に対する性能向上が期待できる．ベクトル量子化には k-means クラスタリング，言語モデルには階層 Pitman-Yor 言語モデル，およびリカレントニューラルネットワーク言語モデルを用いて本枠組みを検討する [5, 6]．

## 2 関連研究

音声の言語識別において言語モデルを利用する枠組みとして，音素認識器に基づく手法が検討されている [7]．具体的には，入力音声を任意のある言語の音素認識器で音素系列にデコードし，その系列を言語モデルでモデル化しておくことで言語識別を実現している．

また，ニューラルネットワークの事後確率に基づく特徴量は，タンデム特徴量と呼ばれ，音声認識で検討されている [8]．実際に，音声認識における音響モデルの特徴量として利用することで，通常の音響特徴量と比較して識別性能の向上が確認されている．

提案法は，両者のアプローチの組み合わせと位置づけることができる．DNN 事後確率系列を言語モデルでモデル化しようという試みはこれまでなされていない．提案法では離散化の処理が必要ではあるものの，DNN 事後確率系列の変動を直接捉えるためには言語モデルの利用が適していると考えられる．

## 3 DNN に基づく言語識別

### 3.1 音声フレーム単位の DNN

ここでは，音声フレーム単位の DNN に基づく言語識別について述べる [3]．この枠組みでは，ある短時間

の音声フレームの音響特徴量を $\boldsymbol{x}$ とした際に，ある言語 $l$ に対する事後確率 $P(l|\boldsymbol{x},\boldsymbol{\Theta})$ を直接推定することが可能な DNN をモデル化する．ここで，$\boldsymbol{\Theta}$ は DNN のモデルパラメータを表す．

DNN は，フィードフォワード型のニューラルネットワークであり，入力層と出力層の間に複数の隠れ層を持つ．ニューラルネットワークの各ノードにおける活性化関数は様々なものを利用可能であるが，本稿における DNN では，出力層を除く各層においてシグモイド関数を活性化関数として用いる．出力層を除くある層における $j$ 番目のノードへの入力を $\boldsymbol{z}$ と置くと，$j$ 番目のノードの出力 $y_j$ は (1) 式で定義できる．

$$y_j = \frac{1}{1+\exp(-\boldsymbol{w}_j^\top \boldsymbol{z} + b_j)}, \tag{1}$$

ここで，$\boldsymbol{w}_j$，$b_j$ は $j$ 番目のノードのパラメータであり，$\boldsymbol{\Theta}$ の一部分に当たる．ある層の $J$ 個のノード全てに対して出力値を求め結果を $\boldsymbol{y} = y_1, \cdots, y_J$ とすると，$\boldsymbol{y}$ が次の層の各ノードへの入力となる．なお次の層が出力層でなければ，$\boldsymbol{y}$ が次の層における $\boldsymbol{z}$ に該当する．また入力層では，$\boldsymbol{x}$ が $\boldsymbol{z}$ に該当する．

出力層では各クラスに対する確率値を得るために，ソフトマックス関数を活性化関数に用いる．出力層は各ノードが各言語ラベルに対応している．この時，ある言語 $l$ に対する事後確率 $P(l|\boldsymbol{x},\boldsymbol{\Theta})$ は (2) 式で定義できる．

$$P(l|\boldsymbol{x},\boldsymbol{\Theta}) = \frac{\exp(\boldsymbol{w}_l^\top \boldsymbol{y} + b_l)}{\sum_m \exp(\boldsymbol{w}_m^\top \boldsymbol{y} + b_m)}, \tag{2}$$

ここで，$\boldsymbol{w}_l$，$b_l$ は，言語 $l$ に対応するノードのパラメータである．DNN では，全ての層の全てのノードの重みパラメータを学習により決定する．学習は，確率的勾配法を利用してクロスエントロピー基準によって行うことができる．

### 3.2 言語識別方法

音声フレーム単位の DNN を用いた言語識別方法について述べる．ある入力発話 $\boldsymbol{X}$ の言語識別は (3) 式に従う．

$$\hat{l} = \arg\max_{l \in L} P(l|\boldsymbol{X}), \tag{3}$$

ここで，$L$ は言語識別器が対象とする全言語の集合，$P(l|\boldsymbol{X})$ は $\boldsymbol{X}$ がある言語 $l$ である確率値を表す．学習した DNN は，入力発話の各フレーム単位での事後確率を出力することができるので，各フレーム単位の DNN を事後確率を利用して $P(l|\boldsymbol{X})$ を算出する．ここで，$\boldsymbol{X}$ をフレーム単位に分割した特徴量系列を $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_K$ と表すと，$P(l|\boldsymbol{X})$ は (4) 式で得られる．

$$P(l|\boldsymbol{X}) = \prod_{k=1}^{K} P(l|\boldsymbol{x}_k, \boldsymbol{\Theta}). \tag{4}$$

