強調音声の効率的サンプリングによるDNNの不確定性学習とデコード法*

○太刀岡勇気 (三菱電機・情報総研), 渡部晋治 (MERL)

1 はじめに

深層神経回路網 (Deep Neural Network; DNN) により、音声認識の性能は大きく向上した。我々も騒音下の音声認識タスクにより、DNNの有効性を確認した [1]。一方で、ガウス混合分布 (Gaussian mixture model; GMM) のために開発された手法をDNNに適用する研究も盛んである。本報では、GMMに基づく騒音下の音声認識では有効性が確認されている不確定性手法をDNNに適応することをめざす。(INTERSPEECH 2015 ではこれに関連するスペシャルセッションが開かれた。)

騒音条件では、DNNを使ったシステムであっても、音声強調により、音声認識性能は大きく改善する [2, 3]。しかしながら、音声に騒音抑圧に起因する歪みが加わることで音声認識性能が大きく低下する。特に学習時とデコード時の騒音条件にミスマッチがある場合もしくはデコード時のみに音声強調をした場合、音響モデルのミスマッチと音声の歪みによって音声認識性能が低下するため、この問題は顕著になる。

この問題に対処するため、音声強調による歪みを表す信頼性指標に基づき特徴量を調整する方法がいくつか提案されている。GMMにおいては、特徴量の不確定性がガウス分布で表されるため、GMMの尤度はそれらの特徴量の不確定性の観点で期待値操作に基づき計算される。期待値は確率変数の周辺化により解析的に計算される。これにより、音声強調によりもたらされる歪みに対して音響モデルをより頑健にすることができる。この方法は不確定性デコーディング手法と呼ばれている。結果として、入力特徴量に対する音響モデルのガウス分布の共分散行列は、不確定性 (すなわち信頼性) の程度に応じて調整される。多くの不確定性を扱う手法が提案され、GMMに対するそれらの有効性が実証されている [4, 5]。例えば、文献 [4, 5] では騒音音声と強調音声の特徴量の差分ベクトルを使っている。しかし、DNNには非線形の活性化関数が含まれているため、不確定性の伝播を解析的に扱うことは困難である。

本報ではDNNのための、不確定性学習及びデコーディング手法を提案する。DNNのスコア計算とDNNの学習のために近似的な期待値演算を行う文献 [6] とは異なり、提案法はモンテカルロ法により不確定性に基づく入力特徴量をサンプルする。ただし、DNNの学習には多大な計算が必要なため、効率的なサンプリングが欠かせない。提案法では、音声強調前後の内挿ベクトルに特化する。確率的に内挿係数をサンプリングすることで、強調音声の特徴量のベクトルの分布を効率的に表すことができる。加えて、デコーディング時にもサンプリングを行い、各サンプルに対する複数の認識仮説を統合することでさらに認識性能を向上させる。

2 DNN不確定性学習/デコーディング

不確定性手法の背景にある理論は、式 (1) に示す条件付き期待値操作に基づいている。

$$\mathbb{E}[f(\boldsymbol{y}_{1:T})|\boldsymbol{x}_{1:T}] \triangleq \int f(\boldsymbol{y}_{1:T})p(\boldsymbol{y}_{1:T}|\boldsymbol{x}_{1:T})d\boldsymbol{y}_{1:T}, \tag{1}$$

ここで、$\boldsymbol{x}_{1:T} = \{\boldsymbol{x}_t|t = 1, \ldots, T\}$ は、T フレームからなる騒音音声の特徴量の時系列ベクトルであり、$\boldsymbol{y}_{1:T}$ は強調音声の時系列ベクトルである。$f()$ は、適用対象に応じてデコーディング (2.1 節参照) や学習 (2.2 節参照) を示す。$p(\boldsymbol{y}_{1:T}|\boldsymbol{x}_{1:T})$ は、不確定性を含んだ形での強調音声の時系列の確率的表現である (2.3 節参照)。

