時間領域低ランクスペクトログラム近似法に基づくマスキング音声の欠損成分復元*

◎関翔悟（名大），亀岡弘和（NTT），戸田智基，武田一哉（名大）

1 はじめに

音声強調は雑音もしくは非目的音声の重畳した混合音から目的音を抽出し分離する技術であり，音声認識における前段処理等で利用される．その有効な手法の一つとして，時間周波数マスクに基づく手法がある．この手法では，混合音のスペクトログラムの各時間周波数点においてどの音源が最も優勢であるかを識別することで，同一音源に属するとみなされた時間周波数点の成分のみを通過させるような時間周波数マスクを構成する．例えば，深層ニューラルネットワークに基づく手法 [1] を用いることで高い信号対雑音比 (Signal to Noise Ratio; SNR) が得られる音声強調処理が可能である．一方で，強調音声のスペクトログラムは，雑音源が優勢であるとみなされた時間周波数点の成分が欠損したものとなるため，しばしば後段で用いられる音声情報処理の性能劣化を招く．したがって，欠損があるスペクトログラムからいかにして欠損した成分を復元するかが重要課題となる．

スペクトログラムの欠損成分を復元する手法として非負値行列因子分解 (Non-negative Matrix Factorization; NMF) に基づく方法が提案されている [2]．この方法は，非負値行列とみなせる振幅スペクトログラムを，二つの非負値行列の積で非欠損箇所において近似することにより欠損箇所の成分を補完しようというものである．これは，振幅スペクトログラムが低ランクな行列で近似できるという仮定を手がかりとしていることに相当する．この方法では欠損箇所における振幅成分しか推定されないため，信号を構成するためには別手法で位相成分も推定する必要がある．

本稿では，時間周波数マスキングにより生じた音声スペクトログラムにおける欠損成分の復元を目的とし，時間領域低ランクスペクトログラム近似法 (Time-domain Spectrogram Factorization; TSF) [6] に基づく欠損成分復元法を提案する．提案法では，従来法で考慮される振幅スペクトログラムの低ランク構造の仮定の下，

- 時間周波数表現が時間領域信号の冗長表現であることより導かれる各時間周波数点の振幅成分および位相成分の局所的な制約条件
- 目的音源のケプストラム特徴量の事前情報

のいずれかまたは両方を新たな手がかりとして考慮し，欠損成分が復元された時間領域信号を直接推定する．理想バイナリマスクに基づく音声スペクトログラムを対象とした実験的評価により，提案法の有効性を示す．

2 問題設定

一部の時間周波数点の成分が欠損した複素スペクトログラムを $\boldsymbol{Y} \in \mathbb{C}^{K\times M}$ とし，その各要素を $Y_{k,m}$ と表す．ただし，$k \in \{1,\dots,K\}$ 及び $m \in \{1,\dots,M\}$ はそれぞれ周波数ビン，時間フレームを表すインデックスである．非欠損箇所の集合を Γ とすると，$\boldsymbol{Y}$ は欠損箇所において

$$Y_{k,m} = 0, \quad ((k,m) \notin \Gamma) \tag{1}$$

となっているものとする．

3 従来法：NMF に基づく欠損成分復元

[2] では，NMF により振幅スペクトログラムを二つの非負値行列の積（低ランクな非負値行列）で近似できると仮定し，$|Y_{k,m}|$ を非欠損箇所 Γ において

$$X_{k,m} = \sum_{l=1}^{L} H_{k,l}U_{l,m} \tag{2}$$

で表すことで，$X_{k,m}$ を欠損成分の推定値とする方法が提案されている．これは各フレームの振幅スペクトル $X_{k,m}$ を L 個の基底振幅スペクトル $H_{k,1},\dots,H_{k,L}$ と非負係数 $U_{1,m},\dots,U_{L,m}$ による線形和で表現することに相当する．

NMF に基づく欠損成分復元手法は，対象とする振幅スペクトログラムの大域的な構造（すなわち低ランク構造）に関する仮定を手がかりに欠損成分を補完しようとするものである．推定された振幅スペクトログラムに対して位相復元手法 [4] による後段処理を行うことで，複素スペクトログラムを復元する．

4 提案法：TSF に基づく欠損成分復元

提案法では欠損成分が復元された複素スペクトログラムに対応する時間領域信号 $\boldsymbol{s} \in \mathbb{R}^N$（$N$ は全サンプル点数）を直接推定する．$\boldsymbol{s}$ の時間周波数表現（短時間 Fourier 変換やウェーブレット変換など）は一般に $\boldsymbol{\psi}_{k,m}^{\mathsf{H}}\boldsymbol{s}$ と表される．ここで，$\boldsymbol{\psi}_{k,m} \in \mathbb{C}^N$ は中心周波数が ω_k の複素正弦波に対して，時間フレーム t_m に対応する時間窓を掛けあわせたものである．

