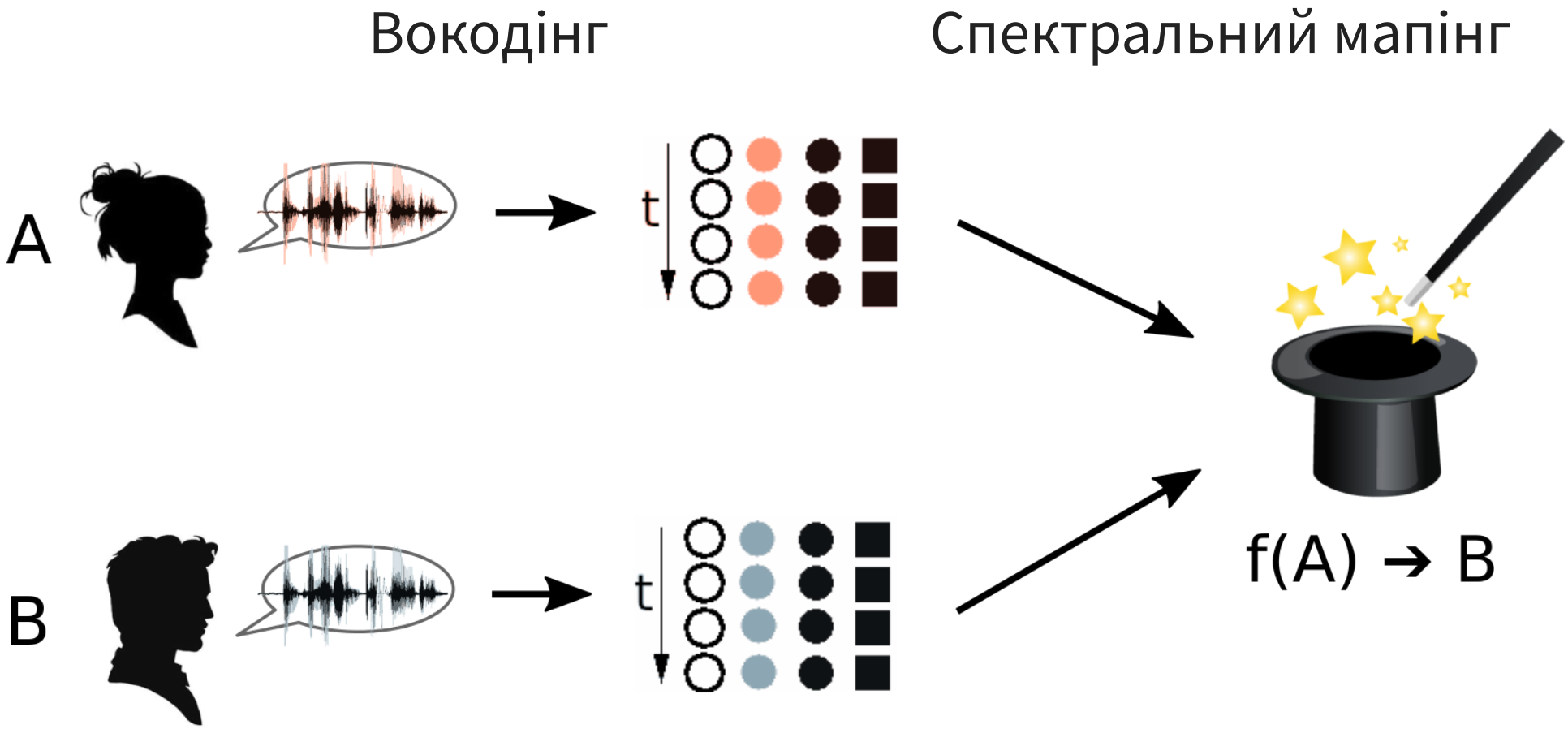
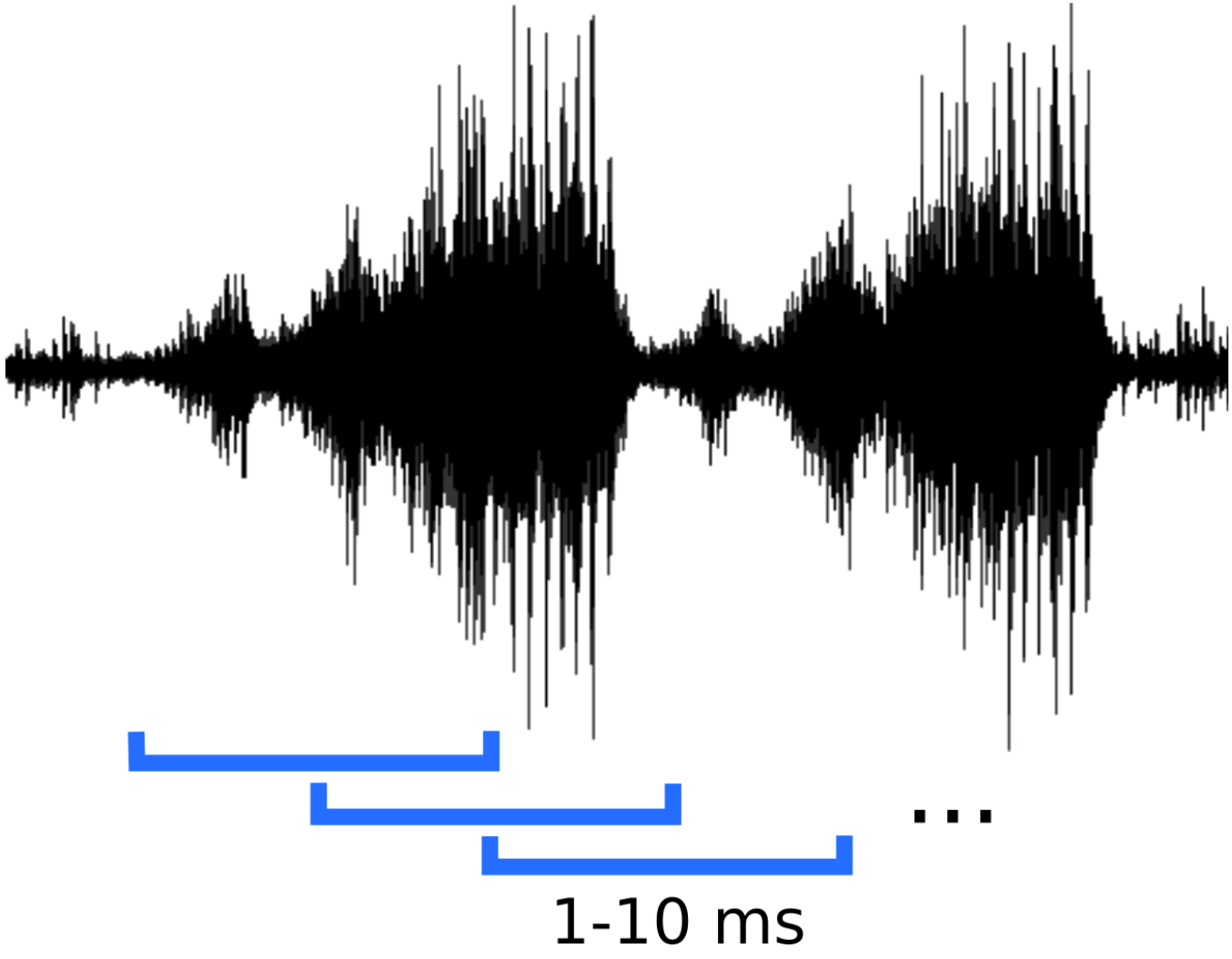
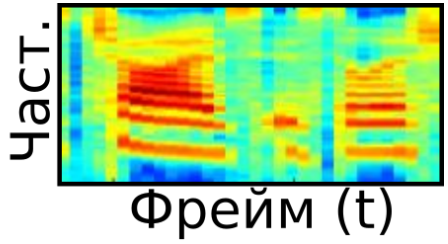
VOICE CONVERSION