2.1 DNN不確定性デコーディング

まず隠れマルコフモデル (Hidden Markov model; HMM) とDNNを統合したハイブリット構造のDNNのための不確定性デコーディングに焦点を当てる。この枠組みでは、式 (1) の $f()$ は以下の実際のデコーディング過程で表される。

$$\begin{aligned}\hat{W} &= \mathbb{E}\left[\arg\max_W p(\boldsymbol{y}_{1:T}|\mathcal{H}_W)p(W)\middle|\boldsymbol{x}_{1:T}\right], \\ &= \mathbb{E}\left[W_{\boldsymbol{y}_{1:T}}|\boldsymbol{x}_{1:T}\right],\end{aligned} \tag{2}$$

W は単語系列、$\mathcal{H}_W$ は W が与えられた時に起こりうるHMMの状態系列、$W_{\boldsymbol{y}_{1:T}}$ は入力特徴量 $\boldsymbol{y}_{1:T}$ が与えられたときのデコードされた単語系列である。いくつかのGMMに基づく従来の不確定性手法は、式 (2) に対する解析解を与える。$p(\boldsymbol{y}_{1:T}|\boldsymbol{x}_{1:T})$ に対するガウス分布に基づく不確定性を用いて、$\mathbb{E}[p(\boldsymbol{y}_{1:T}|\mathcal{H}_W)|\boldsymbol{x}_{1:T}]$ についての期待値操作を積分消去することに注意されたい。しかしながら、DNNに基づく音響モデルに対しては非線形の活性化関数があるために、このような解析解を得ることは出来ず、なんらかの近似が必要となる [6]。

*Uncertainty training and decoding methods for DNN based on efficient sampling of enhanced features, by TACHIOKA, Yuuki (Mitsubishi Electric Corporation); WATANABE, Shinji (MERL).

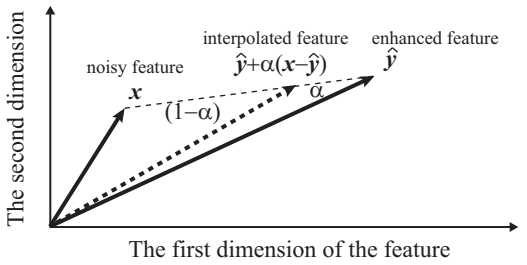

Fig. 1　Noisy feature $\boldsymbol{x}$ and enhanced feature $\hat{\boldsymbol{y}}$, and the sampling of feature $\boldsymbol{y}$ based on an interpolation between them.

提案法では、近似を使うというよりは、モンテカルロサンプリングに基づき、式 (2) から直接的な期待値操作をする。ただし積分操作というよりは、仮説レベルで複数の出力を平均化する。これらの出力は、異なる特徴量サンプルに対して別々にデコーディングすることで得られる。この方法の欠点はすべてのサンプルについて音声認識のデコーディングの計算が必要となることである。これはラティスリスコアリングで多少計算量を減らすことができるにしても問題となりうる。加えて、起こりうる入力特徴量空間を完全に被覆するように $\boldsymbol{y}_{1:T}$ をサンプルすることは非常に困難である。系列の入力特徴量 $p(\boldsymbol{y}_{1:T}|\boldsymbol{x}_{1:T})$ の分布を直接考慮するのではなくて、t フレームのサンプルされた入力特徴量 $\boldsymbol{y}_t$ の関係性を、$\boldsymbol{x}_t$ と $\hat{\boldsymbol{y}}_t$ の間の線形補間に基づき、

$$\boldsymbol{y}_t = \hat{\boldsymbol{y}}_t + \alpha(\boldsymbol{x}_t - \hat{\boldsymbol{y}}_t) \text{ for } t = 1, \ldots, T, \qquad (3)$$

