Learning and controlling the source-filter representation of speech with a VAE

Simon Leglaive

CentraleSupélec, IETR (UMR CNRS 6164), France

Workshop "Methods and Tools for Audio-Visual Processing and Human Robot Interaction" - Inria centre at the University Grenoble Alpes

August 23, 2022

Joint work with

Samir Sadok[1]

Laurent Girin[2]

Xavier Alameda-Pineda[3]

Renaud Séguier[1]

[1] CentraleSupélec, IETR UMR CNRS 6164, France

[2] Univ. Grenoble Alpes, CNRS, Grenoble-INP, GIPSA-lab, France

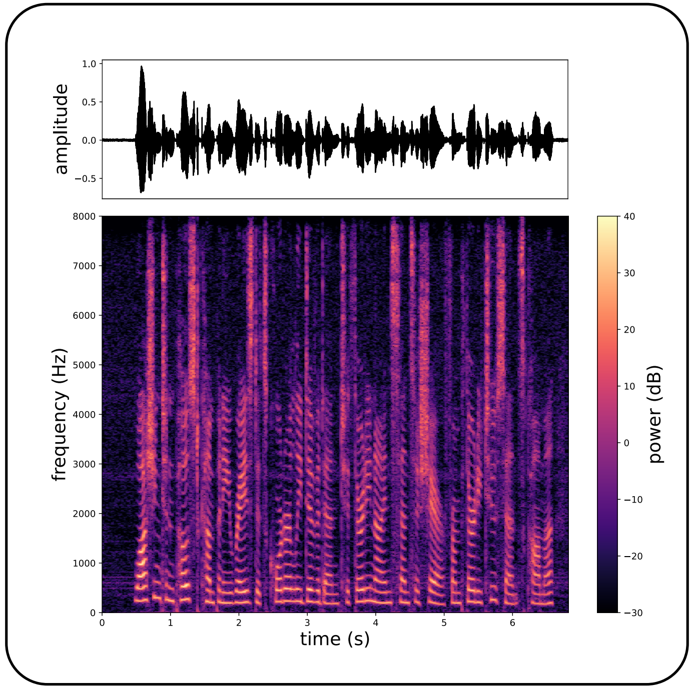
[3] Inria, Univ. Grenoble Alpes, CNRS, LJK, France

S. Sadok, S. Leglaive, L. Girin, X. Alameda-Pineda, R. Séguier, Learning and controlling the source-filter representation of speech with a variational autoencoder, arXiv preprint arXiv:2204.07075, 2022.

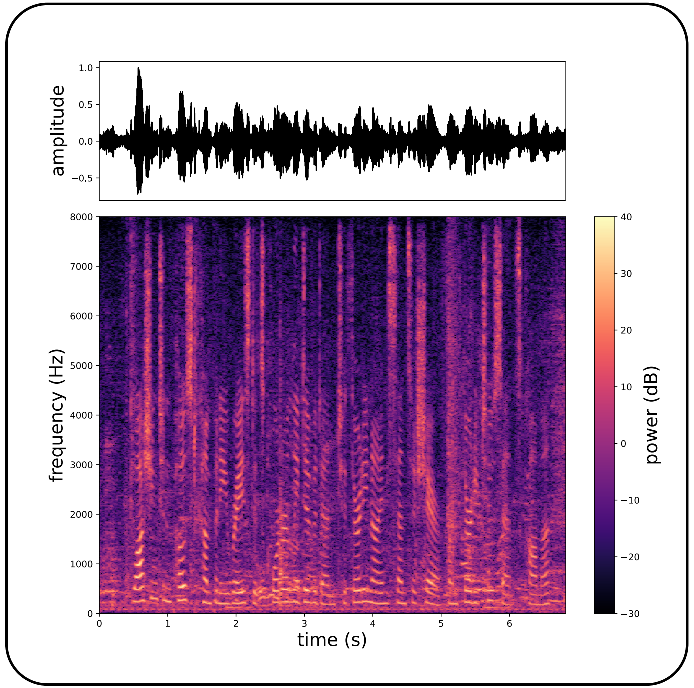
Motivation

Inverse problems in audio signal processing

Latent signal of interest $\mathbf{s}$



Noisy and/or incomplete observations $\mathbf{x}$



$$\underset{\text{posterior}}{\boxed{p(\mathbf{s}|\mathbf{x})}} \propto \underset{\text{prior}}{\boxed{p(\mathbf{s})}} \; \underset{\text{likelihood}}{\boxed{p(\mathbf{x}|\mathbf{s})}}$$

Source separation, speech enhancement, inpainting, phase retrieval, bandwidth extension, ...

