# Informational Speech Factorization

## by Factorial Discriminative Normalization Flow

孙浩然

# Introduction

- Informational speech factorization
- Why generative model
- Why discriminative normalization flow

# VAE and NF

- VAE $$\tilde{\mathcal{L}}(\theta, \phi) = \sum_i \tilde{\mathcal{L}}(\mathbf{x}_i) \leq \sum_i \log p(\mathbf{x}_i) = \mathcal{L}(\theta, \phi).$$
  - ELBO
  - Information loss
- NF $$\log p(\boldsymbol{x}) = \log p(\boldsymbol{z}) + \log \left| \det\left(\frac{\mathrm{d} f_\theta{}^{-1}(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}}\right) \right|$$


**Figure 3:** Distribution transform with normalization flow.

# DNF

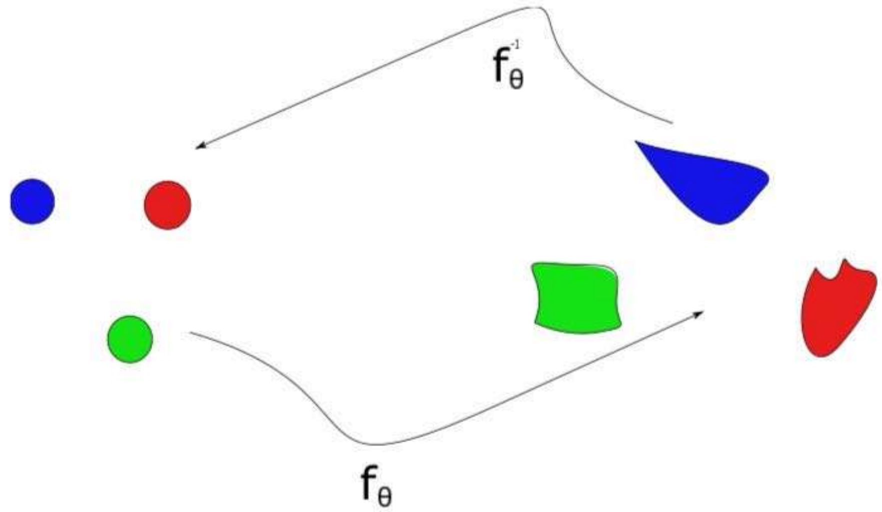



**Figure 5:** The DNF architecture, where each class has its unique prior distribution.

- Supervision $\log p(\boldsymbol{x}) = \log \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{y(\boldsymbol{x})}, \boldsymbol{I}) + J(\boldsymbol{x}),$

# Factorial DNF

- Take more than one information factors into consideration. (two factors as example)
- split the latent code into two partial codes (different dimensions). $\mathbf{z} = [\mathbf{z}^A \ \mathbf{z}^B]$
- The two codes are independent because of the prior's diagonal Gaussian.

$$p(\mathbf{z}) = p(\mathbf{z}^A)p(\mathbf{z}^B),$$

# Factorial DNF

$$p(\boldsymbol{z}^A) = \mathcal{N}(\boldsymbol{z}^A; \boldsymbol{\mu}_{y_A(\boldsymbol{z})}, \boldsymbol{I})$$
$$p(\boldsymbol{z}^B) = \mathcal{N}(\boldsymbol{z}^B; \boldsymbol{\mu}_{y_B(\boldsymbol{z})}, \boldsymbol{I})$$

$$p(\boldsymbol{z}) = p(\boldsymbol{z}^A)p(\boldsymbol{z}^B),$$

$$\log p(\boldsymbol{x}) = \log p(\boldsymbol{z}) + \boldsymbol{J}(\boldsymbol{x}).$$

$$\log p(\boldsymbol{x}) = \log p(\boldsymbol{z}^A) + \log p(\boldsymbol{z}^B) + \boldsymbol{J}(\boldsymbol{x}).$$

# Experimental settings

## Data

- TIMIT, 462 speakers in training set, the original 58 phones are mapped to 39 phones(using 38 phones without 'sil', which is 'silence') by Kaldi's phone mapping tool.
- Phone segments with a 200ms duration, guaranteeing that main phones are in the middle of these segments.
- 4000 dimensions, $20 \times 200$ time-frequency spectrograms, where 20 is the number of frames in the segment, and 200 is the number of frequency bins.