のように決定的に仮定する。α は線形補間係数である。この線形補間の幾何学的な意味解釈を図 1 に示す。本手法は、騒音音声の特徴量と音声強調された音声の特徴量の差異から得られる共分散行列による観測量の近似的な分布 $p(\boldsymbol{y}_{1:T}|\boldsymbol{x}_{1:T}) \approx \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{y}_t|\hat{\boldsymbol{y}}_t, \left[\alpha(\boldsymbol{x}_t - \hat{\boldsymbol{y}}_t)(\boldsymbol{x}_t - \hat{\boldsymbol{y}}_t)^\top\right])$ [4, 5] に基づく不確定性デコーディングに想を得ている。次に線形内挿係数 α を確率変数であるとみなすことによって、1 次元の α を比較的少ない数のサンプルにより効率的にサンプルすることができる。

よって、提案の N 個のモンテカルロサンプルを用いた不確定性デコーディングの方法は式 (2) から以下のように表される。

$$\hat{W} = R\left[\left\{W_{\boldsymbol{y}^n_{1:T}}\right\}_{n=1}^{N}\right],$$
$$\boldsymbol{y}^n_t = \hat{\boldsymbol{y}}_t + \alpha^n(\boldsymbol{x}_t - \hat{\boldsymbol{y}}_t) \text{ for } t = 1, \ldots, T,\ \alpha^n \sim p(\alpha), \qquad (4)$$

$R[\cdot]$ は仮説レベルでの統合により実現され、例えば、Recognizer Output Voting Error Reduction (ROVER) [7] が使える。$\alpha^n \sim p(\alpha)$ の意味は、n 番目の α が分布 $p(\alpha)$ からサンプルされるということである。節 2.3 で、$p(\alpha)$ についてより詳細に議論する。

2.2　DNN 不確定性学習

節 2.1 での記述と同様にして、与えられた単語系列 W に基づく不確定性学習は、式 (1) における $f()$ を学習手順で置き換えることにより、

$$\hat{\Theta} = \mathbb{E}\left[\arg\min_{\Theta} \mathcal{F}_\Theta(\boldsymbol{y}_{1:T}, W) \middle| \boldsymbol{x}_{1:T}\right], \qquad (5)$$

のようにあらわすことができる。モデルパラメータを Θ とするとき、$\mathcal{F}_\Theta$ は DNN の評価関数であり、例えばクロスエントロピー (cross entropy; CE) 基準や、系列の識別的基準が考えられる。

入力特徴量は、2.1 節に提案の不確定性デコーディングと同じ方法で、線形内挿係数の分布 $p(\alpha)$ に基づきサンプリングする。式 (5) のパラメータについての期待値操作の代わりに、評価関数について

$$\begin{aligned}\hat{\Theta} &= \arg\min_{\Theta} \mathbb{E}\left[\mathcal{F}_\Theta(\boldsymbol{y}_{1:T}, W) | \boldsymbol{x}_{1:T}\right], \\ &\approx \arg\min_{\Theta} \sum_{n=1}^{N} \mathcal{F}_\Theta(\boldsymbol{y}^n_{1:T}, W), \\ &\text{where} \quad \boldsymbol{y}^n_t = \hat{\boldsymbol{y}}_t + \alpha^n(\boldsymbol{x}_t - \hat{\boldsymbol{y}}_t) \forall t, \quad \alpha^n \sim p(\alpha).\end{aligned} \qquad (6)$$

のように、モンテカルロサンプリングを導入する。CE 学習に対しては、モンテカルロサンプリングの評価関数は

$$\sum_{n=1}^{N} \mathcal{F}^{\mathrm{CE}}_\Theta(\boldsymbol{y}^n_{1:T}, W) = -\sum_{t=1}^{T}\sum_{n=1}^{N} \log p_\Theta(s_t|\boldsymbol{y}^n_t), \qquad (7)$$

のように表される。s_t はフレーム t における HMM 状態であり、W が与えられた時の Viterbi アライメントにより得られる。よって、単にサンプルされた学習データを入力特徴量として使うことで、評価関数についての加法性により、期待値操作が可能となる。この手法は DNN 学習における系列の識別学習 (例 [8]) にも適用可能である。

