

ソースフィルタ非負値行列因子分解によるボコードを用いない声質変換の実験的検討

2-Q-26 ☆ 須田仁志 齋藤大輔 峯松信明 (東京大学)

非負値行列因子分解 (NMF)

[Lee+, 2000]

- ▶ 非負値行列を、2つの非負値行列の積として近似する

$$X \approx AH$$

- ▶ X としてパワースペクトログラムを考えると、NMF はある時刻のスペクトル x_t を I 個の基底 $a_1, a_2, \dots, a_I$ の和で表すことにあたる

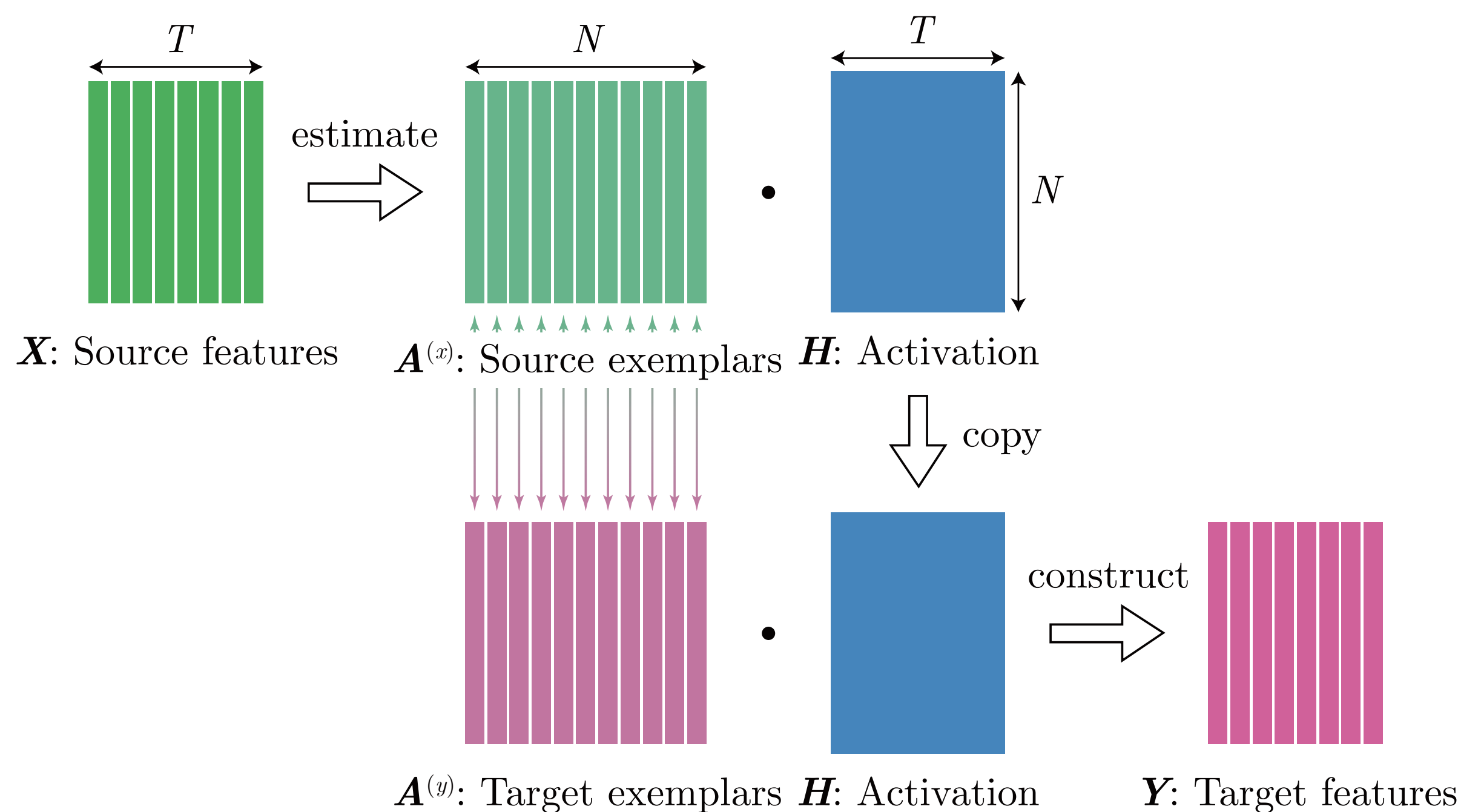
$$x_t \approx \sum_i h_{i,t} a_i$$

- ▷ 生起状態 $h_{i,t}$ は時刻 t での基底 a_i の重みにあたる
- ▶ 補助関数法を用いることで近似アルゴリズムを導ける
 - ▷ 目標行列との距離 $D(X | AH)$ を最小に近づける反復計算を行う
 - ▷ 同じ手法でより複雑な NMF の発展モデルを構築することが可能