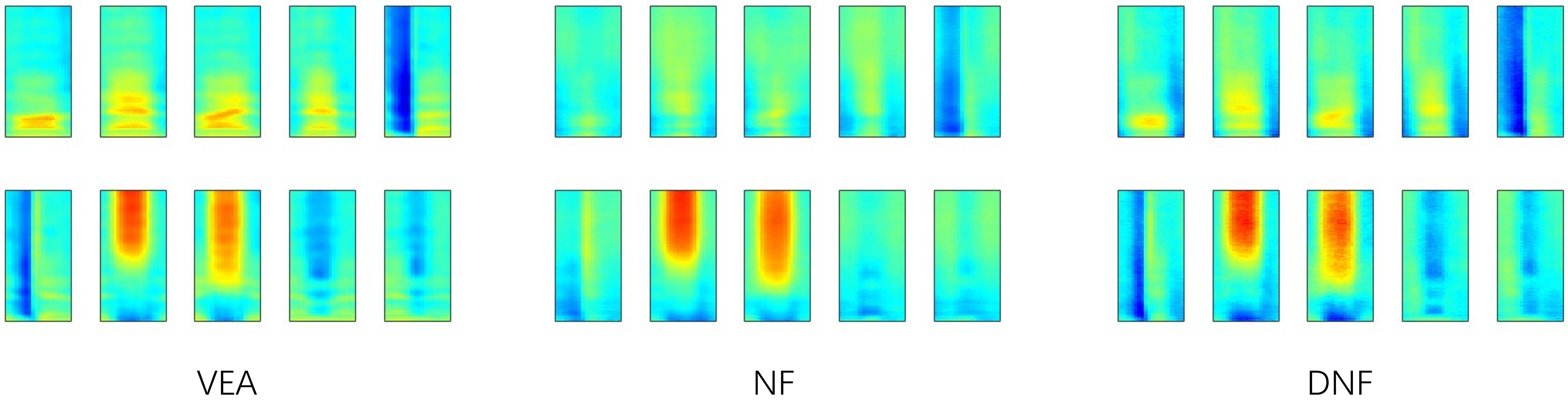
# Experimental settings

## Model training

- VAE involves three convolutional layers followed by a fully connected layers, and the dimension of the latent space is 128
- NF, DNF and f-DNF models follow the RealNVP structure, there are 6 blocks in every model, and each block has a coupling layer and a batch norm layer.
- Class means of DNF and f-DNF are initialized by 0-I normal distribution, and within variances are set as 1.0.
- Each partial codes of f-DNF has 2000 dimensions.

