Vol.2019-MUS-124 No.15
2019/8/28

時間周波数再帰型ニューラルネットワークを用いたスペクトログラムの低ランク近似のための基礎検討

丹治 寛樹[1,a)]　鎌田 弘之[1]

概要：時間周波数領域における音響信号処理において観測信号のスペクトログラムを低ランク近似する手法が広く用いられている．さまざまな近似手法のうち，観測信号のスペクトログラムの振幅またはパワーを基底スペクトルと非負の重みに分解する非負値行列因子分解（nonnegative matrix factorization; NMF）が低ランク近似を得るために最もよく利用される．NMF の一つの派生として，本稿では，観測信号のスペクトログラムから NMF における基底および重みを推定する再帰型ニューラルネットワーク（recurrent neural network; RNN）を提案する．提案するネットワークはスペクトログラムを周波数方向に走査して基底を出力する frequency-RNN と時間方向に走査して重みを出力する time-RNN の 2 つの独立した RNN により構成されており，これらの RNN の出力の線形結合を用いて観測信号のスペクトログラムを再現する．本稿では encoder-decoder モデルおよび variational autoencoder に基づく 2 つのネットワーク構造を検討し，簡単な近似例を示す．

Towards low-rank approximation of spectrograms using time-frequency recurrent neural network

1. はじめに

非負値行列因子分解（nonnegative matrix factorization; NMF）[1] は入力された非負の観測行列の低ランク近似を得るための枠組みであり，音響信号処理においては，音源の振幅またはパワースペクトログラムをモデリングするためによく利用される [2, 3]．音響信号のスペクトログラムに NMF を適用することで，その信号を構成するスペクトルのパターンと混合重みを教師なしで獲得できる [4] ため，ニューラルネットワークと組み合わせた半教師あり音源強調においても活用されている [5, 6].

ニューラルネットワークとスペクトログラムの低ランク近似との融合が進む一方，NMF の最適化問題をニューラルネットワークに基づいて解く試みもなされている [7–9]. 例えば文献 [7] では，NMF の重みを RNN を用いて学習する手法が提案されている．この手法では，与えられたスペクトログラムに対して NMF を適用し，得られた基底を固定して RNN を学習する．音響信号処理において，観測信号は調波構造を持つことが多く，観測信号のスペクトログラムに NMF を適用して得られた基底および重みはそれぞれスパース性および連続性を持つことが期待される [10]. そのため，RNN を用いることで重みの連続性を再現しようとするアプローチは有効である．文献 [7] ではこの手法により半教師あり信号分離に有益な重みが得られることが報告されているが，NMF の基底を出力する RNN は明らかになっていない．

そこで，本研究ではスペクトログラムを入力として与えたとき NMF の基底および重みを出力するニューラルネットワークの構築を目指す．とりわけ，本稿ではニューラルネットワークの構成について検討を行う．

基底および重みを学習するために, 本稿では, スペクトログラムを周波数方向に走査して基底を出力する frequency-RNN（F-RNN）と時間方向に走査して重みを出力する time-RNN（T-RNN）の 2 つの独立した RNN を用いる．これらの RNN の出力のを用いてスペクトログラム近似を得て，観測信号のスペクトログラムとの乖離度を計算することで RNN を学習する．スペクトログラムの近似を得るために，本稿では encoder-decoder および variational autoencoder（VAE）に基づく 2 つのモデルを検討する．

[1] 明治大学理工学部 電気電生命学科
Meiji University, Tama-ku, Kawasaki-shi, Kanagawa 214–8571, Japan
[a)] htanji@meiji.ac.jp

