Изменение скрытых состояний в задаче переноса стиля в аудио

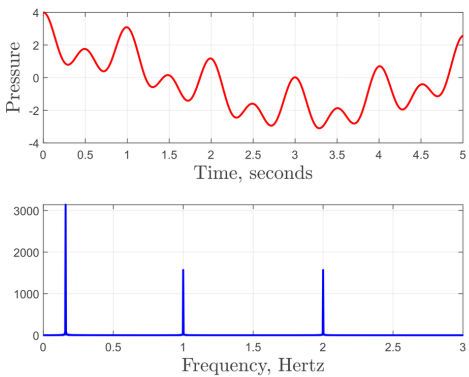
Анна Ремизова
научный руководитель: к. ф.-м. н. Андрей Грабовой

МФТИ

16/12/2023

Постановка задачи

- Пространство аудио(S) - дискретные сигналы конечного размера.



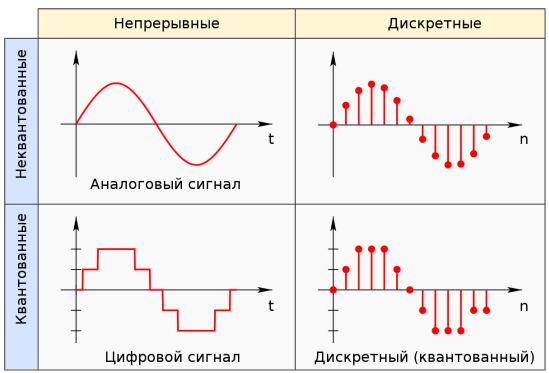


Рис. 1: Представление сигналов

Note

Любой дискретный сигнал длины N в пространстве времени может быть представлен однозначно конечным рядом синусоид.

Постановка задачи

For a length N complex sequence $x(n)$, $n = 0, 1, 2, \ldots, N-1$, the *discrete Fourier transform* (DFT) is defined by

$$X(\omega_k) \triangleq \sum_{n=0}^{N-1} x(n) e^{-j\omega_k t_n} = \sum_{n=0}^{N-1} x(n) e^{-j2\pi kn/N}, \quad k = 0, 1, 2, \ldots N-1$$

$$\begin{aligned} t_n &\triangleq nT = n\text{th sampling instant (sec)} \\ \omega_k &\triangleq k\Omega = k\text{th frequency sample (rad/sec)} \\ T &\triangleq 1/f_s = \text{time sampling interval (sec)} \\ \Omega &\triangleq 2\pi f_s/N = \text{frequency sampling interval (rad/sec)} \end{aligned}$$

We are now in a position to have a full understanding of the transform *kernel*:

$$e^{-j\omega_k t_n} = \cos(\omega_k t_n) - j \sin(\omega_k t_n)$$

Рис. 2: "Mathematics of the discrete fourier transform (dft) with audio applications"second edition, Julius O. Smith

После преобразования Фурье имеем дело с тензором(спектрограммой), с которым мы работаем и который после обработки нужно отобразить обратно в пространство сигналов.

Постановка задачи

Тогда, нашу задачу можно сформулировать несколькими способами:

- Построить отображение из пространства пар дискретных сигналов стиль и содержание в пространство дискретных сигналов $\Psi : S \times S \to S$
- Построить отображение из пространства дискретных сигналов в пространство дискретных сигналов, т.е. без референсного аудио со стилем, но с целевым аудио $\Psi' : S \to S$

Note

Для некоторых методов используется не пространство тензоров, а некоторое пространство скрытых признаков, которые получают из дискретных сигналов с помощью различных методов feature extraction