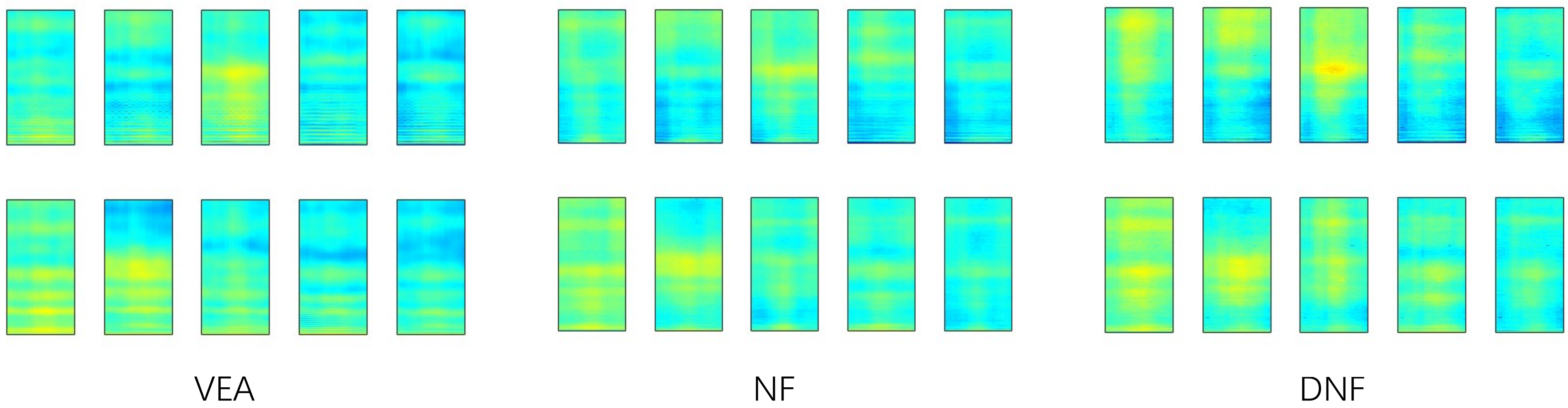
# Speech encoding

## Phone class means as representations



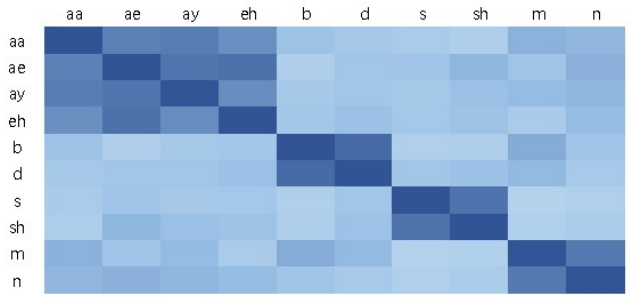
# Speech encoding

## Speaker class means as representations

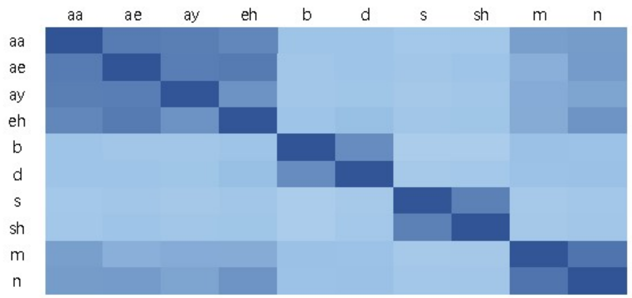


# Speech encoding

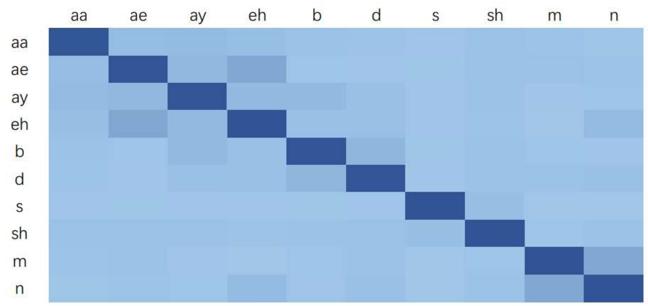
## Phone class mean distance




VEA



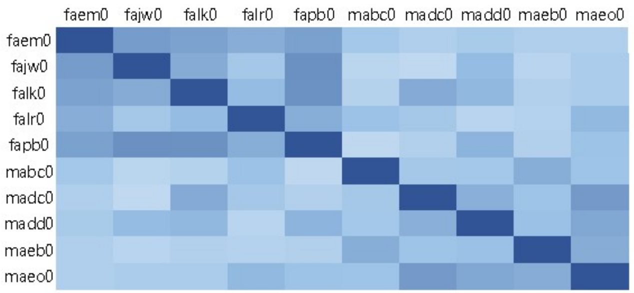

NF

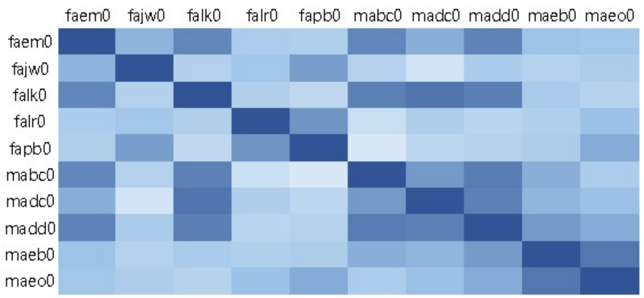


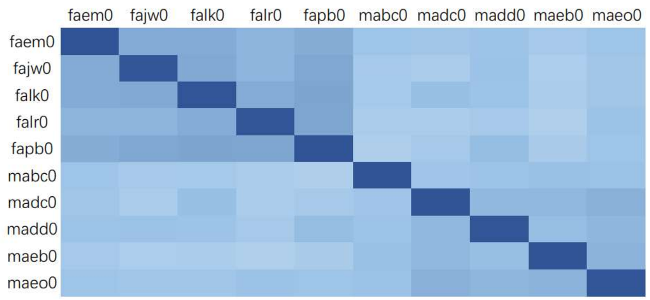

DNF