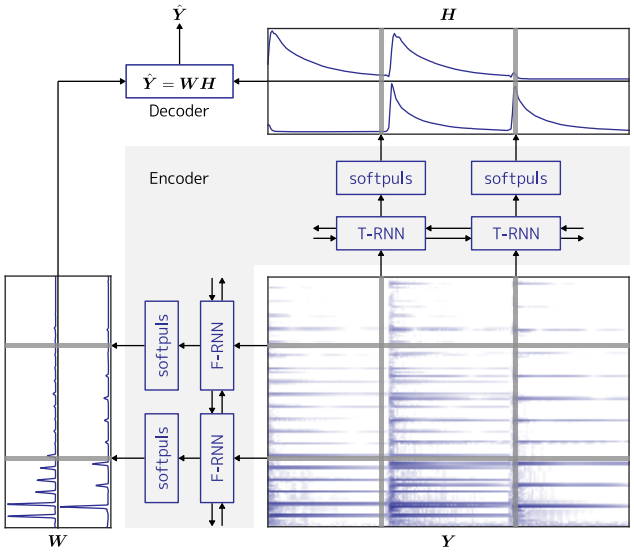

図 1　Encoder-decoder ネットワークの構成

2. 関連研究

再帰ユニットの出力を周波数方向および時間方向の双方に伝搬させる RNN を用いてスペクトログラムを解析する手法がいくつか提案されている [11–13]．文献 [11, 12] では，周波数方向の long short-term memory（LSTM）[14] によって獲得したスペクトログラムの特徴量を時間方向の LSTM に入力するネットワークを提案し，それぞれ音声認識や音高推定に適用している．また，文献 [13] では，時間方向および周波数方向に別の RNN を構築するのではなく，時間方向および周波数方向の双方に出力や状態を伝搬できる再帰ユニットが提案されている．

3. 非負値行列因子分解

本節では NMF の定式化について述べる．NMF では，観測信号の振幅またはパワースペクトログラム $\boldsymbol{Y} \in \mathbb{R}_+^{M\times N}$ を K 個の基底スペクトル $\boldsymbol{W} \in \mathbb{R}_+^{M\times K}$ と重み行列 $\boldsymbol{H} \in \mathbb{R}_+^{K\times N}$ の積で近似する．$\boldsymbol{Y}$ の要素を y_{mn} とすると，その近似 $\hat{y}_{mn}$ は $\hat{y}_{mn} = \sum_{k=1}^{K} w_{mk}h_{kn}$ で与えられる．ここで，w_{mk} および h_{kn} はそれぞれ $\boldsymbol{W}$ および $\boldsymbol{H}$ の要素である．$\hat{y}_{mn}$ の要素を並べた行列 $\hat{\boldsymbol{Y}}$ は $\hat{\boldsymbol{Y}} = \boldsymbol{WH}$ のように求められる．

観測信号の複素スペクトログラムが複素正規分布に従うと仮定すると，NMF は複素正規分布の分散の最尤推定問題として定式化される [15]．この定式化に基づけば，NMF の推定問題は次式で表される Itakura-Saito（IS）divergence 規範に基づく評価関数の最適化問題に帰着する．

$$f_{\mathrm{IS}}(\boldsymbol{W}, \boldsymbol{H}) = \frac{1}{MN}\sum_{m,n}\left(\frac{y_{mn}}{\hat{y}_{mn}} - \log\frac{y_{mn}}{\hat{y}_{mn}} - 1\right) \quad (1)$$

4. 提案モデル

本稿では，ネットワークの入力にスペクトログラムを与え，出力からスペクトログラムの近似を得るネットワークを構築する．そこで，図 1 に示すような encoder-decoder ネットワークと同様のモデルを考える．このネットワークは周波数方向および時間方向に 2 つの双方向 RNN をもっており，それぞれの RNN は入力されたスペクトログラムを周波数方向および時間方向に走査する．RNN の出力は全結合層に入力され，全結合層は基底および重みの推定値を出力する．図 1 において，ニューラルネットワークは $\boldsymbol{Y}$ のコーディング $\boldsymbol{W}$，$\boldsymbol{H}$ を得る encoder，$\boldsymbol{W}$ と $\boldsymbol{H}$ の線形結合を求める部分は $\boldsymbol{Y}$ を再構成する decoder とみなせる．

