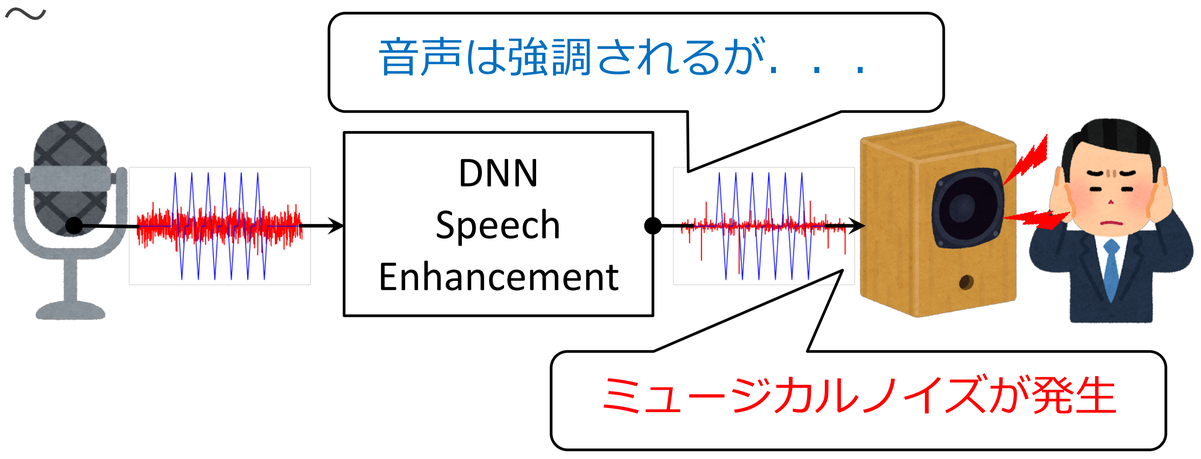
カートシスマッチングに基づく 低ミュージカルノイズ DNN 音声強調の評価

溝口 聡，齋藤 佑樹，高道 慎之介，猿渡 洋（東京大学）

聴覚的に好印象な音声強調とは？

- 音声強調 ～聴き心地の良いハンズフリー音声通信のために～
  - 聴覚的に好印象な音声強調が満たすべき条件
    - 雑音抑圧性能：高い
    - 音声の歪み発生量：小さい
    - ミュージカルノイズ発生量：小さい
- DNN 音声強調
  - DNN の表現力の高さによって，強力な雑音抑圧を達成可能
  - しかし，ミュージカルノイズが発生する可能性あり
    - ミュージカルノイズ：聴覚的に不愉快な残留雑音



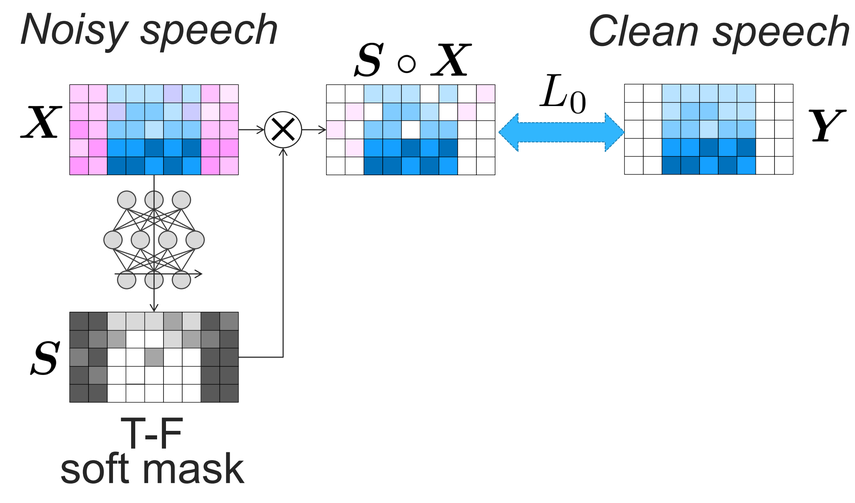
本発表の主題：低ミュージカルノイズな DNN 音声強調の実現方法とその実験的評価

従来法：ソフトマスクベースの DNN 音声強調 [1]

- 入力：観測信号のスペクトログラム
- 出力：雑音抑圧のための時間周波数ソフトマスク
- 損失関数：ターゲットの音声と強調後の音声の距離

$$L_0\left(\boldsymbol{S}\circ\boldsymbol{X},\boldsymbol{Y}\right)=\left\|\boldsymbol{S}\circ\boldsymbol{X}-\boldsymbol{Y}\right\|_{1,1}$$