複素スペクトログラム $\boldsymbol{\psi}_{k,m}^{\mathsf{H}}\boldsymbol{s}$ は，$N < KM$ の場合には $\boldsymbol{s}$ の冗長表現になることから，非欠損箇所の $Y_{k,m}$ の振幅および位相を，周辺の欠損箇所の振幅・位相成分を推定する手がかりとすることができる．また，対象スペクトログラムのケプストラム特徴量の事前情報が得られる場合には，その統計分布も欠損成分を補完するための手がかりとなりうる．以上より一部欠損したスペクトログラムから時間領域信号を直接構成する問題は

$$\begin{aligned}\mathcal{I}(\boldsymbol{\theta}) = &\sum_{(k,m)\in\Gamma} |\boldsymbol{\psi}_{k,m}^{\mathsf{H}}\boldsymbol{s} - Y_{k,m}|^2 \\ &+ \lambda_1 \sum_{k,m} \mathcal{D}.\left(|\boldsymbol{\psi}_{k,m}^{\mathsf{H}}\boldsymbol{s}| \,|\, \sum_l H_{k,l}U_{l,m}\right) \\ &+ \lambda_2 \sum_{(k,m)\in\Gamma} \mathcal{D}.\left(|Y_{k,m}| \,|\, \sum_l H_{k,l}U_{l,m}\right) \\ &- \lambda_3 \mathcal{K}(\boldsymbol{\mathcal{X}})\end{aligned} \tag{3}$$

を最小化する最適化問題として定式化することができる．ただし，λ_1, λ_2, λ_3 は正則化パラメータ，$\boldsymbol{\theta} = \{\boldsymbol{s}, \boldsymbol{H}, \boldsymbol{U}\}$ は推定したい未知パラメータの集合であり，$\boldsymbol{H} = \{H_{k,l}\}$，$\boldsymbol{U} = \{U_{l,m}\}$ である．$\mathcal{D}.$ は誤差関数であり，ここでは Euclid 距離（二乗誤差）

$$\mathcal{D}_{\mathrm{EU}}(x|y) = (y-x)^2 \tag{4}$$

---

*Missing Component Restoration for Masked Speech Signals Based on Time-domain Spectrogram Factorization by SEKI, Shogo (Nagoya University), KAMEOKA, Hirokazu (NTT), TODA, Tomoki, TAKEDA, Kazuya (Nagoya University)

Table 1: Experimental condition

サンプリング周波数	16 kHz
フレームサイズ	32 ms
シフトサイズ	16 ms
NMF 基底数	30
更新回数	100 回
フィルタバンク	20 次
MFCC	0–13 次
GMM 混合数	30

または Kullback-Leibler (KL) ダイバージェンス

$$\mathcal{D}_{\mathrm{KL}}(x|y) = y \log \frac{y}{x} - (y - x) \tag{5}$$

を用いる．音源の時間領域信号 $\boldsymbol{s}$ とその振幅スペクトログラムの低ランク表現 $\{\boldsymbol{H}, \boldsymbol{U}\}$ を推定する枠組みは時間領域低ランクスペクトログラム近似法 (Time-domain Spectrogram Factorization; TSF) [6] と同様である．

式 (3) の第一項は観測複素スペクトログラムと $\boldsymbol{s}$ の複素スペクトログラムの非欠損領域における二乗誤差，第二項は式 (2) と $\boldsymbol{s}$ の振幅スペクトログラムの誤差の大きさ，第三項は式 (2) と観測振幅スペクトログラムの誤差の大きさを表す．第四項は以下で与えられる正則化項である．

$$\mathcal{K}(\boldsymbol{\mathcal{X}}) = \log \prod_m \sum_p w_p \prod_q \mathcal{N}\left(\mathcal{X}_{q,m}; \mu_{p,q}, \sigma^2_{p,q}\right) \tag{6}$$

$$\mathcal{X}_{q,m} = \sum_r c_{q,r} \log \sum_k f_{r,k} X_{k,m} \tag{7}$$

ただし $\mathcal{X}_{q,m}$ はメル周波数ケプストラム係数（Mel-Frequency Cepstrum Coefficients; MFCC）であり，$f_{r,k}$ は $r(\in \{1, \ldots, R\})$ 番目のメルフィルタバンク係数，$\{c_{q,r}\}_{0 \le q \le Q-1, 1 \le r \le R}$ は逆離散コサイン変換係数である．式 (6) はパラメータ $\left\{w_p, \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p\right\}_{1 \le p \le P}$ の混合ガウス分布 (Gaussian Mixture Model; GMM) の対数尤度を表し，ケプストラム距離正則化項として知られている [5]．ただし，w_p，$\boldsymbol{\mu}_p = (\mu_{p,1}, \ldots, \mu_{p,Q})^{\mathsf{T}}$，$\boldsymbol{\Sigma}_p = \mathrm{diag}\left(\sigma^2_{p,1}, \ldots, \sigma^2_{p,Q}\right)$ は p 番目のガウス分布の重み，平均及び分散を表す．