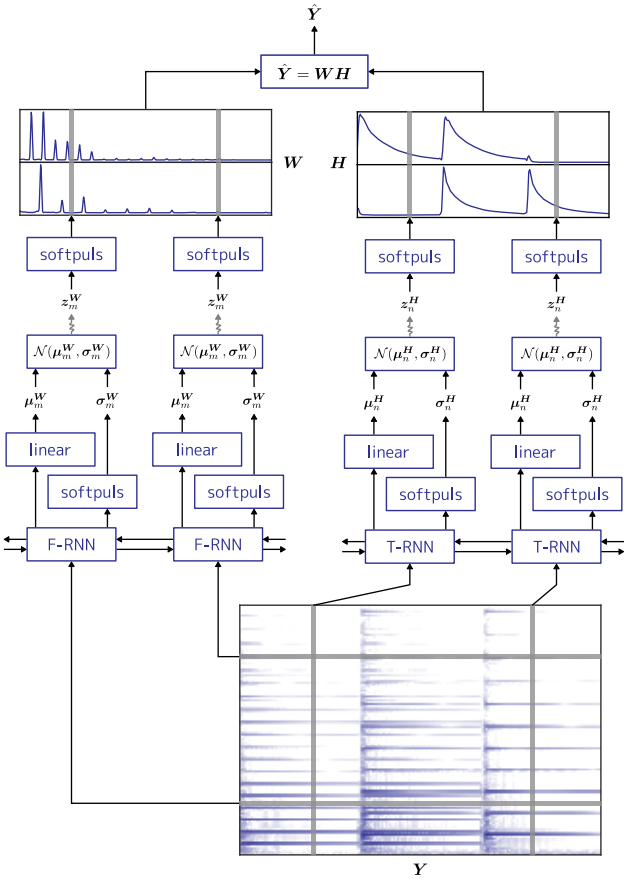

図 2　VAE の構成

図 1 に示すネットワークの学習手順は以下の通りである．

(1) 観測信号のスペクトログラム $\boldsymbol{Y}$ に NMF を適用し，基底 $\boldsymbol{W}_{\mathrm{NMF}}$ および重み $\boldsymbol{H}_{\mathrm{NMF}}$ を得る
(2) F-RNN および T-RNN を独立に学習する
- $\boldsymbol{W}_{NMF}$ を固定して T-RNN のみ学習する
- $\boldsymbol{H}_{NMF}$ を固定して F-RNN のみ学習する

(3) F-RNN および T-RNN の出力を用いて F-RNN および T-RNN を同時に学習する

NMF における基底と重みのスケールの任意性を解決するために，F-RNN の後段の全結合層の出力 w_{mk} が $\sum_m w_{mk} = 1$ を満たすように正規化する．

図 1 のモデルに VAE [16] で用いられているような生成モデルを導入することもできる．図 2 に生成モデルを導入したネットワークの構成を示す．VAE [16] では，潜在変数モデルにおける潜在変数の分布の母数をニューラルネッ

トワークで与える．図 2 では，基底および重みの潜在変数 $\boldsymbol{z}_m^{\boldsymbol{W}}$，$\boldsymbol{z}_n^{\boldsymbol{H}}$ が正規分布に従うと仮定し，RNN の後段の全結合層で正規分布の母数を得る．得られた母数を用いて $\boldsymbol{z}_m^{\boldsymbol{W}}$ および $\boldsymbol{z}_n^{\boldsymbol{H}}$ をサンプリングし，基底および重みを出力する．

VAE において最大化する評価関数は次式で与えられる．

$$
\begin{aligned}
f_{\mathrm{VAE}} = &\frac{1}{MN}\mathbb{E}_{q(\boldsymbol{z}_m^{\boldsymbol{W}}|\boldsymbol{Y})q(\boldsymbol{z}_n^{\boldsymbol{H}}|\boldsymbol{Y})}[\log p(y_{mn}|\boldsymbol{z}_m^{\boldsymbol{W}}, \boldsymbol{z}_n^{\boldsymbol{H}})] \\
&+ \frac{1}{M}\mathbb{E}_{q(\boldsymbol{z}_m^{\boldsymbol{W}}|\boldsymbol{Y})}\left[\log \frac{q(\boldsymbol{z}_m^{\boldsymbol{W}}|\boldsymbol{Y})}{p(\boldsymbol{z}_m^{\boldsymbol{W}})}\right] \\
&+ \frac{1}{N}\mathbb{E}_{q(\boldsymbol{z}_n^{\boldsymbol{H}}|\boldsymbol{Y})}\left[\log \frac{q(\boldsymbol{z}_n^{\boldsymbol{H}}|\boldsymbol{Y})}{p(\boldsymbol{z}_n^{\boldsymbol{H}})}\right] \qquad (2)
\end{aligned}
$$