NMF による統計的声質変換

[Takashima+, 2012]

- ▶ 発話音声の中のあるフレームの (パワー) スペクトル包絡が“いくつかのパターン”と“そのパターンの重み”で表せると仮定
 - ▷ パターンは話者に対して用意される (時間非依存)
 - ▷ パターンの各時刻での重みが発話内容を表す
 - ▷ パワースペクトルは非負の特徴量なので、パターンを基底に、その重みを生起状態とみなすことで NMF を用いることができる
- ▶ 学習時には、パラレルデータに対して両者の生起状態が一致する仮定をして、それぞれの基底を推定する
 - ▷ これにより両話者の基底の対応付けを行うことができる
- ▶ 変換時には、入力音声のスペクトル包絡を、学習済み入力話者基底と生起状態に分解し、得られた生起状態と出力話者基底を掛け合わせて目標となるスペクトル包絡が推定できる



▲ NMF による統計的声質変換の概念図

生起状態フリー学習

- ▶ パラレルデータから基底を学習する際、通常は 2 話者のパラレルデータの生起状態は同じと仮定して固定する
- ▶ しかし DTW 後でも完全に時間的対応付けが取れているとはいえず、NMF ではこの時間的ミスマッチの影響を強く受けてしまう
- ▶ 両者の生起状態の距離を距離規準に含めることで、基底の対応の破綻を防ぎつつ、ミスマッチによる品質低下を抑える
 - ▷ たとえば目的関数を以下のように設定できる (μ は正の定数)

$$D = D(Y | A^{(y)} H^{(y)}) + \mu D(H^{(y)} | H^{(x)})$$

- ▷ 第 1 項は NMF の最適化規準を表す
- ▷ 第 2 項は生起状態が離れすぎないようにするための項
- ▷ 自然性は高まるが話者性は犠牲になってしまう

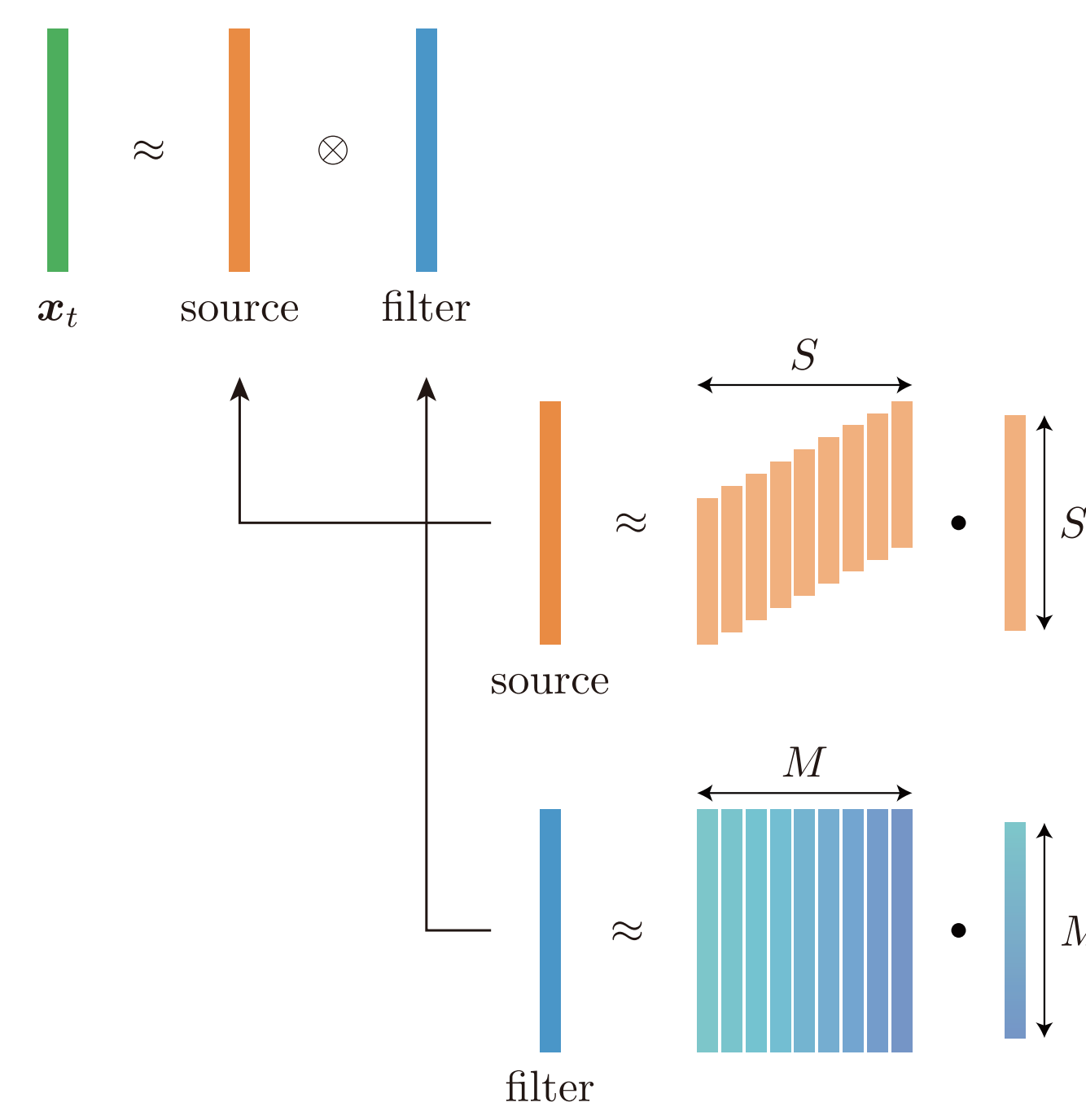
ソースフィルタ NMF (SF-NMF)