2.3　線形内挿係数の確率過程

各発話に対して、複数の α をサンプルするために、式 (8) の 1 次元の K 混合 GMM を用いる。

$$p(\alpha) = \sum_{k=1}^{K} w_k \mathcal{N}(\alpha|\mu_k, \sigma), \qquad (8)$$

ここで平均 μ_k は経験的に区間 $[0, 1]$ の内のいくつかの値に決定した。これにより、入力特徴量 $\boldsymbol{y}_t$ は騒音特徴量 $\boldsymbol{x}_t$ と音声強調された特徴量 $\hat{\boldsymbol{y}}_t$ の中間でサンプルされることとなる。分散 σ と混合重み $w_k (= 1/K)$ は固定とし、いくつかの実験においてはさらに $\alpha \in \{\mu_k\}_{k=1}^{K}$ も固定した (すなわち $\sigma \to 0$)。

3 実験の設定

3.1 コーパス

ここでは 2 つの騒音下および残響下音声認識タスクを用いて、提案法の有効性を示す。初めに用いるのは、第 2 回 CHiME チャレンジ　トラック 2 [9] である。これは中程度の語彙のタスクであり、音声発話は *Wall Street Journal* (*WSJ*) のデータベースより採られている。これに非定常騒音が、信号対雑音比 (signal-to-noise ratio; SNR) で −6 から 9 dB になるように混ぜられている。多チャンネル非負値行列因子分解 (multi-channel non-negative matrix factorization; MNMF) アルゴリズム [10] により、音声強調した。

2 つめのコーパスは、REVERB チャレンジ [11] のシミュレーションデータである。これは残響環境下におけるタスクで、発話は同じく *WSJ* のデータベースより採られている。音声データは、クリーン音声に 6 種の室内インパルス応答を畳み込むことで生成されている。6 つのインパルス応答の内訳は、3 つの残響時間が 0.25、0.5、0.75 秒と異なる室において、マイクと音源の距離が 0.5 m (near) もしくは 2 m (far) の 2 種収録されている。これに比較的定常な騒音が SNR20 dB で重畳されている。8 つのマイクが半径 0.1 m の円上に配置されている。到来方向推定に基づく多チャンネルビームフォーミングおよび単チャンネルの残響除去が適用されている。

3.2 音声認識の設定

音声認識の設定は 2 つのタスクに共通である。言語モデル重みと言った、いくつかのチューニングが必要なパラメータは開発セットの単語誤り率 (word error rate; WER) に基づいて最適化した。語彙サイズは 5k であり、トライグラムの言語モデルを使った。音声認識システムは、Kaldi ツールキット [12] を用いて構築した。提案の不確定性学習を行う際には、過学習を防ぐため、学習率を低減している。内挿学習データは元の学習データと似ているので、過学習しやすいためである。40 次元のフィルタバンク特徴量とその動的特徴量 (Δ と ΔΔ) を特徴量として用いた。DNN 音響モデルを CE 基準により学習した後に、系列のベイズリスク最小化 (sequential minimum Bayes risk; sMBR) 識別学習を行った [8]。

3.3 6 つのシステム設定

以下に示す 6 つのシステムを用意した。

1. noisy: $\boldsymbol{x}$ で学習し、$\boldsymbol{x}$ をデコードする。
2. enhan (強調音声; enhanced): $\boldsymbol{y}$ で学習し、$\hat{\boldsymbol{y}}$ をデコードする。
3. diff (差異; difference): $\left[\hat{\boldsymbol{y}}^\top, [\boldsymbol{x}-\hat{\boldsymbol{y}}]^\top\right]^\top$ をデコードする。
4. uncert(t) (不確定性学習; uncertainty training): $\hat{\boldsymbol{y}}$ をデコードする。ただしモデルは $\hat{\boldsymbol{y}}+\alpha[\boldsymbol{x}-\hat{\boldsymbol{y}}]$ に対し、$\mu_k \in \{0, 0.1, 0.2\}$ で学習する。
5. uncert(d) (不確定性デコーディング; uncertainty decoding): $\hat{\boldsymbol{y}}+\alpha[\boldsymbol{x}-\hat{\boldsymbol{y}}]$ を $\mu_k \in \{0, 0.1, 0.2\}$ でデコードする。ただしモデルは $\hat{\boldsymbol{y}}$ で学習する。複数の仮説を ROVER により統合する。
6. uncert(t,d) (不確定性学習/デコーディングの統合): $\hat{\boldsymbol{y}}+\alpha[\boldsymbol{x}-\hat{\boldsymbol{y}}]$ を $\mu_k \in \{0, 0.1, 0.2\}$ でデコードする。モデルも同じ特徴量で学習する。複数の仮説を ROVER により統合する。

