Noisy Channel Model に基づく音声特徴量強調に関する検討*

☆ バン フクアンフイ, 齋藤大輔, 柏木陽佑, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

携帯端末などの普及により音声認識システムは身近な物となりつつあるが、実環境では多様な雑音が混入するため、音声認識システムの精度が低下してしまう。そこで、実環境における音声認識のための、耐雑音処理の研究が注目されている。

耐雑音処理は大きく 2 つのアプローチに分けることができる。1 つ目は、音響モデルを雑音に対して適応するアプローチであり、代表的な物としては Parallel Model Combination (PMC)[1] や Vector Taylor Series (VTS) 適応 [2] が挙げられる。2 つ目は、ノイジー音声特徴量から雑音成分を除去し、クリーン音声特徴量に近づけることで音響モデルとのミスマッチを低減する音声特徴量強調アプローチである。音声特徴量強調の代表的な手法として、Stereo-based Piecewise LInear Compensation for Environments (SPLICE)[3] や Stereo-based Stocastic Mapping (SSM)[4] などがあるが、これらはノイジー音声特徴量とクリーン音声特徴量の関係を学習するために、パラレルデータが必要となる。しかし、パラレルデータはその性質上、十分な量の学習データを確保することが困難であり、モデルの複雑度を上げると過学習が起きてしまう。

そこで、この問題を回避するため、本稿は Noisy Channel Model に基づく新たな特徴量強調手法を提案する。Noisy Channel Model は声質変換などで利用されており [5]、これを用いることでクリーン音声特徴量の分布を事前分布として利用することが可能となる。そのため、従来の手法で利用することが困難であった非パラレルなクリーン音声を効率的にモデル学習に用いることができる。これによる学習データ数の増加により、より高い精度でのクリーン音声特徴量の推定が期待される。

2 Stereo-based Stocastic Mapping (SSM) 手法

特徴量強調はノイジー特徴量からクリーン特徴量を推定する技術である。ノイジー特徴量系列を $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_x}]$、クリーンの特徴量系列を $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{n_y}]$ とする。特徴量強調は以下で定式化される。

$$\hat{\boldsymbol{y}}_t = \underset{\boldsymbol{y}_t}{\operatorname{argmax}}\, p(\boldsymbol{y}_t|\boldsymbol{x}_t) \tag{1}$$

まず、結合ベクトル系列 $\boldsymbol{z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n]$ を作る。ここで $\boldsymbol{z}_t = [\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top]$ である。結合確率密度は以下のように、Gaussian Mixture Model (GMM) でモデル化する。

$$p(\boldsymbol{z}_t|\boldsymbol{\lambda}^{(z)}) = \sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}) \tag{2}$$

ここで $\boldsymbol{\lambda}^{(z)}$ は重み w_m、平均ベクトル $\boldsymbol{\mu}_m^{(z)}$、分散共分散行列 $\boldsymbol{\Sigma}_m^{(z)}$ からなる結合モデルのパラメータである。平均ベクトル、分散共分散行列は以下のように書くことができる。

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix} \tag{3}$$

各パラメータは EM アルゴリズムで推定される。

クリーン音声の特徴量の推定値 $\hat{\boldsymbol{y}}_t$ は以下のように求められる。

$$\hat{\boldsymbol{y}}_t = \left(\sum_{m=1}^{M} \beta_{m,t} \boldsymbol{D}_m^{(y)-1} \right)^{-1} \times \left(\sum_{m=1}^{M} \beta_{m,t} \boldsymbol{D}_m^{(y)-1} \boldsymbol{E}_{m,t}^{(y)} \right) \tag{4}$$

ここで、

$$\beta_{m,t} = p\big(m|\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{\lambda}^{(z)}\big) \tag{5}$$

$$\boldsymbol{E}_{m,t}^{(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} \big(\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}\big) \tag{6}$$

$$\boldsymbol{D}_m^{(y)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} \boldsymbol{\Sigma}_m^{(xy)} \tag{7}$$