[Virtanen+, 2006] の簡略化

- ▶ ソースフィルタモデル「ソース $\otimes$ フィルタ」の概念を NMF で表す

$$x_t(k) \approx \left(\sum_{n,s=1}^{N,S} h_{n,s,t}^{(s)} a_n^{(s)}(k-s) \right) \left(\sum_{m=1}^M h_{m,t}^{(f)} a_m^{(f)}(k) \right)$$

- ▷ n はソース基底、 m はフィルタ基底のインデックスを表す
- s はソース基底の周波数方向のシフト量を表す
- ▷ ソースの基底をシフトでき、様々な F_0 を持つソーススペクトルを 1 つの基底で表現できる
- ▶ スペクトル包絡ではなく、調波構造を備えたスペクトルそのものを扱う



[SF-NMF を用いた声質変換]

- ▶ 本研究ではソースとフィルタそれぞれの基底の対応を学習
 - ▶ 変換時には NMF による変換と同様に生起状態をまず推定し、その生起状態と学習済み基底から目標スペクトログラムを再構成
- ▲ $N = 1$ のときの SF-NMF の概念図

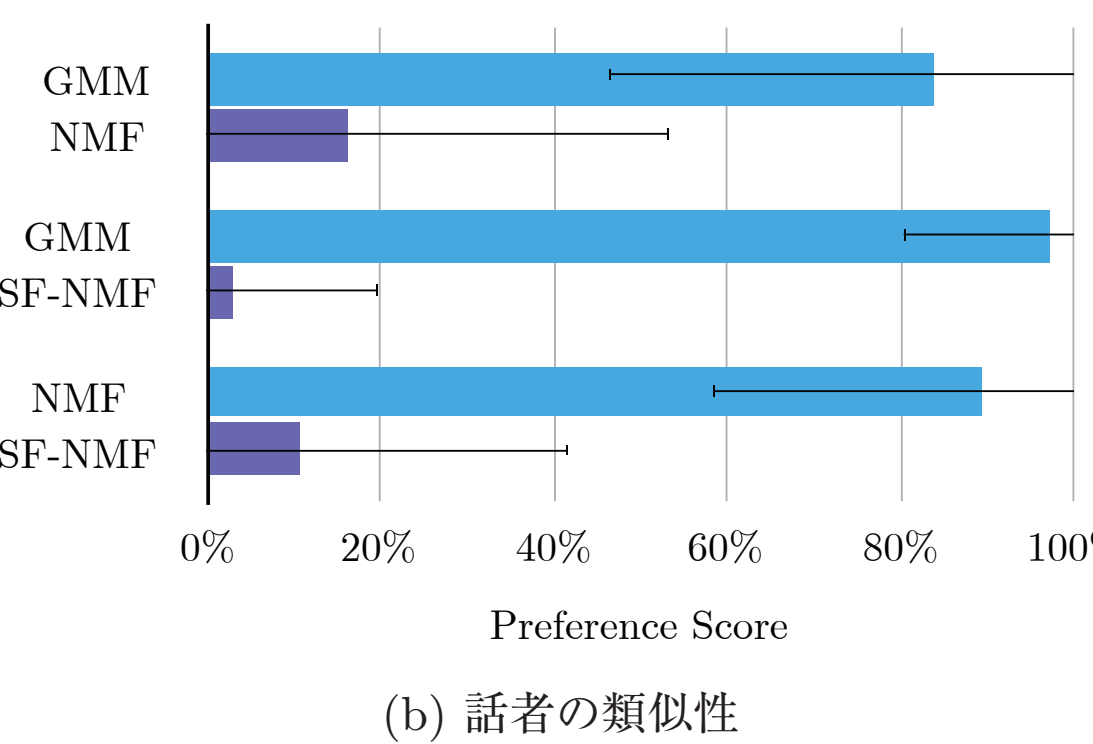
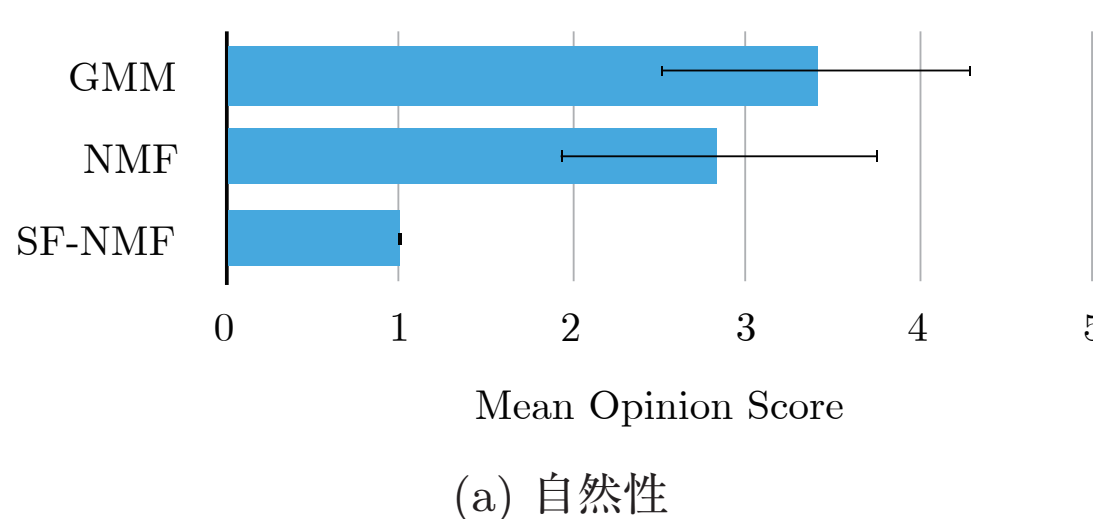
連続ウェーブレット変換と位相推定

[Nakamura+, 2014]

- ▶ SF-NMF で扱うスペクトルは対数スケールの周波数軸が必要