- 問題点：ミュージカルノイズの発生が考慮されていない
  - ミュージカルノイズ [2] とは
    - 非線形な信号処理によって発生するノイズ
    - その正体は推定エラーによるアーチファクト
    - 音程を持っているため聴覚的に不愉快



提案法：カートシスマッチングによる正則化つき DNN 音声強調

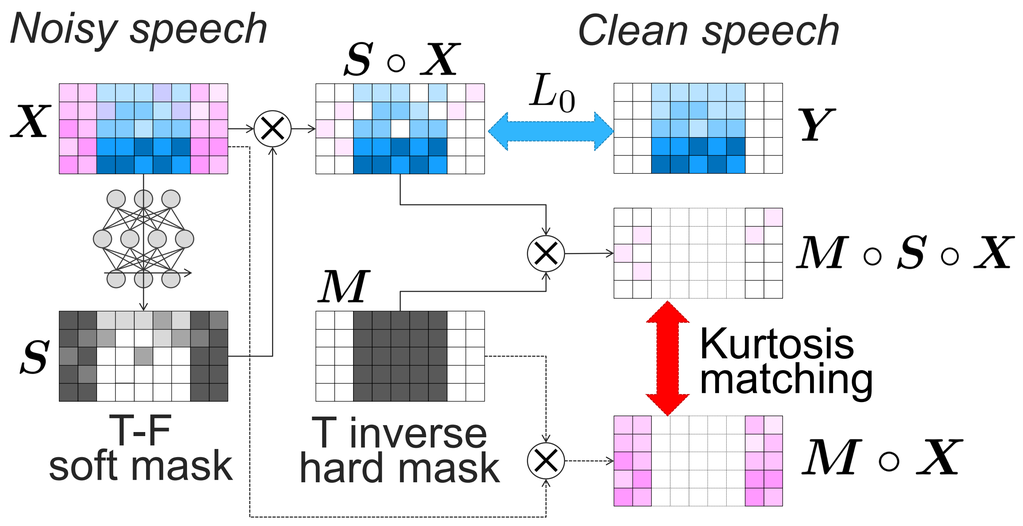
- カートシス

$$\mathrm{kurt}\left(\boldsymbol{X}\right)=\frac{\sum_{k,t}\left(X_{k,t}-\mathbb{E}_{k,t}\left[X_{k,t}\right]\right)^4}{\left(\sum_{k,t}\left(X_{k,t}-\mathbb{E}_{k,t}\left[X_{k,t}\right]\right)^2\right)^2}$$

  - 処理前後の変化量とミュージカルノイズの発生量と正の相関あり [3]
- カートシスマッチング
  - ターゲット信号から特定した非音声区間について下式により正則化

$$L(\boldsymbol{S}\circ\boldsymbol{X},\boldsymbol{Y})=L_0(\boldsymbol{S}\circ\boldsymbol{X},\boldsymbol{Y})+\lambda\left|\frac{\mathrm{kurt}\left(\boldsymbol{M}\circ\boldsymbol{X}\right)-\mathrm{kurt}\left(\boldsymbol{M}\circ\boldsymbol{S}\circ\boldsymbol{X}\right)}{\mathrm{kurt}\left(\boldsymbol{M}\circ\boldsymbol{X}\right)}\right|$$

  - カートシスの変動を抑制してミュージカルノイズの発生を低減！



実験的評価

実験条件

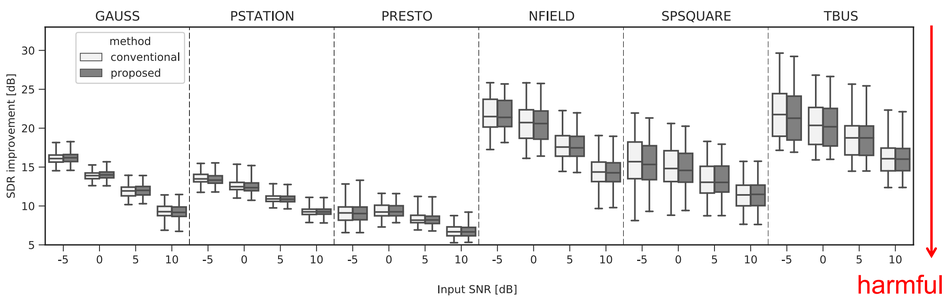
学習データ	JNAS より 31890 文
テストデータ	JSUT より 200 文 × 24
サンプルレート	16 kHz
雑音	DEMAND より 5 種 + ガウス性雑音
入力 SN 比	-5, 0, 5, 10 dB
窓関数	Hanning
FFT 長	1024
ホップ 長	80
DNN アーキテクチャ （詳細な構造）	U-Net [4] （[5] に倣う）
パッチ長	256
最適化手法	Adam
バッチサイズ	32