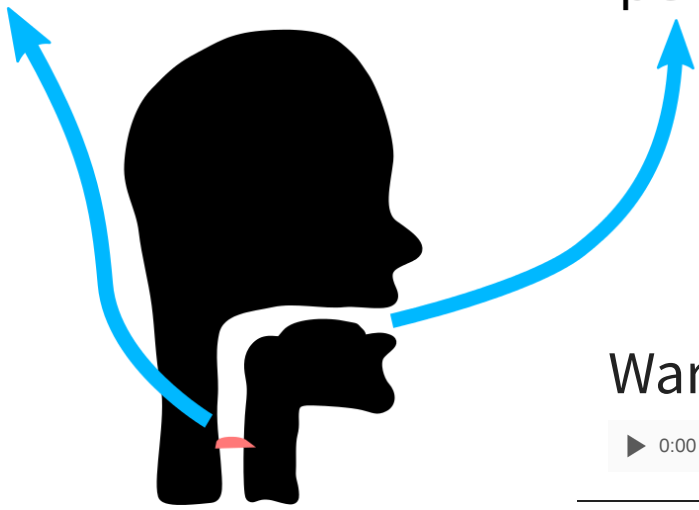
ЩО ЦЕ?



ВОКОДІНГ







Звичайний голос:

0:00 / 0:03

$F_0 = F_0 \cdot 0.5$:

0:00 / 0:03

$F_0 = F_0 \cdot 6$:

0:00 / 0:03

Крокова F_0:

0:00 / 0:03

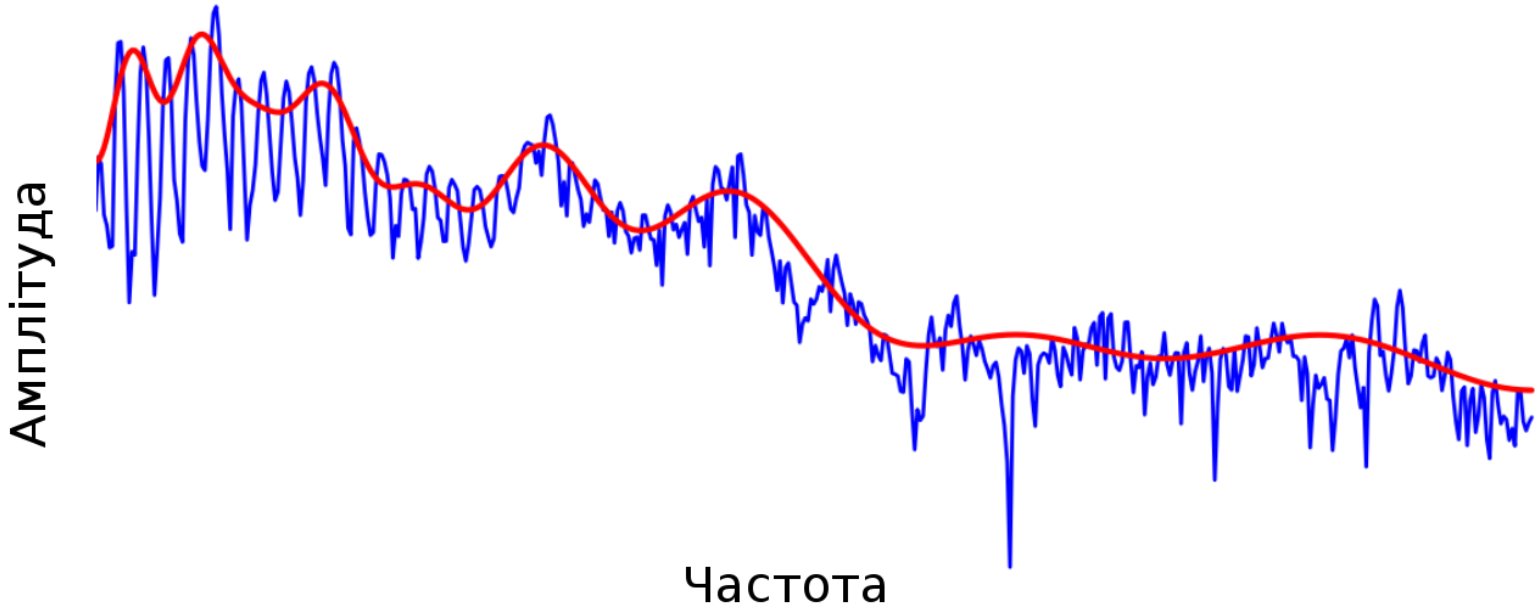
Warping #1:

0:00 / 0:03

Warping #2:

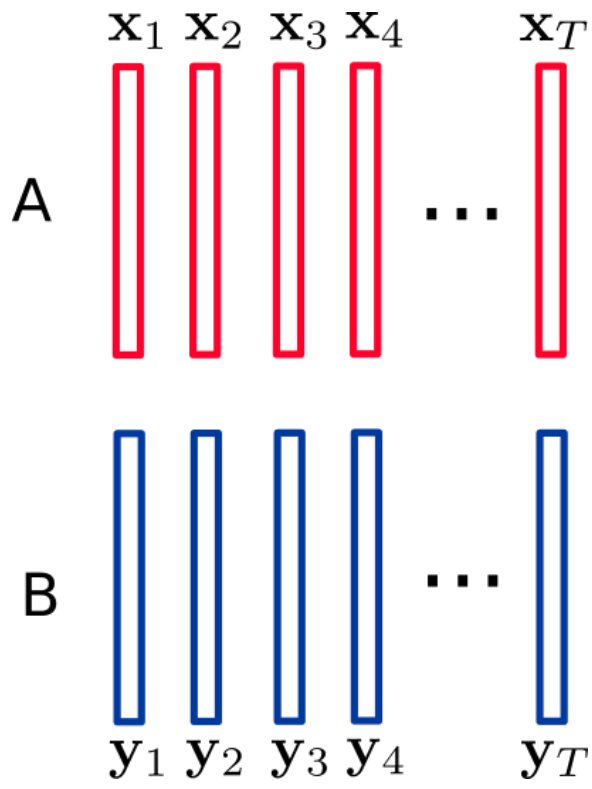
0:00 / 0:03

СПЕКТРАЛЬНІ ФІЧІ



Фіча	**Розмір**	**Алгоритми**
Спектральна густина:	$\mathbb{R}^{513}$	NMF, FW
MCEP (Mel-cepstrum envelope):	$\mathbb{R}^{24}$	GMM, Kernel, DNN

ПАРАЛЕЛЬНИЙ КОРПУС



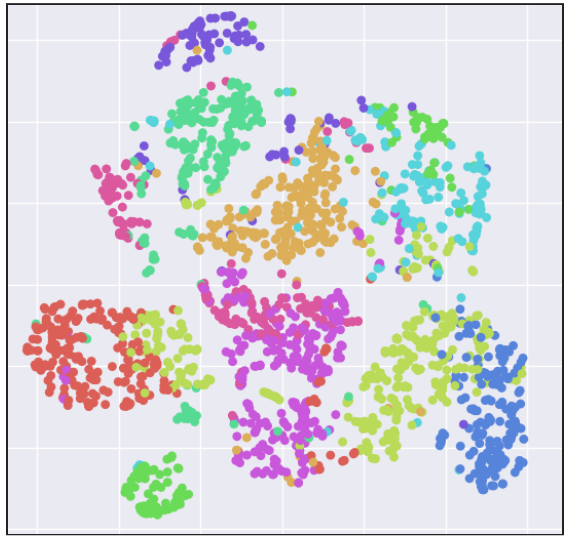
СПЕКТРАЛЬНИЙ МАПІНГ

- Наївне застосування регресії призводить до незадовільних результатів
- JDGMM: Joint Density Gaussian Mixture Models