# Speech encoding

## Speaker class mean distance



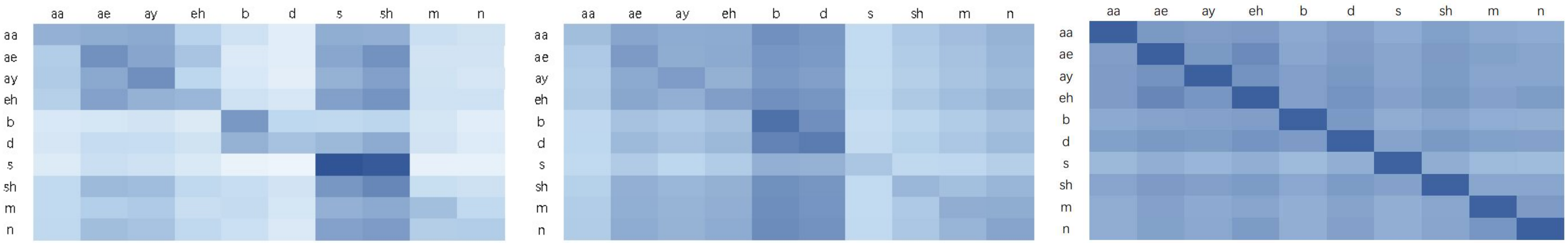




VEA NF DNF

# Speech encoding

## Phone class likelihood

## (Between-class likelihood)



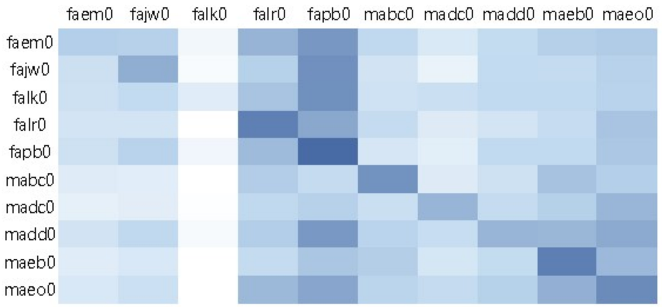

VEA NF DNF

Each column represents a class of data, and each row represents a class distribution

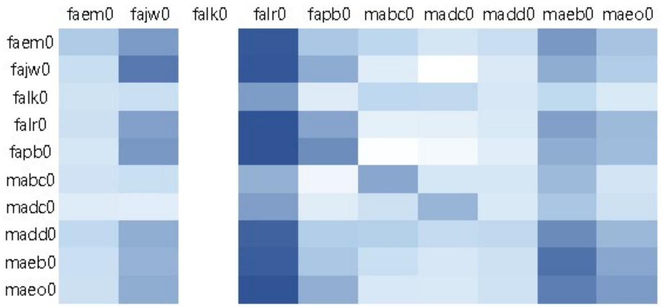
# Speech encoding

## Speaker class likelihood

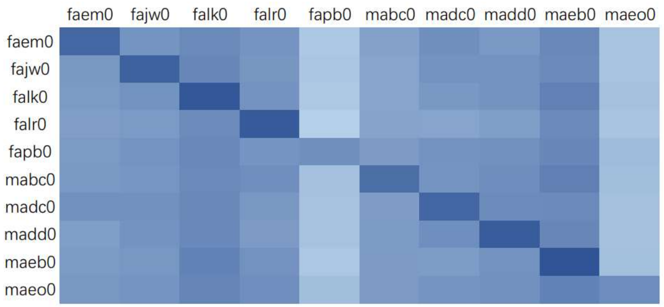
## (Between-class likelihood)