We need a probabilistic/generative model of the latent signal of interest

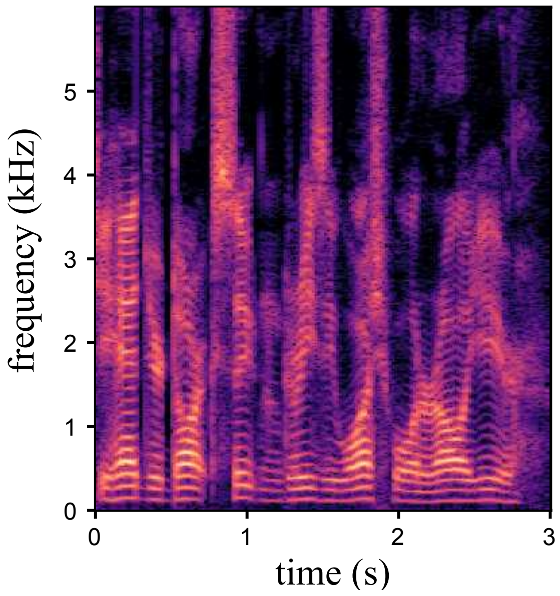
Non-stationary Gaussian model (Ephraim and Malah, 1984)

- Let $\mathbf{S} \in \mathbb{C}^{F \times T}$ denote an audio/speech signal in the short-time Fourier transform (STFT) domain, with

$$p(\mathbf{S}) = \prod_{t=1}^{T} p(\mathbf{s}_t) = \prod_{t=1}^{T} \mathcal{N}_c\left(\mathbf{s}_t; \mathbf{0}, \mathrm{diag}\{\mathbf{v}_{s,t}\}\right).$$

- $\mathbf{s}_t \in \mathbb{C}^F$ denotes the complex-valued spectrum of the signal at time frame t.

- $\mathbf{v}_{s,t} \in \mathbb{R}_+^F$ represents the expected power spectrum of the signal at time frame t.



The variance is usually constrained to encode specific spectro-temporal characteristics.

This Gaussian model implies that the entries of $|\mathbf{s}_t|^{\odot 2}$ follow an exponential or Gamma distribution parametrized by $\mathbf{v}_{s,t}$.

Y. Ephraim and D. Malah, Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, IEEE TASSP 1984.

The variance modeling framework (Vincent et al., 2010)

From "explicit" signal models to data-driven approaches:

- Structured-sparsity-inducing priors for modeling tonal and transient sounds

  (Févotte et al., 2007)

- Non-negative matrix factorization (NMF) for modeling spectrograms as non-negative linear combinations of learned spectral templates

  (Benaroya et al., 2003; Févotte et al., 2009; Ozerov et al., 2012)

- (Dynamical) variational autoencoder (VAE) for learning (spectro-temporal) spectral structures