ここで，$\mathbb{E}_{p(x)}[x]$ は分布 $p(x)$ の x についての期待値である．この式において対数尤度関数 $\log p(y_{mn}|\boldsymbol{z}_m^{\boldsymbol{W}}, \boldsymbol{z}_n^{\boldsymbol{H}})$ は y_{mn} および $\hat{y}_{mn}$ についての IS divergence と定数項を除いて等しい．また，q および p はそれぞれ潜在変数の事後分布および事前分布で，$\boldsymbol{z}_m^{\boldsymbol{W}}$ および $\boldsymbol{z}_n^{\boldsymbol{H}}$ の各要素 $z_{mj}^{\boldsymbol{W}}$，$z_{nj}^{\boldsymbol{H}}$ について以下の正規分布を設定する．

$$
\begin{aligned}
p(\boldsymbol{z}_m^{\boldsymbol{W}}) &= \textstyle\prod_j \mathcal{N}(z_{mj}^{\boldsymbol{W}}; 0, 1) && (3)\\
p(\boldsymbol{z}_n^{\boldsymbol{H}}) &= \textstyle\prod_j \mathcal{N}(z_{nj}^{\boldsymbol{H}}; 0, 1) && (4)\\
q(\boldsymbol{z}_m^{\boldsymbol{W}}|\boldsymbol{Y}) &= \textstyle\prod_j \mathcal{N}(z_{mj}^{\boldsymbol{W}}; \mu_{mj}^{\boldsymbol{W}}, {\sigma_{mj}^2}^{\boldsymbol{W}}) && (5)\\
q(\boldsymbol{z}_n^{\boldsymbol{H}}|\boldsymbol{Y}) &= \textstyle\prod_j \mathcal{N}(z_{nj}^{\boldsymbol{H}}; \mu_{nj}^{\boldsymbol{H}}, {\sigma_{nj}^2}^{\boldsymbol{H}}) && (6)
\end{aligned}
$$

式 (2) において 1 項目の期待値の計算に必要な積分は解析的に計算できないため，文献 [16] で述べられているモンテカルロ積分を用いて近似する．

以下のように図 2 のネットワークを学習する．

(1) 観測信号のスペクトログラム $\boldsymbol{Y}$ に NMF を適用し，基底 $\boldsymbol{W}_{\mathrm{NMF}}$ および重み $\boldsymbol{H}_{\mathrm{NMF}}$ を得る

(2) F-RNN および T-RNN を独立に学習する
- $\boldsymbol{W}_{NMF}$ を固定して T-RNN のみ学習する
- $\boldsymbol{H}_{NMF}$ を固定して F-RNN のみ学習する

(3) すべてのニューラルネットワークのパラメータを固定し，潜在変数のみ最適化する

潜在変数の最適化には majorization-minimizatioin（MM）アルゴリズム [17] を用いる．アルゴリズムの詳細は付録 A.1 に示す．

5. 近似例

図 1 および図 2 のネットワークによりどのような近似結果が得られるか確認するために，本節では RWC 音楽データベース [18] に収録されたクラリネット音（311CLNOM）およびピアノ音（011PFNOM）から生成した 3 重音のスペクトログラムを低ランク近似する．観測信号のスペクトログラムを図 3(a) および図 4(a) に示す．観測信号のサンプリング周波数は 11025 [Hz] である．短時間フーリエ変換（STFT）のフレーム長およびフレーム周期はそれぞれ 1024 [point]，256 [point] であり，窓関数には Hamming 窓を用いる．また，観測信号のパワースペクトログラムを得た後で，パワースペクトログラムの最大値が 1 になるように正規化する．再帰ユニットには LSTM [14] および gated recurrent unit（GRU）[19] を用いる．基底の数は 3 とし，再帰ユニットの隠れ状態の次元は 30，VAE における潜在変数の次元は 15 とする．また，モンテカルロ積分に用いるサンプル数は 10 とする．このシミュレーションでは，正確な近似を得るために式 (2) の 1 項目のみを用いてパラメー