これは，各音声フレームがそれぞれ独立であることを仮定していることに相当する．この枠組みは，音声フレーム単位での結果を利用するため，言語識別の早期確定なども容易に実現できる．

## 4 DNN 事後確率系列の言語モデル化に基づく言語識別

### 4.1 DNN 事後確率系列の離散化

提案法では，従来法で出力できる DNN 事後確率系列を離散系列化し，言語ごとに離散系列をモデル化しておくことで言語識別を実現する．

離散系列化は，$k$ 番目のフレームの DNN 事後確率分布を $\boldsymbol{p}_k$ とおくと，DNN 事後確率系列 $\boldsymbol{P} = \boldsymbol{p}_1, \cdots, \boldsymbol{p}_K$ を離散系列 $\boldsymbol{S} = s_1, \cdots, \boldsymbol{s}_K$ に変換する．本稿では，この離散系列化のために k-means クラスタリングを利用する．k-means クラスタリングでは，ベクトル集合から指定した数のセントロイド (代表ベクトル) を学習することができる．そこで，DNN の学習データについても DNN 事後確率系列を求め，その集合からセントロイドを学習する．ここでは，求めた $T$ 個のセントロイドを $\boldsymbol{c}_1, \cdots, \boldsymbol{c}_T$ とする．このセントロイドを利用した DNN 事後確率 $\boldsymbol{p}_k$ の離散化は (5) 式に従う．

$$s_k = \arg\min_t D(\boldsymbol{c}_t, \boldsymbol{p}_k), \tag{5}$$

ここで $D$ は，2 ベクトル間のユークリッド距離を表す．つまりこの処理では，各事後確率を最近傍のセントロイドのインデクス番号に離散化する．

### 4.2 言語モデル化に基づく言語識別

離散系列化された DNN 事後確率系列を，言語モデルを用いて言語ごとにモデル化する．ある言語 $l$ についての言語モデルは，DNN 事後確率の離散系列 $\boldsymbol{S}$ の生成確率 $P(\boldsymbol{S}|\boldsymbol{M}_l)$ をモデル化するものとする．ここで $\boldsymbol{M}_l$ は，言語 $l$ の言語モデルのモデルパラメータを表す．DNN 事後確率系列の言語モデル化することで，$\boldsymbol{S}$ がある言語 $\boldsymbol{l}$ である確率を (6) 式で定義できる．

$$P(l|\boldsymbol{S}) = \frac{P(\boldsymbol{S}|\boldsymbol{M}_l)}{\sum_m P(\boldsymbol{S}|\boldsymbol{M_m})}. \tag{6}$$

このとき，DNN 事後確率系列の言語モデル化に基づく言語識別は (7) 式に従う．

$$\hat{l} = \arg\max_{l \in L} P(l|\boldsymbol{S}). \quad (7)$$

この枠組みでは，DNN 事後確率の離散系列をどのように言語モデルでモデル化するかが重要となる．

n-gram 言語モデルの 1 種である階層 Pitman-Yor 言語モデルであれば，$k$ 番目のフレームの確率値を求める際に $n-1$ フレームのコンテキスト情報 $\boldsymbol{s}_{k-n+1}^{k-1}$ を利用する．階層 Pitman-Yor 言語モデルによる $P(\boldsymbol{S}|\boldsymbol{M}_l)$ は (8) 式で定義される．

$$P(\boldsymbol{S}|\boldsymbol{M}_l^{\mathrm{HPY}}) = \prod_{k=1}^{K} P(s_k|\boldsymbol{s}_{k-n+1}^{k-1}, \boldsymbol{M}_l^{\mathrm{HPY}}). \quad (8)$$