VEA




NF




DNF

# Speech encoding

## Recognition

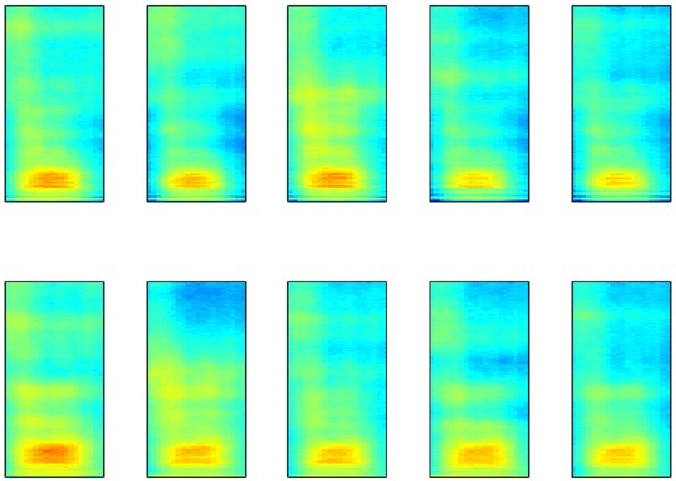
- Top-1 log(p(z))
- Short-term phone and long-term speaker information bias
- Speaker verification based on average likelihood of *K* sequential segments

**Table 1**
Accuracies of phone recognition and speaker verification on VAE, NF, DNF.

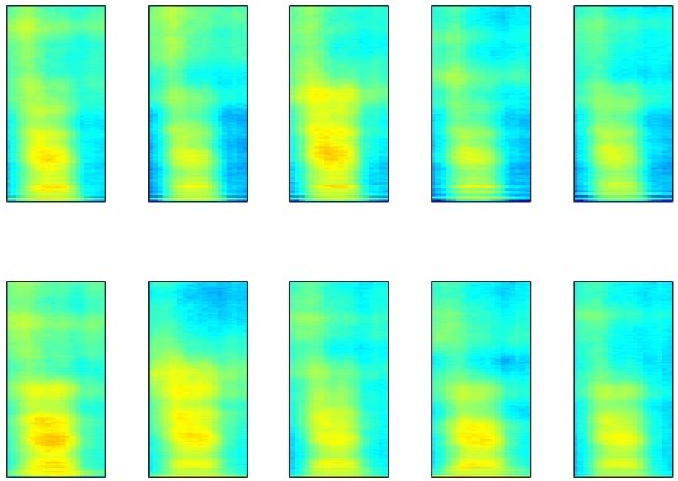
| | VAE | NF | DNF |
|---|---|---|---|
| Phone | 0.5289 | 0.5192 | **0.9986** |
| Speaker (K=1) | 0.3567 | 0.2985 | **0.9318** |
| Speaker (K=3) | 0.7242 | 0.6051 | **0.9963** |
| Speaker (K=5) | 0.8766 | 0.7555 | **0.9991** |
| Speaker (K=10) | 0.9822 | 0.9197 | **1.0** |