である。

3 提案手法

式 (1) の $p(\boldsymbol{y}_t|\boldsymbol{x}_t)$ を直接モデル化する代わりに、ベイズ定理を用いて、以下のように表す。

$$\hat{\boldsymbol{y}}_t = \underset{\boldsymbol{y}_t}{\operatorname{argmax}}\, p(\boldsymbol{x}_t|\boldsymbol{y}_t) p(\boldsymbol{y}_t) \tag{8}$$

式 (8) は Noisy Channel Model とよばれ、入力モデル $p(\boldsymbol{y}_t)$ とチャネルモデル $p(\boldsymbol{x}_t|\boldsymbol{y}_t)$ の二つの要素から構成される。Noisy Channel Model の利点として入力モデルとチャネルモデルを独立にモデル化できるため、特に、入力モデルの学習にパラレルデータを用いる必要がない。

モデル学習について、まずチャネルモデル $p(\boldsymbol{x}_t|\boldsymbol{y}_t)$ に関して、SSM 手法と同じく結合モデルからパラメータを抽出する。ただし、$\boldsymbol{x}_t$ と $\boldsymbol{y}_t$ は SSM と逆方向の変換モデルとなる。

次に式 (8) の入力モデル $p(\boldsymbol{y}_t)$ を大量のクリーンデータから Universal Background Model - Gaussian Mixture Model (UBM-GMM) としてモデル化する。

$$p(\boldsymbol{y}_t|\boldsymbol{\lambda}^{(c)}) = \sum_{n=1}^{N} w_n \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_n^{(c)}, \boldsymbol{\Sigma}_n^{(c)}) \tag{9}$$

式 (8) に基づき、尤度関数は以下のように定義する。

$$\mathcal{L}(\boldsymbol{y}_t; \boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}, \boldsymbol{\lambda}^{(c)}) = p(\boldsymbol{x}_t|\boldsymbol{y}_t, \lambda^{(z)}) p(\boldsymbol{y}_t|\lambda^{(c)})^\alpha \tag{10}$$

---

*Speech feature enhancement based on the Noisy Channel Model by Van Phu Quang Huy, Daisuke Saito, Yosuke Kashiwagi, Nobuaki Minematsu, Keikichi Hirose (The University of Tokyo)

ここで α は結合モデルとクリーンモデルのバランスをコントロール重みである。

Noisy Channel Model に基づく声質変換 [5] と同様、更新式は以下のようになる。

$$\hat{\boldsymbol{y}}_t = \Big(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{'(y)-1} + \alpha \sum_{n=1}^{N} \zeta_{n,t} \boldsymbol{\Sigma}_n^{(c)-1} \Big)^{-1} \times \Big(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{'(y)-1} \boldsymbol{E}_{m,t}^{'(y)} + \alpha \sum_{n=1}^{N} \zeta_{n,t} \boldsymbol{\Sigma}_n^{(c)-1} \boldsymbol{\mu}_n^{(c)} \Big) \tag{11}$$

ここで、

$$\gamma_{m,t} = p(m|\boldsymbol{y}_t, \boldsymbol{\lambda}^{(z)}), \quad \zeta_{n,t} = p(n|\boldsymbol{y}_t, \boldsymbol{\lambda}^{(c)}) \tag{12}$$

$$\boldsymbol{E}_{m,t}^{'(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yy)} \boldsymbol{\Sigma}_m^{(xy)+} (\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}) \tag{13}$$

$$\boldsymbol{D}_m^{'(y)-1} = \boldsymbol{D}_m^{(y)-1} - \boldsymbol{\Sigma}_m^{(yy)-1} \tag{14}$$

記号 $(\cdot)^+$ は模擬逆行列を表す。音声特徴量強調の場合、ノイジー特徴量の空間が縮退するため、相互共分散が小さくなる結果、式 (13) の第 2 項が発散することで音声認識性能の低下をまねく恐れがある[1]。そのため、式 (13) の第 2 項の係数行列 $\boldsymbol{\Sigma}_m^{(yy)} \boldsymbol{\Sigma}_m^{(xy)+}$ を単位行列に近似し、発散を抑制した更新式についても比較実験を行う。