4 結果と考察

4.1 第 2 回 CHiME チャレンジ　トラック 2

表 1 には、第 2 回 CHiME チャレンジ開発セットでの WER を示す。MNMF による音声強調により、DNN の音声認識システムの性能が堅調に向上した。差異特徴量を入力特徴量に結合した場合 (表の “diff”、これは文献 [4, 5] を模したもの) には、CE モデルに対して WER が 0.23%低減、sMBR(識別学習) モデルに対しては 0.31%の低減となった。この実験では固定の α を用いている。すなわち、$\alpha \in \{0, 0.1, 0.2\}$ である (式 (8) の $\sigma \to 0$)。提案の不確定性デコーディング (表中 “uncert(d)”) により、WER は CE モデル、sMBR モデルそれぞれに対して 0.37%、0.38%低減した。この場合、モデル再学習が不必要となるが、デコーディングのための計算時間は増加する。提案の不確定性学習 (表中 “uncert(t)”) により、WER は CE モデル、sMBR モデルそれぞれに対して 0.75%、0.55%低減した。この場合、学習にかかる時間は増加するものの、デコーディングにかかる時間は “enhan” や “diff” とほぼ同じである。DNN 音響モデルにおいては、不確定性をデコーディング時に考慮するよりも学習時に考慮する方が効率が良い。不確定性を学習時デコーディング時の双方に導入すると (表中 “uncert(t,d)”)、WER は顕著に向上し、CE モデル、sMBR モデルそれぞれに対して 0.92%、1.12%低減した。

表 1 には、内挿点に対してランダムな外乱を導入することの効果を示している (表中 ‘+p’)(式 (8) の $\sigma = 0.015$)。不確定性デコーディング (“uncert(d)”) ではこの手法はすべての σ に対して性能を向上させたわけではなかったが、不確定性学習 (“uncert(t)”) と両者の組み合わせ (“uncert(t,d)”) に対しては性能が向上した。$\sigma = 0.015$ の場合、CE 音響モデルに対して、学習では WER は 0.31%向上、学習/デコーディ

Table 1　Average WER [%] on the second CHiME challenge (Track 2).

	dt		et	
	CE	sMBR	CE	sMBR
noisy	31.58	28.90	26.56	24.59
enhan	27.89	24.92	23.09	20.29
diff	27.66	24.61	22.97	20.70
uncert(t)	27.14	24.37	22.40	20.51
+p	26.83	24.95	22.21	20.38
uncert(d)	27.52	24.54	22.69	19.99
+p	27.56	24.53	22.69	20.00
uncert(t,d)	26.97	**23.80**	22.11	20.10
+p	**26.26**	24.24	**21.96**	**19.86**

ングの併用では0.71%の向上が見られた。しかしながら、この手法ではsMBRモデルの音声認識性能を向上させることはできなかった。

表1には、評価セットでのWERも示している。この場合、不確定性の導入により、デコーディングのよりも学習の際の性能が向上し、両者の組み合わせ“uncert(t,d)”の場合に最も良い性能を達成した。この傾向は開発セットでの場合と同様であった。この場合、不確定性学習および学習/デコーディングの併用に対して、ランダムな外乱を導入することで、性能はsMBRモデルのときでさえも向上した。これにより、外乱によって音響モデルの未知データに対する頑健性を向上させることができることを示した。最後に、提案法は“enhan”からWERを、CEモデルの場合1.13%、sMBRモデルの場合0.43%低減させ、同様にして“diff”を0.12%、−0.41%上回った。これらの結果から提案法の有効性を確認できた。