# Speech factorization

## Class means as representations



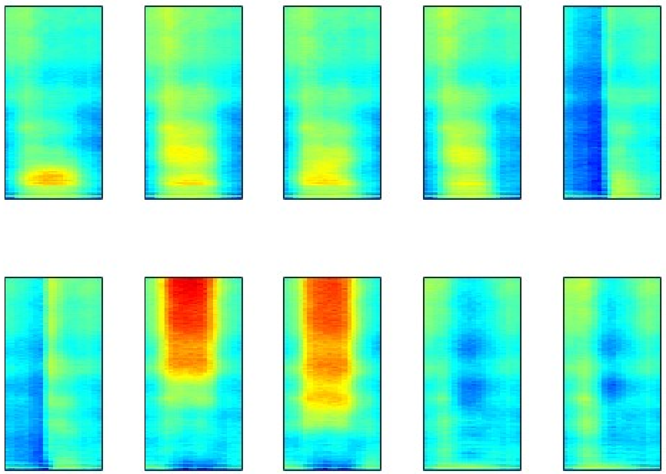
Different speakers with /aa/



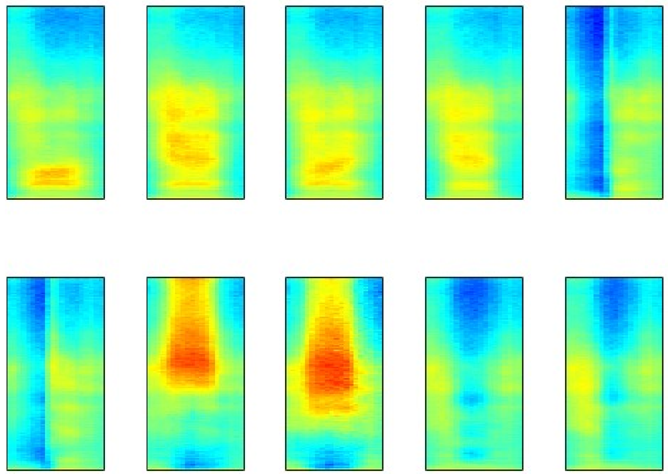
Different speakers with /iy/

# Speech factorization

## Class means as representations



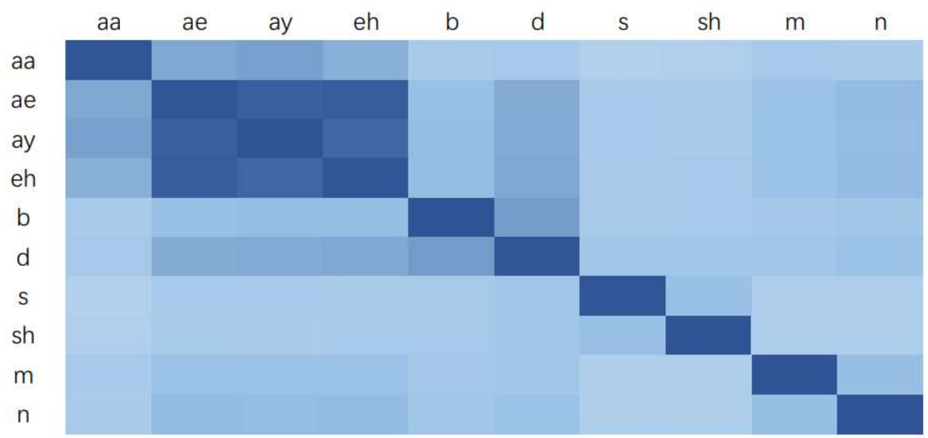
Different phones with a female



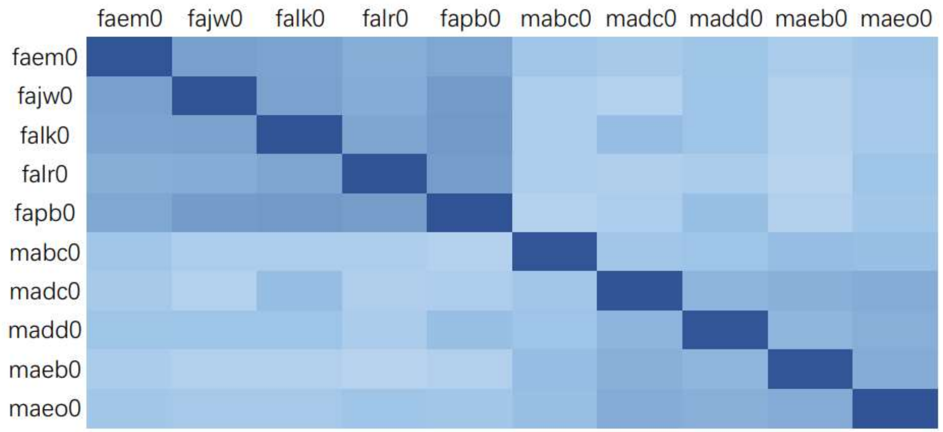
Different phones with a male

# Speech factorization

## Class mean distance



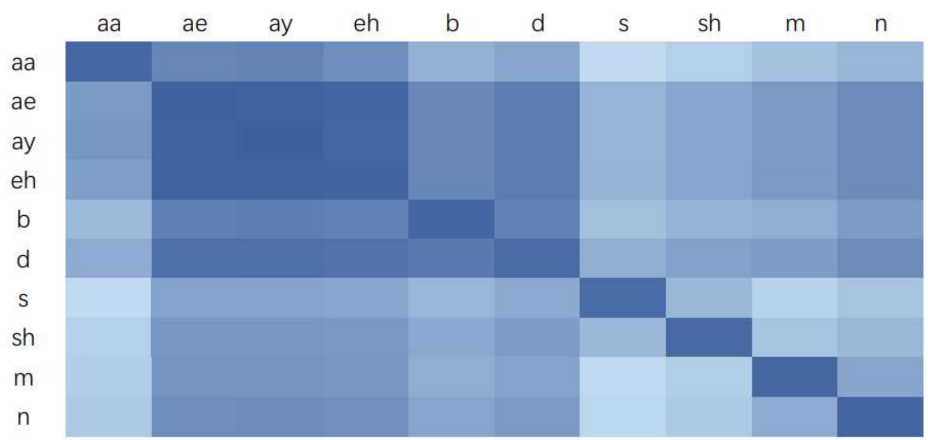

Phones



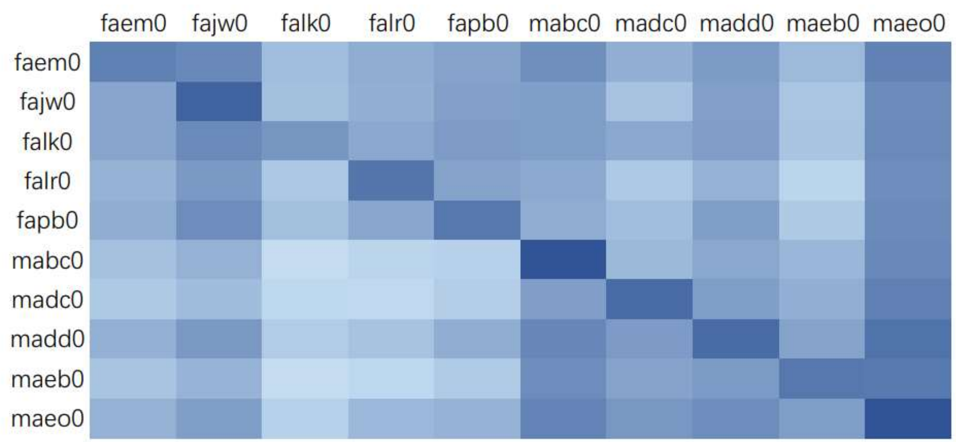

Speakers

# Speech factorization

## Class likelihood




Phones