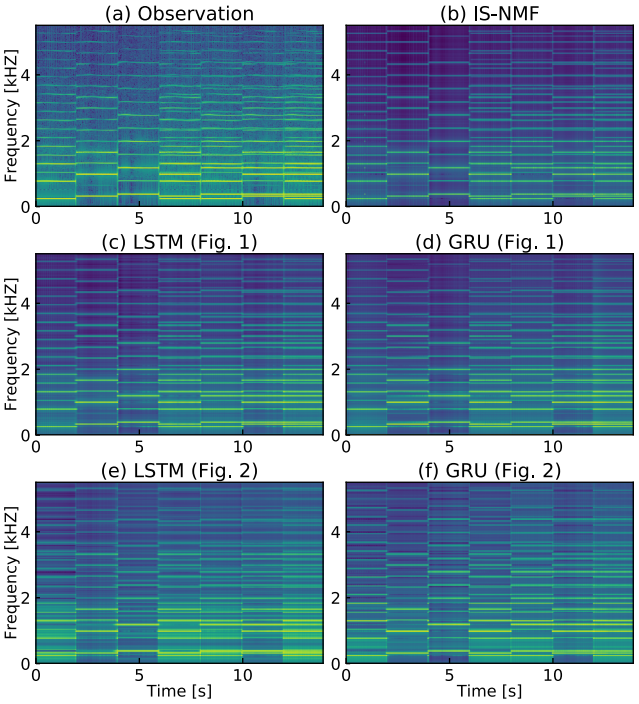

図 3　3 重音の近似例（Clarinet，311CLNOM）

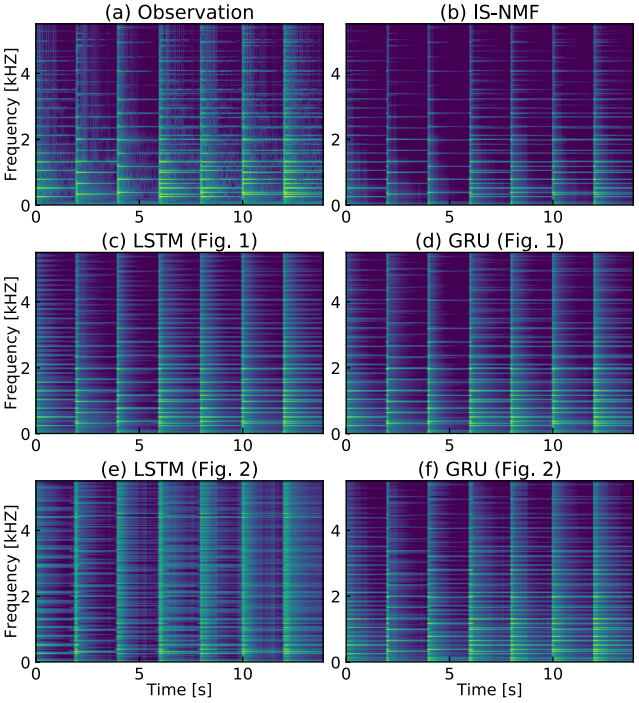

図 4　3 重音の近似例（Piano，011PFNOM）

タの勾配を計算する．

RNN の学習を効率化させるために，スペクトログラムの次元を圧縮したものを RNN に入力する．本節では，対数パワースペクトログラム $\tilde{\boldsymbol{Y}}$ に対して特異値分解を適用し，$\tilde{\boldsymbol{Y}} = \boldsymbol{X}_{\mathrm{F}}\boldsymbol{\Sigma}\boldsymbol{X}_{\mathrm{T}}^{\top}$ のように分解する．$\boldsymbol{X}_{\mathrm{F}}$ および $\boldsymbol{X}_{\mathrm{T}}$ から寄与率の大きなベクトルを 3 つ取り出し，それぞれ F-RNN, T-RNN の入力に用いる．