リカレントニューラルネットワーク言語モデルは，隠れ層に再帰的なループを持つリカレントニューラルネットワークを応用したものであり，$k$ 番目のフレームの確率を求める際に $k-1$ 番目のフレームと $k-1$ 番目のフレームの隠れ層の出力ベクトル $\boldsymbol{h}_{k-1}$ を利用する．リカレントニューラルネットワーク言語モデルによる $P(\boldsymbol{S}|\boldsymbol{M}_l)$ は (9) 式で定義される．

$$P(\boldsymbol{S}|\boldsymbol{M}_l^{\mathrm{RNN}}) = \prod_{k=1}^{K} P(s_k|s_{k-1}, \boldsymbol{h}_{k-1}, \boldsymbol{M}_l^{\mathrm{RNN}}). \quad (9)$$

各フレームの確率の予測時は，n-gram モデルではコンテキスト情報が $n-1$ フレームに限定される一方で，リカレントニューラルネットワーク言語モデルではリカレント構造によって長距離のコンテキスト情報を利用できる．

# 5 実験

## 5.1 実験条件

提案法の有効性を検証するために，多言語音声データベース Globalphone を利用して評価実験を行った [9]．Globalphone は各言語の母国語話者により発話された短文発話の音声データであり，各発話は 5s から 10s 程度である．今回は，フランス語 (FR)，ドイツ語 (GE)，韓国語 (KO)，中国語 (MA)，ポルトガル語 (PO)，ロシア語 (RU)，上海語 (SH)，スペイン語 (SP)，スウェーデン語 (SW)，タイ語 (TH)，トルコ語 (TU)，ベトナム語 (VI) の計 12 言語の音声データを利用した．我々は，話者オープンになるように，学習データ，開発データ，テストデータに分割した．それぞれのデータの音声ファイル数は，116041，2371，3181 である．

まず本実験で利用する DNN は，中間層 5 層，1024 ノードとした．音響特徴量には，フレームサイズ 20ms，フレームシフト 10ms として抽出した 38 次元の MFCC (12MFCC+12ΔMFCC+12ΔΔMFCC+Δ 対数パワー +ΔΔ 対数パワー) を用いた．なお DNN の入力時には，対象フレームの特徴量に加え，前後 10 フレームを結合した 798 次元のベクトルを利用した．DNN の学習時は，最初に Discriminative プリトレーニングに基づき初期のネットワーク構造を構築した後に，開発データに対して最適になるようにファインチューニングを実施した．その際の確率的勾配法におけるミニバッチサイズは 1024，初期学習率は 0.01，モーメンタムは 0.9 とした．

提案法における DNN 事後確率系列の離散化のための k-means クラスタリングでは，セントロイドの数を 32 個，および 64 個とし，2 種類のセントロイド集合を求めて利用した．DNN 事後確率系列のモデル化のための言語モデルとしては，階層 Pitman-Yor 言語モデル (HPYLM) は 3-gram モデルを学習した．一方，リカレントニューラルネットワーク言語モデル (RNNLM) は，中間層 100 ノードで開発データに最適になるように学習した．

## 5.2 実験結果

従来法，提案法について，開発データおよび，テストデータに対する言語識別の性能を評価した．我々は発話時間が短時間の場合の性能も評価するために，開発データ，およびテストデータの前半部分 (1s および 3s) のみを使った場合の性能も検証した．なお，前半部分のみを使った場合でも，評価データの総数は変わらない．評価指標には，Equal Error Rate (EER) を使用した．評価結果を表 1 に示す．

まず，従来法，提案法に関わらず，音声データが長いほど高い識別性能を実現できている．これは音声データが長いほど，識別に有用な情報が含まれる可能性が高いことに起因すると考えられる．音声データの長さが 1s の時は，従来法である DNN のみの場合と，提案法である言語モデルを利用する場合と比較して，ほとんど性能が変わらなかった．一方で，音声データの長さが 3s 以上の時は，提案法により大きく性能改善を達成できた．これは，事後確率の系列としての情報を言語モデルにより捉えることができたことに起因すると考えられる．

表 1: EER(%) による言語識別性能

| | | セントロイドの数 | 言語モデル | 開発データ | | | テストデータ | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | 1s | 3s | 全体 | 1s | 3s | 全体 |
| 従来法 | DNN | - | - | 7.72 | 1.24 | **0.43** | 11.01 | 4.47 | 3.12 |
| 提案法 | DNN | 32 | HPYLM | **7.36** | 0.89 | 0.53 | **10.22** | 2.87 | 1.68 |
| 提案法 | DNN | 64 | HPYLM | 7.61 | **0.81** | **0.43** | 10.45 | 3.09 | 1.51 |
| 提案法 | DNN | 32 | RNNLM | 7.61 | 0.97 | **0.43** | 10.30 | 2.68 | 1.36 |
| 提案法 | DNN | 64 | RNNLM | 7.68 | 0.89 | 0.48 | 10.35 | **2.33** | **1.07** |