Linear Transformation

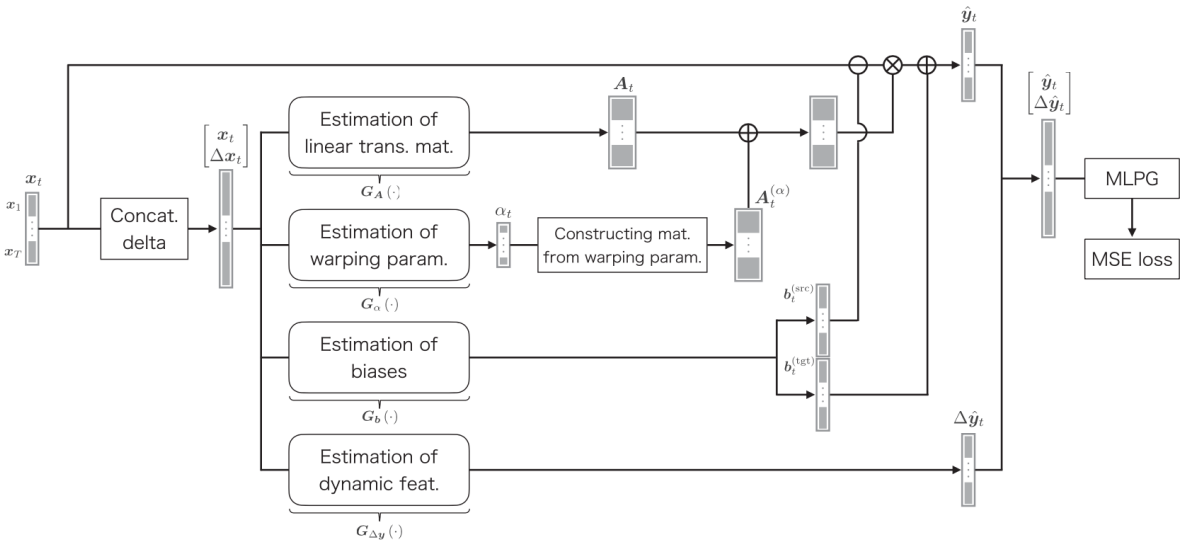


Рис. 3: Framework for time-variant linear transformations

Gaussian Mixture Model

$$P\left(\boldsymbol{z}_t \mid \boldsymbol{\lambda}^{(z)}\right) = \sum_{m=1}^{M} w_m \mathcal{N}\left(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}\right)$$

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}$$

$$\hat{\boldsymbol{y}}_{\boldsymbol{t}} = \Psi'\left(\boldsymbol{x}_t\right) = \sum_{m=1}^{M} P\left(m \mid \boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) \boldsymbol{E}_{m,t}^{(y)}, \text{ where}$$

$$\boldsymbol{E}_{m,t}^{(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} \left(\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}\right)$$

Note

where $\boldsymbol{x_t}$ and $\boldsymbol{y_t}$ - feature vectors of time index t from utterances of source and target speakers, respectively. And $\boldsymbol{z}_t = \left[\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top\right]^\top$ - joint vector

Variational Autoencoder

Note

denote **x** - observed dataset and **z** - latent variables

$$\begin{aligned}\log p_\theta(\mathbf{x}) &= KL\left[q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z} \mid \mathbf{x})\right] + \mathcal{L}(\theta, \phi; \mathbf{x}) \\ &\geq \mathcal{L}(\theta, \phi; \mathbf{x}) \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] - KL\left[q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z})\right]\end{aligned}$$

Итого, общая функция потерь:

$$\text{Loss} = KL\left[q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z})\right] - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z}, \mathbf{t})\right] + l_{\text{stop}}$$

Note

Для стохастического **z** здесь также применяется "reparamentrization trick"