図 3 および図 4 にそれぞれクラリネット音，ピアノ音の近似結果を示す．これらの図において，図 3(b) および図 4(b) は複素正規分布に基づく NMF（IS-NMF） [17] の結果である．NMF では基底および重みがスパースになるような結果が得られる．しかし，提案モデルに基づく手法ではこのスパース性が緩和され，図 4(d) のように推定される重みはより連続性を持つようになる．基底の推定値においてはスペクトルのピーク形状が広がる傾向にあるため，今後，さらなる改善が求められる．

6. おわりに

本稿では，RNN を用いてスペクトログラムを低ランク近似するためのモデルを構築した．シミュレーションでは，3 重音の近似問題を通して提案したモデルに基づいてどのような近似表現を得られるか確認した．今後は大量の音楽音響データを用いたネットワークの学習や提案モデルに基づく音声強調や自動採譜のアルゴリズムの開発を行う．

謝辞　本研究は，JSPS 科研費（特別研究員奨励費 18J14238）の助成を受けた．

参考文献

[1] Lee, D. and Seung, H.: Learning the parts of objects with nonnegative matrix factorization, *Nature*, Vol. 401, No. 6755, pp. 788–791 (1999).
[2] Sawada, H., Kameoka, H., Araki, S. and Ueda, N.: Multichannel extensions of non-negative matrix factorization with complex-valued data, *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 21, No. 5, pp. 971–982 (2013).
[3] Kitamura, D., Ono, N., Sawada, H., Kameoka, H. and Saruwatari, H.: Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization, *IEEE/ACM Trans. Audio, Speech, and Language Processing*, Vol. 24, No. 9, pp. 1626–1641 (2016).
[4] Smaragdis, P. and Brown, J.: Non-negative matrix factorization for polyphonic music transcription, *Proc. 2003 IEEE International Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New York, USA, pp. 177–180 (2003).
[5] Bando, Y., Mimura, M., Itoyama, K., Yoshii, K. and Kawahara, T.: Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization, *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 716–720 (2018).
[6] Leglaive, S., Girin, L. and Horaud, R.: Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization, *Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 101–105 (online) (2019).
[7] Boulanger-Lewandowski, N., Mysore, G. and Hoffman, M.: Exploiting long-term temporal dependencies in NMF using recurrent neural networks with application to source separation, *Proc. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6969–6973 (2014).
[8] Le Roux, J., Hershey, J. and Weninger, F.: Deep NMF for speech separation, *Proc. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 66–70 (2015).
[9] Wisdom, S., Powers, T., Pitton, J. and Atlas, L.: Deep recurrent NMF for speech separation by unfolding iterative thresholding, *Proc. 2017 IEEE International Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 254–258 (2017).
[10] Virtanen, T.: Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria, *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 15, No. 3, pp. 1066–1074 (2007).
[11] Li, J., Mohamed, A., Zweig, G. and Gong, Y.: LSTM time and frequency recurrence for automatic speech recognition, *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 187–191 (2015).
[12] Liu, Y. and Wang, D.: Time and frequency domain long short-term memory for noise robust pitch tracking, *Proc. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5600–5604 (2017).
[13] Sainath, T. and Li, B.: Modeling time-frequency patterns with LSTM vs. convolutional architectures for LVCSR tasks, *Google AI*, (online), available from ⟨https://ai.google/research/pubs/pub45401⟩ (2016).
[14] Hochreiter, S. and Schmidhuber, J.: Long short-term memory, *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780 (1997).
[15] Févotte, C., Bertin, N. and Durrieu, J.: Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis, *Neural Computation*, Vol. 21, No. 3, pp. 793–830 (2008).
[16] Kingma, D. and Welling, M.: Auto-encoding variational Bayes, *Proc. 2nd International Conference on Learning Representations (ICLR)*, Alberta, Canada (2014).
[17] Févotte, C.: Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization, *Proc. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1980–1983 (2011).
[18] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC music database: popular, classical, and jazz music databases, *Proc. 3rd International Conference on Music Information Retrieval (ISMIR)*, Paris, France, pp. 287–288 (2002).
[19] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation, *Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, pp. 1724–1734 (2014).
[20] Parikh, N. and Boyd, S.: Proximal algorithms, *Foundations and Trends in Optimization*, Vol. 1, No. 3, pp. 127–239 (2014).