使用した雑音とそのカートシス

GAUSS	PSTATION	PRESTO	NFIELD	SPSQUARE	TBUS
3.00	5.56	12.1	13.3	29.8	35.8

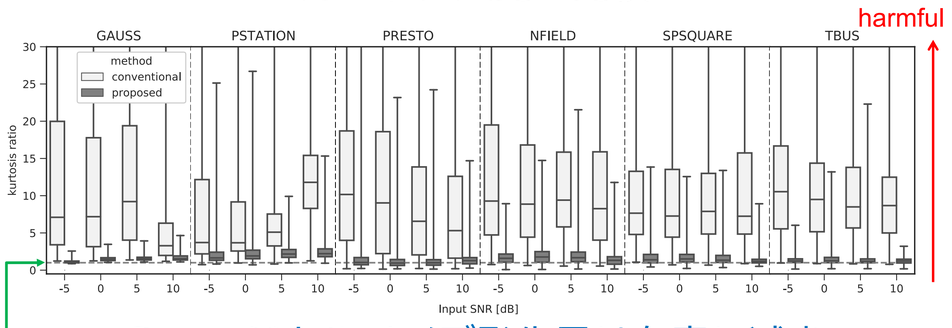
1 が理想値

Signal-to-Distortion Ratio 改善量の比較



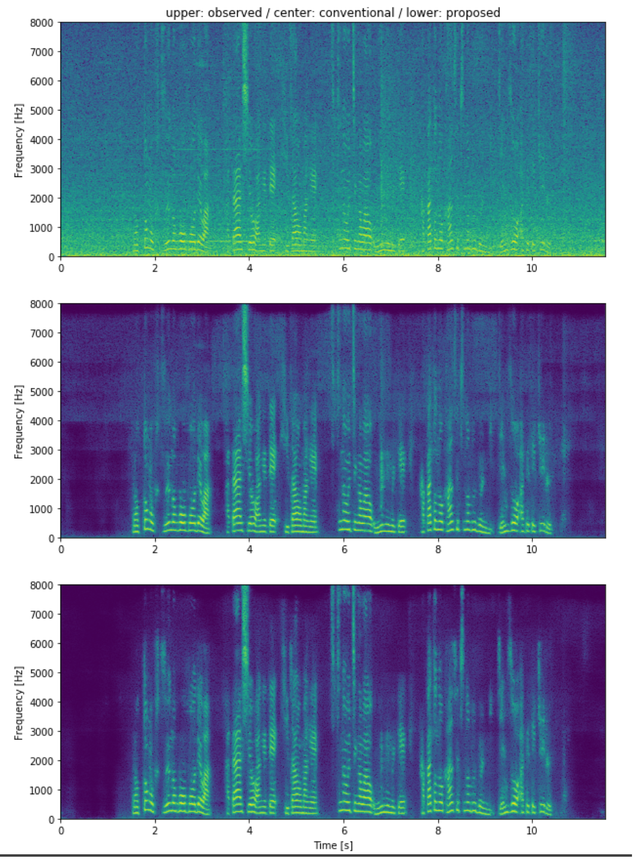
雑音抑圧性能，音声歪み発生量に有意差なし

カートシス比の比較



ミュージカルノイズ発生量は有意に減少

強調音声のスペクトログラム



参考文献

[1] Y. Xu et al., IEEE/ACM Trans. on ASLP, vol. 23, no. 1, pp. 7–19, Jan. 2015.
[2] S. Boll, IEEE Trans. ASSP, vol. 27, no. 2, pp. 113–120, 1979.
[3] Y. Uemura et al., Proc. IWAENC, 2008.
[4] O. Ronneberger et al., Proc. MICCAI, 2015, pp. 234–241.
[5] N. Jansson et al., Proc. ISMIR, 2017, pp. 745–751.