表 2: テストデータに対する言語ごとの EER(%) による言語識別性能

| | FR | GE | KO | MA | PO | RU | SH | SP | SW | TH | TU | VI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 従来法 | 0.00 | 0.00 | 0.00 | 0.00 | 0.40 | 12.35 | 21.15 | 1.99 | 1.32 | 0.00 | 2.85 | 0.00 |
| 提案法 | 0.00 | 0.00 | 0.00 | 0.00 | 1.57 | 5.68 | 0.89 | 1.49 | 1.84 | 0.00 | 3.28 | 0.00 |

言語モデルとして，HPYLM，および RNNLM の 2 種類を使用したが，開発データに対してはあまり傾向が見られなかった．一方で，テストデータでは RNNLM の方が有効な性能を示した．RNNLM は，長距離の関係を柔軟に捉えることができるモデルであるため，その効果が出ているのではないかと考えれる．

次に，従来法と提案法 (セントロイドの数 64, RNNLM) について，テストデータ (音声全体) についての言語識別の性能を各言語ごとに調査した結果を表 2 に示す．従来法では，ロシア語と上海語で特に性能が低いことが見てとれる．この内訳としては，ロシア語はポルトガル語に，上海語は中国語に間違える場合が多かった．一方で提案法では，ロシア語と上海語の性能を大きく改善できていることが見てとれる．提案法は，従来法で識別誤りが起こりやすい言語について性能改善できる点が特徴であることが分かった．

## 6 まとめ

本稿では DNN に基づく言語識別の高精度化を目指して，DNN 事後確率系列の言語モデル化に基づく方法を提案した．提案法では，DNN 事後確率系列を離散系列として捉えてから言語モデルを適用することで，事後確率系列の変化を捉えた言語識別を行うことが可能となった．

Globalphone を用いた言語識別の評価実験から，提案法は音声の発話長が長い場合に有効であることが示され，特に従来法で識別誤りを起こしやすい言語に対して大幅な性能改善を実現した．

今後は，対象とする言語数を増やした場合の性能を検証するとともに，DNN 自体にリカレント構造を持つものを利用した場合との比較も実施する予定である．

## 参考文献


[1] Eliathamby Ambikairajah, Haizhou Li, Liang Wang, Bo Yin, and Vidhyasaharan Sethu, “Language identification: A tutorial,” *Circuits and Systems Magazine, IEEE*, vol.11, no.2, pp.82108, 2011.

[2] Haizhou Li, Bin Ma and Kong Aik Lee, “Spoken Language Recognition: From Fundamentals to Practice,” *Proceedings of the IEEE*, vol.101, pp.1136-1159,2013.

[3] Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez and Oldrich Plchot, “Automatic Language Identification Using Deep Neural Networks,” *In Proc. ICASSP 2014*, pp.5337-5341, 2014.

[4] Najim Dehak, Pedro A.Torres-Carrasquillo, Douglas Reynolds and Reda Dehak, “Language Recognition via Ivectors and Dimensionality Reduction,” *In Proc. Interspeech*, pp.857-860, 2011.

[5] Yee Whye Teh, “A Hierarchical Bayesian Language Model based on Pitman-Yor Processes,” *In Proc. COLING/ACL 2006*, pp.985-992, 2006.

[6] Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky and Sanjeev Khudanpur, “Recurrent Neural Network based Language Model,” *In Proc. Interspeech 2010*, pp.1045-1048, 2010.

[7] Mark A. Zismann, “Comparison of four approaches to automatic language identification of telephone speech,” *Speech and Audio Processing, IEEE Transactions*, vol.4, pp.31-44, 1996.

[8] Hynek Hermansky, Daniel P.W. Ellis and Sangita Sharma, “Tandem connectionist feature stream extraction for conventional HMM systems,” *In Proc. ICASSP 2000*, pp.1635-1638, 2000.

[9] Tanja Schultz, Ngoc Thang Vu and Tim Schlippe, “Globalphone: A Multilingual Text and Speech Database in 20 Languages,” *In Proc. ICASSP 2013*, pp.8126-8130, 2013.