付　　録

A.1 潜在変数の最適化

本節では，REF. において潜在変数 $\boldsymbol{z}_m^{\boldsymbol{W}}$，$\boldsymbol{z}_n^{\boldsymbol{H}}$ を最適化する方法を述べる．手順 (2) においてニューラルネットワー

クのパラメータが得られた後，手順 (3) では推定された変分事後分布 q を用いて次式を最大化する $\boldsymbol{z}_m^{\boldsymbol{W}}$，$\boldsymbol{z}_n^{\boldsymbol{H}}$ を求める．

$$\mathcal{F}(\boldsymbol{z}_m^{\boldsymbol{W}}, \boldsymbol{z}_n^{\boldsymbol{H}}) = \frac{1}{MN}\sum_{m,n}\log p(y_{mn}|\boldsymbol{z}_m^{\boldsymbol{W}}, \boldsymbol{z}_n^{\boldsymbol{H}}) + \frac{1}{M}\sum_{m}\log q(\boldsymbol{z}_{mj}^{\boldsymbol{W}}) + \frac{1}{N}\sum_{n}\log q(\boldsymbol{z}_{nj}^{\boldsymbol{H}}) \tag{A.1}$$

式 (A.1) の 1 項目は損失関数，2 項目および 3 項目は罰則項とみなせる．したがって，1 項目を最大にする潜在変数を求めた後で，求められた潜在変数に 2 項目および 3 項目の近接作用素 [20] を適用することで式 (A.1) を最大化できる．

式 (A.1) の 1 項目は IS divergence と等価であることから，式 (A.1) の 1 項目を最大にする w_{mk} は以下の反復更新式で与えられる [17].

$$w_{mk} \leftarrow w_{mk}\sqrt{\frac{\sum_n \frac{y_{mn}h_{kn}}{\hat{y}_{mn}^2}}{\sum_n h_{kn}/\hat{y}_{mn}}} \tag{A.2}$$

w_{mk} を出力する全結合層の重みを $\boldsymbol{\Omega}^{\boldsymbol{W}}$, バイアスを $\boldsymbol{b}^{\boldsymbol{W}}$ とすると, $\boldsymbol{w}_m = [w_{m1}, \ldots, w_{mK}]^\top$ は $\boldsymbol{w}_m = \mathrm{softplus}(\boldsymbol{\Omega}^{\boldsymbol{W}}\boldsymbol{z}_m^{\boldsymbol{W}} + \boldsymbol{b}^{\boldsymbol{W}})$ のように求められる．ここで，$\mathrm{softplus}(\boldsymbol{x})$ は softplus 関数で，関数の値は入力ベクトルの要素ごとに計算される．softplus 関数の逆関数 $\mathrm{softplus}^{-1}$ を用いれば，式 (A.1) の 1 項目を最大にする $\boldsymbol{z}_m^{\boldsymbol{W}}$ は次式で与えられる．

$$\boldsymbol{z}_m^{\boldsymbol{W}} = (\boldsymbol{\Omega}^{\boldsymbol{W}})^{+}(\mathrm{softplus}^{-1}(\boldsymbol{w}_m) - \boldsymbol{b}^{\boldsymbol{W}}) \tag{A.3}$$

ここで，$(\boldsymbol{\Omega}^{\boldsymbol{W}})^{+}$ は $\boldsymbol{\Omega}^{\boldsymbol{W}}$ の一般逆行列である．式 (A.3) で求めた $\boldsymbol{z}_{mj}^{\boldsymbol{W}}$ に $\log q(z_{mj}^{\boldsymbol{W}})$ の近接作用素を適用する．すなわち，$z_{mj}^{\boldsymbol{W}}$ は次式によって更新される．

$$z_{mj}^{\boldsymbol{W}} \leftarrow \frac{\gamma\mu_{mj}^{\boldsymbol{W}} + \sigma_{mj}^{2\ \boldsymbol{W}} z_{mj}^{\boldsymbol{W}}}{\gamma + \sigma_{mj}^{2\ \boldsymbol{W}}} \tag{A.4}$$

ここで γ は正の定数である．$\boldsymbol{z}_n^{\boldsymbol{H}}$ についても同様にして最適化できる．