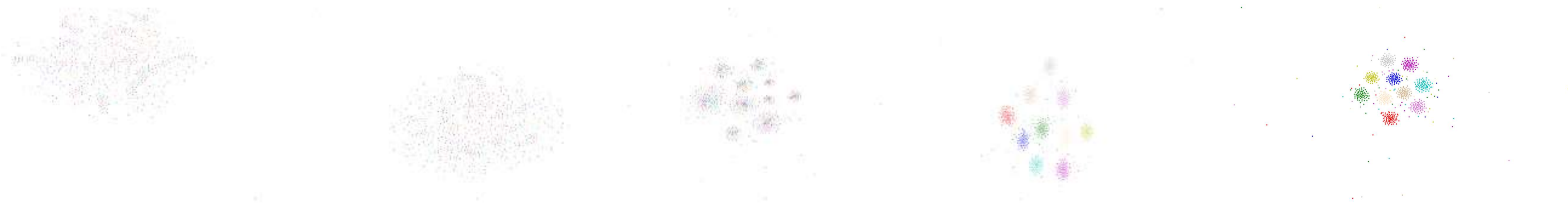



Speakers

# Speech factorization

## Class discrimination (tSNE)

(a) VAE (b) NF (c) Phone DNF (d) Speaker DNF (e) Factorial DNF 40 speakers



(f) VAE (g) NF (h) Phone DNF (i) Speaker DNF (j) Factorial DNF 40 speakers

# Speech factorization

## Phone and speaker conversion

- Posteriors from mlp
- 38 phones & 40 speakers; 4/5 segments for mlp training and 1/5 segments for conversion and mlp test

$$x' = f(f^{-1}(x) + \mu_{A,c2} - \mu_{A,c1}).$$

Table 1. Posteriors on the target class before and after phone/speaker conversion.