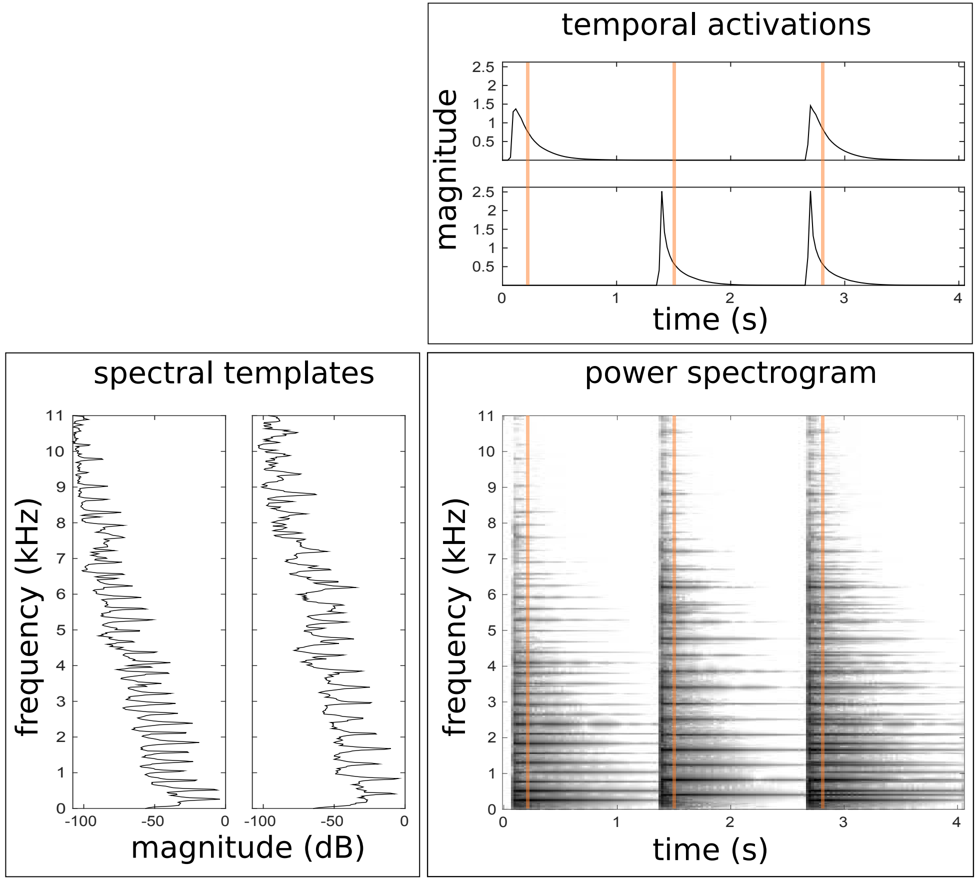
  (Bando et al., 2018; Leglaive et al., 2018; 2020; Girin et al., 2021)

E. Vincent et al., Probabilistic modeling paradigms for audio source separation, In: Machine Audition: Principles, Algorithms and Systems, 2010.
C. Févotte et al., Sparse linear regression with structured priors and application to denoising of musical audio, IEEE TASLP, 2007.
L. Benaroya et al., Non negative sparse representation for Wiener based source separation with a single sensor, IEEE ICASSP 2003.
C. Févotte et al., Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis, Neural Computation, 2009.
A. Ozerov et al., A general flexible framework for the handling of prior information in audio source separation, IEEE/ACM TASLP, 2012.
Y. Bando et al., Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization, IEEE ICASSP 2018.
S. Leglaive et al., A variance modeling framework based on variational autoencoders for speech enhancement, IEEE MLSP 2018.
S. Leglaive et al., A recurrent variational autoencoder for speech enhancement, IEEE ICASSP 2020.
L. Girin et al., Dynamical variational autoencoders: A comprehensive review, Foundations and Trends in Machine Learning, 2021.

NMF-based variance modeling (Févotte et al., 2009)

$$p(\mathbf{s}_t) = \mathcal{N}_c\Big(\mathbf{s}_t; \mathbf{0}, \mathrm{diag}\{\mathbf{v}_{s,t} = \mathbf{W}\mathbf{h}_t\}\Big),$$

- $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ is a dictionary matrix of spectral templates;
- $\mathbf{h}_t \in \mathbb{R}_+^K$ is the low-dimensional activation vector at time frame t;
- K is the rank of the factorization.