Variational Autoencoder

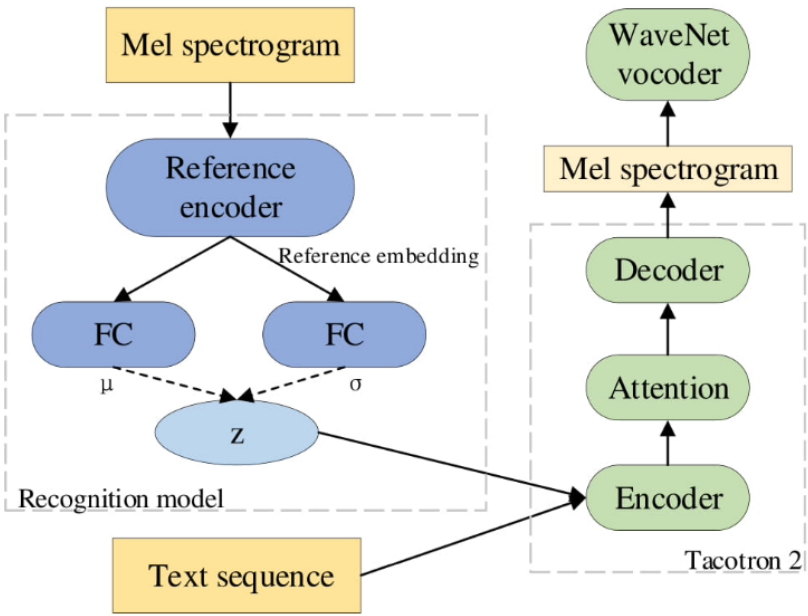


Рис. 4: Прежложенная модель

Image Like Approach

- So let $\vec{p}$ and $\vec{x}$ be the original image and the image that is generated and P^l and F^l their respective feature representation in layer l. We then define the squared-error loss between the two feature representations
  $\mathcal{L}_{\text{content}}(\vec{p}, \vec{x}, l) = \frac{1}{2} \sum_{i,j} \left(F_{ij}^l - P_{ij}^l \right)^2$
- Gram matrix $G^l \in \mathcal{R}^{N_l \times N_l}$, where G_{ij}^l is the inner product between the vectorised feature map i and j in layer l: $G_{ij}^l = \sum_k F_{ik}^l F_{jk}^l$

$$E_l = \frac{1}{4N_l^2 M_l^2} \sum_{i,j} \left(G_{ij}^l - A_{ij}^l \right)^2$$

$$\mathcal{L}_{style}(\vec{a}, \vec{x}) = \sum_{l=0}^{L} w_l E_l$$

Results:

- Работает базовый эксперимент на основе подхода к решению задачи о переносе стиля для изображений

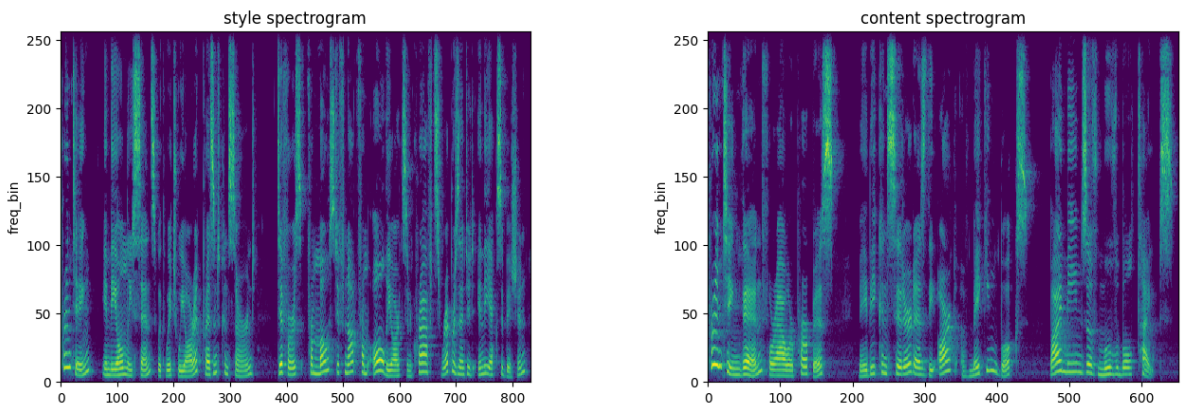


Рис. 6: Входные спектрограммы

Results:

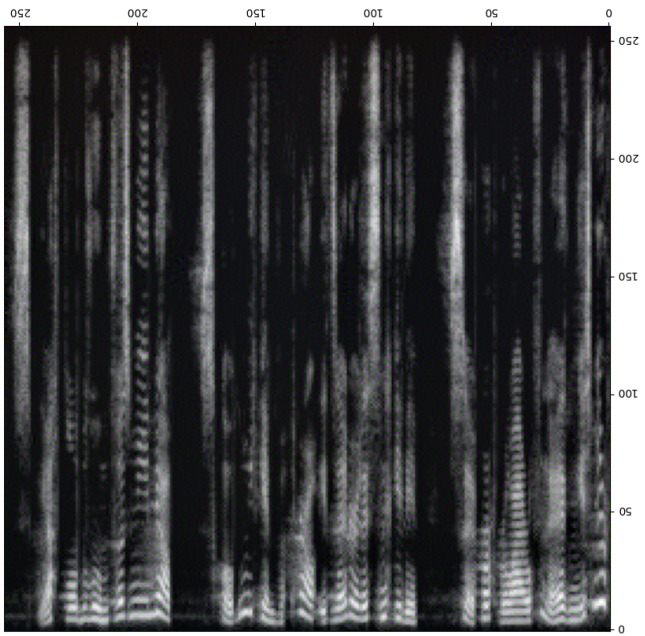


Рис. 7: Итоговая спектрограмма

- Доработка существующего базового эксперимента
- Реализация других подходов с помощью Gaussian Mixture Model и Variational Autoencoder, их улучшение/доработка
- Формализация и оформление результатов