| | Phone Manipulation | | | |
|---|---|---|---|---|
| Model | $p(q2\|x)$ | $p(q2\|x')$ | $p(s\|x)$ | $p(s\|x')$ |
| VAE | 0.0345 | 0.3724 | 0.6117 | 0.4915 |
| NF | 0.0726 | 0.2357 | 0.6117 | 0.4086 |
| DNF | 0.0277 | 0.4161 | 0.6117 | 0.5289 |
| Factorial DNF | 0.0375 | 0.3510 | 0.6117 | 0.5627 |
| | Speaker Manipulation | | | |
| Model | $p(s2\|x)$ | $p(s2\|x')$ | $p(q\|x)$ | $p(q\|x')$ |
| VAE | 0.0330 | 0.3903 | 0.5203 | 0.5134 |
| NF | 0.0124 | 0.5805 | 0.5203 | 0.3871 |
| DNF | 0.0108 | 0.6060 | 0.5203 | 0.3809 |
| Factorial DNF | 0.0295 | 0.4804 | 0.5203 | 0.5051 |

# Speech factorization

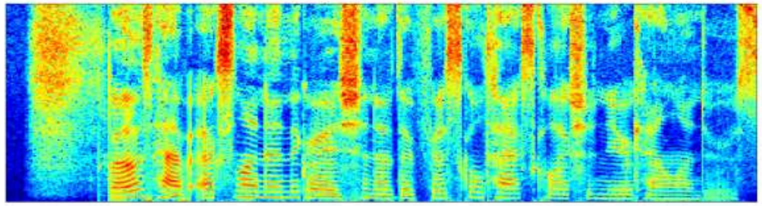
## Speaker conversion example

- Converting a Speech *Xa* spoken by speaker *a* to *Xb* which sounds like coming from speaker *b*.

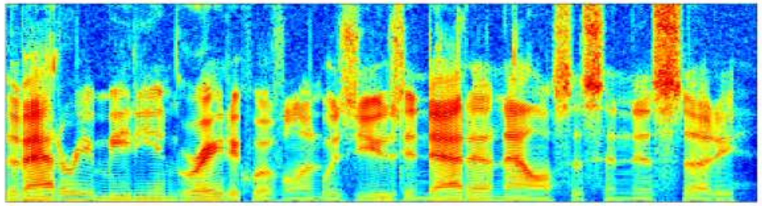
$$\mu_{S,a} = \frac{1}{T_a} \sum_t z_{a,t}^S.$$

$$\Delta_S = \mu_{S,b} - \mu_{S,a}$$

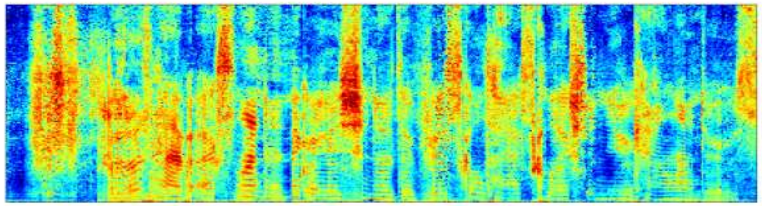
$$z_{c,t} = [z_{a,t}^Q \; z_{a,t}^S + \Delta_S].$$



(a) Original speech $X_a$



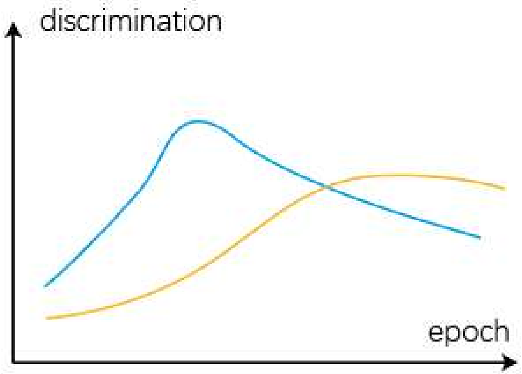
(b) Sample speech for target speaker $X_b$



(c) Converted speech $X_c$

# Discussions

- The phone factor code and speaker factor code are not fully independent.
- Dimension splitting makes differences.
- The two codes couldn't get the best performance at the same time.




- Discrimination of each code space is not so good as we hoped.

# Summary

- Informational speech factorization by factorial DNF is feasible, but there remain some problems to be solved.