4 実験

4.1 実験条件

AURORA2 を用いた雑音環境下の連続数字音声認識において、提案手法の評価を行った [6]。AURORA2 のサブセットについて、A セットが雑音環境クローズドテスト、B セットが雑音環境オープンテスト、C セットがチャネルノイズありのテストとなっている。

各セットにおける音声認識率の平均によって、性能評価を行った。なお各セットには、 4 種類の雑音が 5 種類の SN 比（0〜20）で重畳されたサブセットが用意されている。音声認識においてクリーン音声を学習データとして HMM を学習した。特徴量には、MFCC とそのパワー、およびその Δ、ΔΔ (MFCC_E_D_A) の 39 次元を用いた。結合 GMM とクリーン UBM の混合数が両方 512 とした。初期検討として非パラレルデータは用いず、結合 GMM のクリーンデータと UBM のクリーンデータは同じデータから学習した。

4.2 実験結果

まず、重み $\alpha = 1$ とし、提案手法 (NCM) を特徴強調なし (baseline) 、SSM と比較した。結果を Table 1 に示す。式 (13) の第 2 項に制約を加える事 (NCM-I) で、認識結果が改善することが確認できる。

さらに NCM-I において重み α を変更した。結果を Table 2 に示す。α を調整する事で、一定の認識率の改善が見られた。今回の実験を通して、Noisy Channel Model に基づく音声強調は SSM を用いた音声強調法に比べて十分な改善が見られなかった。原因の一つとして、クリーンモデル (UBM) と結合 GMM のクリーンデータを同じデータで学習したため、事前分布としての効果が期待したほど得られなかったと考えられる。また、声質変換の場合は入力モデル $p(\boldsymbol{y}_t)$ を一人の話者 (話者依存) から構成するのに対して、特徴量強調の場合は複数話者の音声を含むの UBM を作る。このため、音響特徴量空間のモデル化に違いが生じ、UBM が入力されたデータに対して十分な効果を発揮していないという可能性もある。他の原因としては連続数字認識タスクにおいては、語彙が少ないため、クリーンモデルの効果が十分発揮されないとも考えられる。そのため、今後は大語彙音声認識タスクにおいて提案法を検討することが必要となっている。

Table 1　word accuracy の平均（%）

	Set A	Set B	Set C	Average
baseline	55.26	47.88	66.46	54.55
SSM	85.05	79.29	76.28	80.99
NCM	59.34	49.97	69.11	57.55
NCM-I	80.96	74.62	73.47	76.93

Table 2　重み α を変更する場合の word accuracy の平均（%）

α	Set A	Set B	Set C	Average
0.8	81.80	74.91	74.30	77.55
1.0	80.96	74.62	73.47	76.93
1.2	81.52	74.78	74.04	77.32
2.0	79.38	72.88	72.11	75.33

5 まとめ

本稿は Noisy Channel Model に基づく特徴量強調手法を提案し、その初期検討を行った。Noisy Channel Model に基づく変換は声質変換で検討されたものであるが、本稿における特徴量強調実験の結果、タスクの違いによるいくつかの異なる傾向が明らかになった。今後は、音声認識のための音声特徴量強調に適した提案法の改良を検討するほか、結合モデルとクリーンモデルの独立性に着眼し、クリーンモデルを HMM で表現するなど、より精緻なモデルの導入も検討していく予定である。

参考文献

[1] M.J.F. Gales and S.J. Young, IEEE TSAP, 4 (5), 352-359, 1996.
[2] A. Acero *et al.* Proc.ICSLP, 869-872, 2000.
[3] J. Droppo *et al.* Proc.ICSLP, 29-32, 2002.
[4] M. Afify *et al.*, IEEE TSAP, 17 (7),1325-1334, 2009.
[5] D. Saito, IEEE TSAP, 20 (10), 1784-1794, 2012.
[6] H.G. Hirsch and D. Pearce, Proc. ISCA ITRW ASR, 2000

[1] 一方、声質変換の場合は、特徴量系列のダイナミクスを強調する効果として働き、変換の性能の向上につながると考えられる。