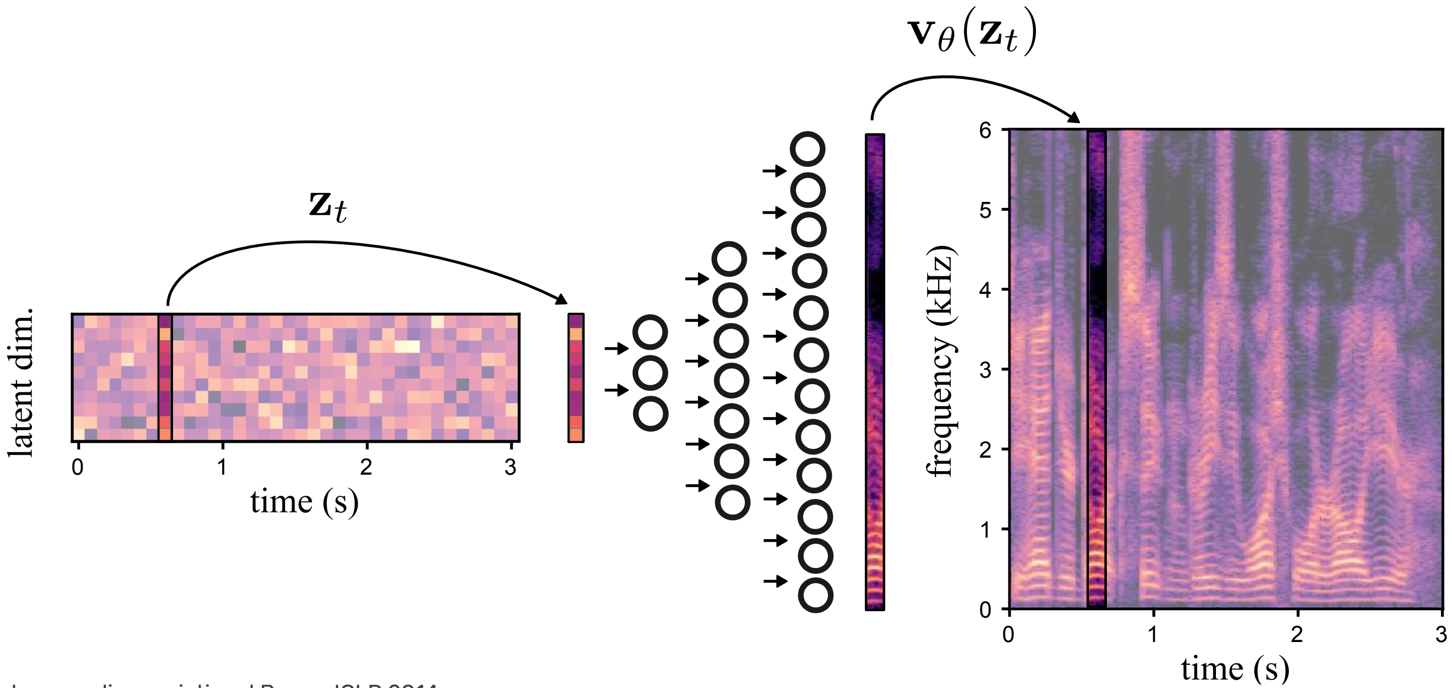
C. Févotte et al., Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis, Neural Computation, 2009.

VAE-based variance modeling (Kingma and Welling, 2014; Bando et al., 2018)

$$p(\mathbf{s}_t \mid \mathbf{z}_t) = \mathcal{N}_c\Big(\mathbf{s}_t; \mathbf{0}, \text{diag}\left\{\mathbf{v}_{s,t} = \mathbf{v}_\theta(\mathbf{z}_t)\right\}\Big),$$

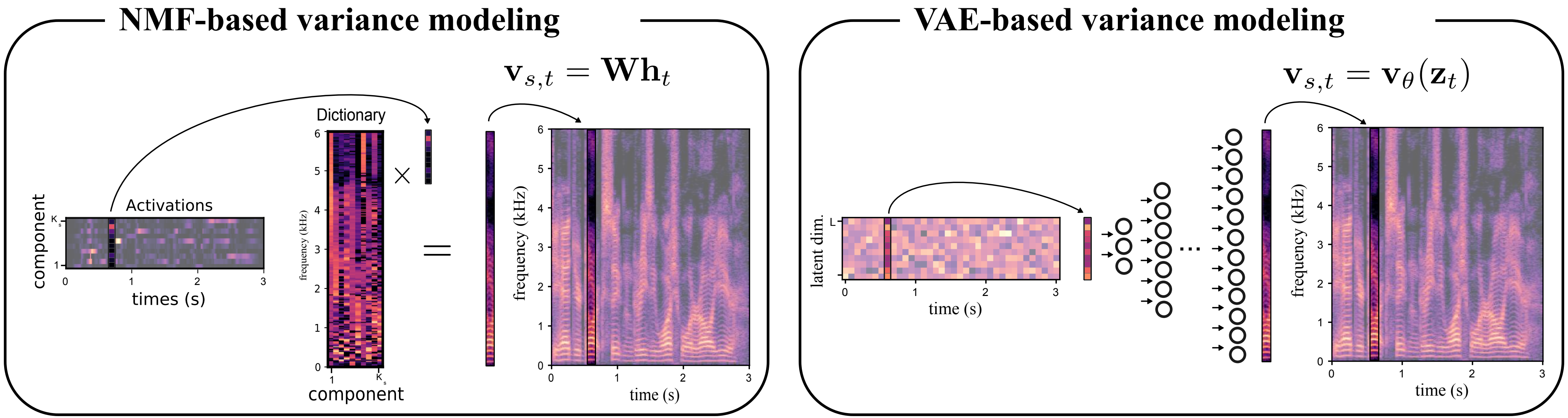
- $\mathbf{z}_t \in \mathbb{R}^K$ is a low-dimensional latent vector with $p(\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_t; \mathbf{0}, \mathbf{I})$.