TSF に基づく欠損成分の推定では，NMF に基づく従来法と同様に観測スペクトログラムの大域的制約を考慮するのに加えて，観測複素スペクトログラムの欠損成分と周囲の成分の依存関係に基づく局所的制約，音声の事前情報がある場合は音声の特徴量空間上での分布に対する制約を考慮することができる．

5　実験的評価

理想バイナリマスク (Ideal Binary Mask; IBM) によりマスキングされた音声に対して，提案法の有効性を調査する．音素バランス 503 文 A セットの男性話者 1 名による計 10 発話に対して，Babble ノイズを SNR を変化させて重畳し，各発話に対して理想バイナリマスクを適用することで，マスキング音声を作成する．提案法として式 (3) に対して Euclid 二乗誤差規準（EU-TSF），ケプストラム距離正則化項の有無を考慮した KL-divergence 規準（KL-TSF w/o Reg., KL-TSF w/ Reg.）の場合を評価する．また従来法として，式 (3) における第三項，第四項で表される NMF に基づく欠損成分復元手法 [2] を，提案法と同様の場合（EU-NMF，KL-TSF w/o Reg.，KL-TSF w/ Reg.）について評価する．ただし，位相復元アルゴリズム [4] についての更新回数は 100 回とする．評価尺度は発話音声に対する復元音声の SNR 及び発話音声と復元音声間の MFCC 距離を用いる．実験条件

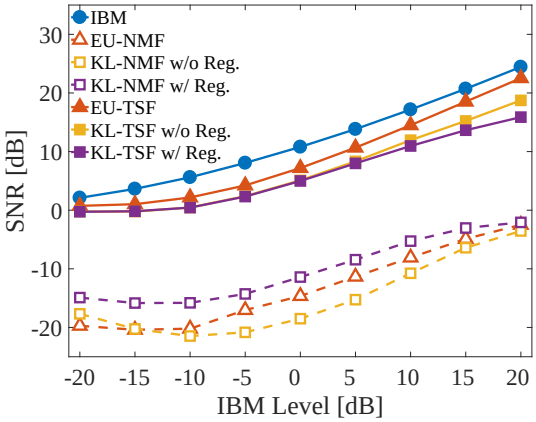

Fig. 1: Experimental result: SNR

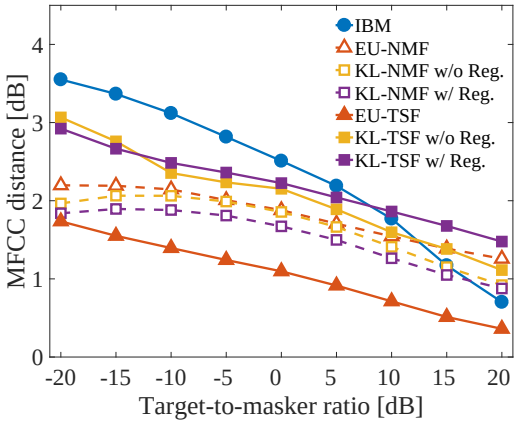

Fig. 2: Experimental result: MFCC distance

を Table. 1 に示す．ケプストラム距離正則化項における GMM パラメータは同一話者の異なる計 100 発話から学習する．

実験結果を Figs. 1，2 に示す．それぞれ SNR，MFCC 距離の結果であり理想バイナリマスクによる結果 (IBM) についても示されている．実験結果より，TSF に基づく提案法が NMF に基づく従来法に比べ高性能であることが確認される．また，提案法では理想バイナリマスクと同様に高い SNR を満たす一方で，欠損成分により生じる MFCC 距離の改善が確認される．

6　おわりに

本稿ではマスキング音声に対するスペクトログラムの欠損成分復元について，TSF に基づく欠損成分復元手法を提案した．欠損成分の復元において，従来法で考慮される観測スペクトログラムの低ランク表現に基づくスペクトログラムの大域的構造に加えて，時間周波数領域表現がもつ冗長性に基づく欠損成分周囲とのの局所的な依存関係や，目的音声がもつ特徴量の事前情報を考慮することが可能である．実験的評価により，TSF に基づく提案法が NMF に基づく従来法より高性能であることが確認されたとともに，提案法では理想バイナリマスクと同様の SNR をもつ一方，特徴量上での歪みを改善することを確認した．

参考文献

[1] Hershey, *et al.*, *ICASSP*, 31–35, 2016.
[2] Smaragdis, *et al.*, *JSPS*, 65(3), 361–370, 2010.
[3] Smaragdis, *et al.*, *WASPAA*, 177–180, 2003.
[4] Griffin & Lim, *IEEE TASSP*, 32(2), 236–243, 1984.
[5] Li *et al.*, *Interspeech*, 3753–3757, 2016.
[6] Kameoka, *ICASSP*, 86–90, 2015.