 - ▷ 周波数軸のスケールに依存しない連続ウェーブレット変換が有効
- ▶ 非負の実数を扱う NMF では位相情報を扱えないためパワースペクトルを推定したのち、位相推定を行うことで再合成
 - ▷ 目標スペクトルと再構成スペクトルの差を目的関数として最適化

変換音声の主観評価

- ▶ 対象音声は Voice Conversion Challenge 2016 [Toda+, 2016] を利用
- ▶ 話者は SM1 (男性) $\rightarrow$ TF1 (女性) ▶ 学習 162 文、テスト 54 文
- ▶ SF-NMF では WORLD 分析で得られた F_0 情報を利用してソースの生起状態を、包絡情報を利用してフィルタの基底をそれぞれ事前に推定し初期値に用いた
- ▶ NMF、SF-NMF ともに生起状態フリー学習を適用
- ▶ 自然性は 5 段階評価、話者類似性は XAB テストにより評価 (被験者 7 名)
- ▶ SF-NMF による変換音声の品質は既存手法のそれに及ばなかった
 - ▷ フィルタ基底のみ交換を行った場合より大きく品質が低下するためソース基底の対応が話者間で取れていないことが原因と考えられる



Sample: <https://www.gavo.t.u-tokyo.ac.jp/~hitoshi/sfnmf-vc/>

今後の展望

- ▶ SF-NMF を用いた声質変換を実現したが、品質には課題が残る
- ▲ SF-NMF によるよい高品質な音声のモデル化を検討する
- △ 声質変換の品質向上のための研究