4.2　REVERBチャレンジ

表2にはREVERBチャレンジにおけるWERを示す。ここでは固定のαを用いた。まず開発セットでのWERを見ると、CHiMEチャレンジの場合よりもベースラインの性能が高いものの、提案法はやはり有効であり、傾向も似ている。すなわち、提案法はデコーディング時よりも学習時に有効であり、それらを組み合わせることでさらに性能が向上した。

さらに評価セットにおけるWERを見ると、提案法により、“enhan”からWERを、CEモデルの場合0.52%、sMBRモデルの場合0.15%低減させ、“diff”をそれぞれ0.13%、−0.01%上回った。これより、提案法は2つのタスクにおいて音声認識性能を改善した。

5　まとめと課題

本報では、DNN音響モデルのための不確定性学習/デコーディング手法を提案した。提案法はDNNの学習とデコーディングの手順や構造を全く変えることなく使うことができる。不確定性を学習時とデコーディング時それぞれに導入した場合を比較することで、学習時に不確定性を導入することが最も効果的であることが分かった。加えて、内挿点を乱数により摂動を与えることで性能をさらに改善することができた。2つの騒音残響環境下の音声認識タスクにより提案法の有効性を確認した。今後の課題は、騒音の種類に応じて最適な内挿点を自動的に決定するアルゴリズムの開発である。

Table 2　Average WER [%] on the REVERB challenge simulation data.

	dt		et	
	CE	sMBR	CE	sMBR
noisy	8.56	6.95	8.84	7.34
enhan	7.66	6.04	7.79	6.57
diff	7.19	5.96	7.66	6.58
uncert(t)	7.21	5.86	7.32	6.56
uncert(d)	7.64	6.04	7.79	6.51
uncert(t,d)	**7.15**	**5.82**	**7.27**	**6.42**

参考文献

[1] Y. Tachioka et. al, Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark, the 2nd CHiME Workshop, pp.19–24, 2013.

[2] T. Yoshioka et. al, The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices, ASRU, pp.436–443, 2015.

[3] T. Hori et. al, The MERL/SRI system for the 3rd CHiME challenge using beamforming, robust feature extraction, and advanced speech recognition, ASRU, pp.475–481, 2015.

[4] M. Delcroix et. al, Static and dynamic variance compensation for recognition of reverberant speech with dereverberation preprocessing, IEEE Trans. on ASLP, 324–334, 2009.

[5] D. Kolossa et. al, Independent component analysis and time-frequency masking for speech recognition in multitalker conditions, EURASIP J. on Audio, Speech, and Music Processing, ID 651420, 2010.

[6] R. Astudillo and J. daSilva Neto, “Propagation of uncertainty through multilayer perceptrons for robust automatic speech recognition,” INTERSPEECH, 2011.

[7] J. Fiscus, “A post-processing system to yield reduced error word rates: Recognizer output voting error reduction (ROVER),” ASRU, pp.347–354, 1997.

[8] K. Veselý et. al, “Sequence-discriminative training of deep neural networks,” INTERSPEECH, pp.2345–2349, 2013.

[9] E. Vincent et. al, “The second ‘CHiME’ speech separation and recognition challenge: Datasets, tasks and baselines,” ICASSP, pp.126–130, 2013.

[10] A. Ozerov et. al, “A general flexible framework for the handling of prior information in audio source separation,” IEEE Trans. on ASLP, **20**, 1118–1133, 2012.

[11] K. Kinoshita et. al, “The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech,” WASPAA, 2013.

[12] D. Povey et. al, “The Kaldi speech recognition toolkit,” ASRU, pp.1–4, 2011.