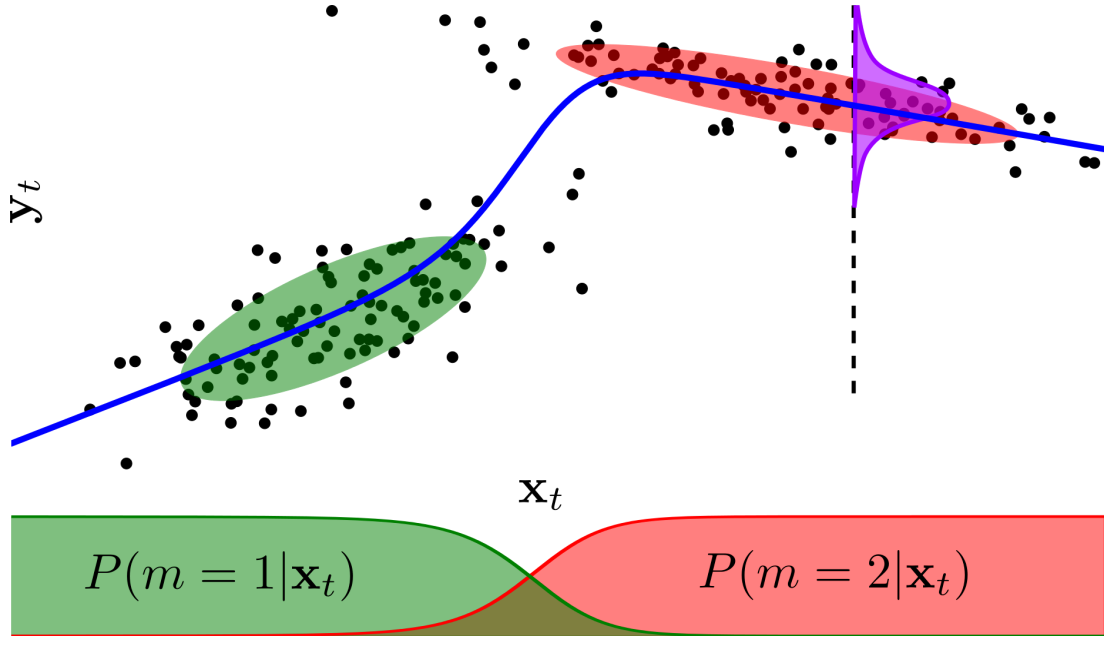
[Toda et al. 2007]

JOIN DENSITY GMM

- GMM - метод кластеризації
- Акустичний простір має природнє розділення на кластери



- JDGMM дозволяє моделювати локальну функцію перетворення для кожного кластера



$\{\mathbf{x}_t\}_1^T$ та $\{\mathbf{y}_t\}_1^T$ початковий та цільовий голоси та $\mathbf{z}_t := (\mathbf{x}_t, \mathbf{y}_t)$

$$P(\mathbf{x}_t, \mathbf{y}_t) = P(\mathbf{z}_t) = w_1 \mathcal{N}(\mathbf{z}_t; \mu_1^{(z)}, \mathbf{C}_1^{(zz)}) + w_2 \mathcal{N}(\mathbf{z}_t; \mu_2^{(z)}, \mathbf{C}_2^{(zz)})$$

$$\mu_m^{(z)} = \begin{bmatrix} \mu_m^{(x)} \\ \mu_m^{(y)} \end{bmatrix}; \mathbf{C}_m^{(z)} = \begin{bmatrix} \mathbf{C}_m^{(xx)} & \mathbf{C}_m^{(xy)} \\ \mathbf{C}_m^{(yx)} & \mathbf{C}_m^{(yy)} \end{bmatrix}$$

$$P(\mathbf{y}_t|\mathbf{x}_t) = P(m=1|\mathbf{x}_t)P(\mathbf{y}|\mathbf{x}_t, m=1) + P(m=2|\mathbf{x}_t)P(\mathbf{y}|\mathbf{x}_t, m=2)$$

$$\hat{\mathbf{y}}_t(\mathbf{x}_t) = E_y[P(\mathbf{y}_t|\mathbf{x}_t)]$$

НЕДОЛІКИ

- Не враховується динамічна природа сигналу
- Надмірне згладжування $\hat{\mathbf{y}}_t$

Приклад звуку:

ДИНАМІЧНІ ФІЧІ

1. Замість $\mathbf{y}_t \in \mathbb{R}^D$ використовуємо $\mathbf{Y}_t = (\mathbf{y}_t, \Delta\mathbf{y}_t) \in \mathbb{R}^{2D}$, де
$$\Delta\mathbf{y}_t := 0.5 \cdot (\mathbf{y}_{t-1} - \mathbf{y}_{t+1})$$

2. Тренуємо JDGMM $\implies P(\mathbf{Y}_t|\mathbf{X}_t)$

3. Збираємо послідовності у один вектор:
$\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_{T_{tst}}) \in \mathbb{R}^{DT_{tst}}$
$\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_{T_{tst}}) = (\mathbf{y}_1, \Delta\mathbf{y}_1, \ldots, \mathbf{y}_{T_{tst}}, \Delta\mathbf{y}_{T_{tst}}) \in \mathbb{R}^{2DT_{tst}}$
Можна записати: $\mathbf{Y} = \mathbf{W}\mathbf{y}$

4. Знаходимо найкращу (maximum likelihood) послідовність для цільового голосу:
$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\, P(\mathbf{W}\mathbf{y}|\mathbf{X}_{tst}) = (\mathbf{W}^T\mathbf{D}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{D}\mathbf{e}$$

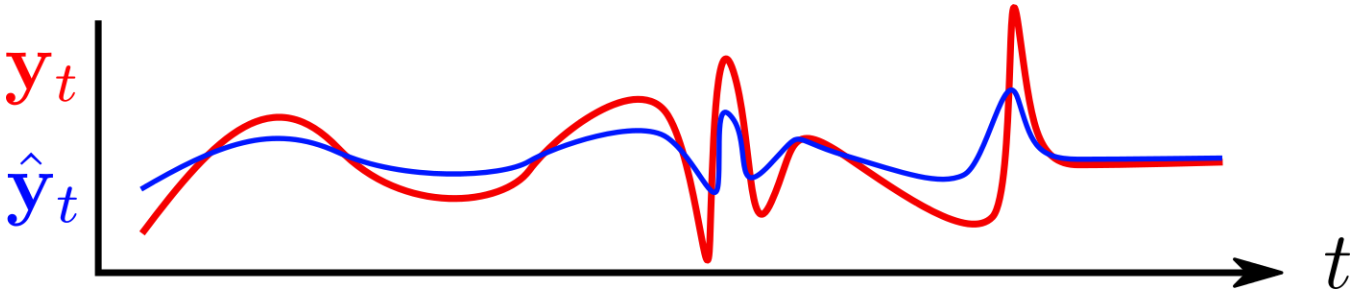
НЕДОЛІКИ

- ~~Не враховується динамічна природа сигналу~~
- Надмірне згладжування $\hat{\mathbf{y}}_t$

Приклад звуку:

Приклад звуку (ДФ):

ГЛОБАЛЬНА ДИСПЕРСІЯ



1. Моделюємо розподіл дисперсій цільового спікера простим нормальним розподілом: $P(\mathbf{v})$

2. Враховуємо дисперсію під час оптимізації:

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\, P(\mathbf{W}\mathbf{y}|\mathbf{X}) \cdot P(var(\mathbf{y}))$$

3. На жаль, немає аналітичного рішення для максимуму ¯_(ツ)_/¯

Приклад звуку:

Приклад звуку (DF):

Приклад звуку (DF + GV):

ІНШІ ПІДХОДИ

- Non-negative Matrix Factorization (NMF) - кожен новий фрейм моделюється як лінійна комбінація фреймів з тренувальної вибірки
- Frequency warping (FW) - кожен новий цільовий фрейм моделюється як викривлення відповідного фрейму початкового спікера
- Kernel Partial Least Squares - Лінійна регресія з нелінійним ядром
- Neural Networks

TAKE HOME MESSAGE:

Use domain knowledge, Luke!

СКРЯБІН

$\mathbf{x}_{train}$ 0:00 / 0:09

$\mathbf{y}_{train}$ 0:00 / 0:11

$\hat{\mathbf{y}}$ 0:00 / 0:27

THANKS!

bielievtsov@ibdi.pro