- $\mathbf{v}_\theta : \mathbb{R}^K \mapsto \mathbb{R}_+^F$ is a neural network (decoder) of parameters θ.
- $p(\mathbf{s}_t) = \int p(\mathbf{s}_t \mid \mathbf{z}_t) p(\mathbf{z}_t) d\mathbf{z}_t$.



D.P. Kingma and M. Welling, Auto-encoding variational Bayes, ICLR 2014.
Y. Bando et al., Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization, IEEE ICASSP 2018.

NMF vs. VAE for variance modeling



In speech enhancement, the VAE model outperforms the NMF model (Leglaive et al., 2018).

However, contrary to NMF, **we cannot directly relate the learned representation to interpretable properties of the signal.**

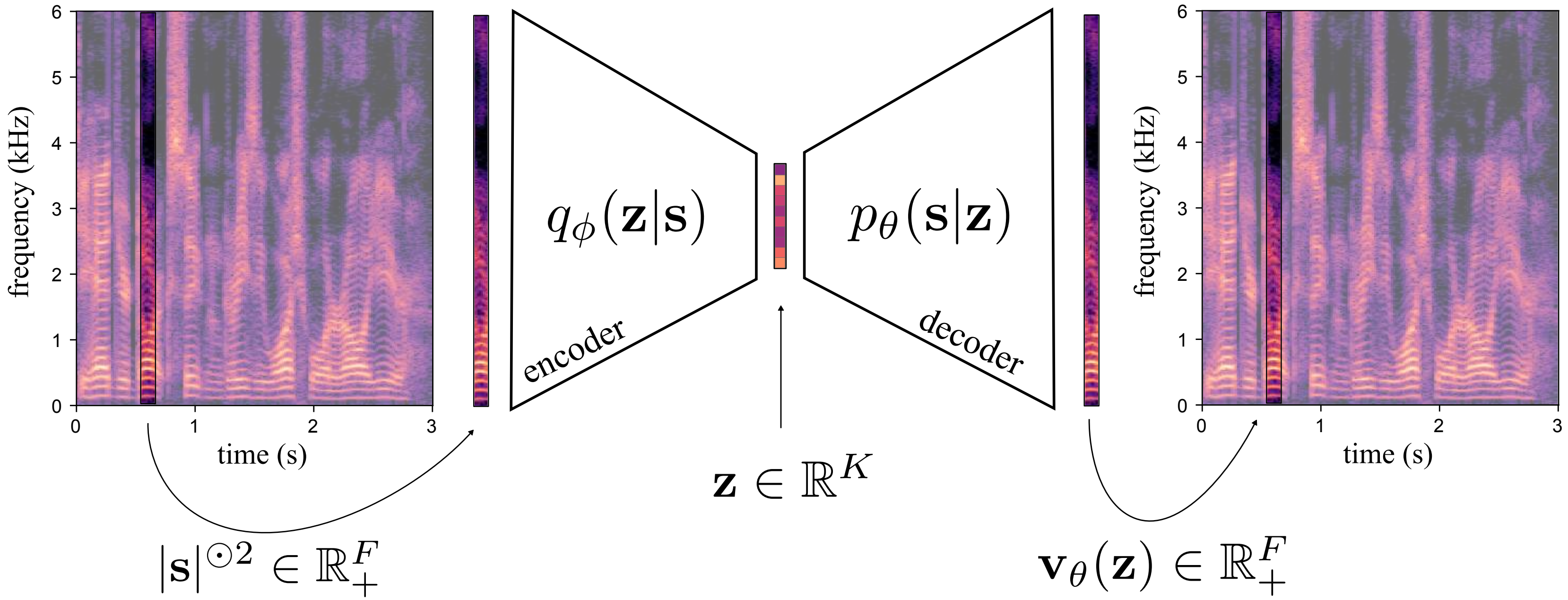
This is the problem we are going to tackle.

S. Leglaive et al., A variance modeling framework based on variational autoencoders for speech enhancement, IEEE MLSP 2018.

Analyzing the VAE latent space

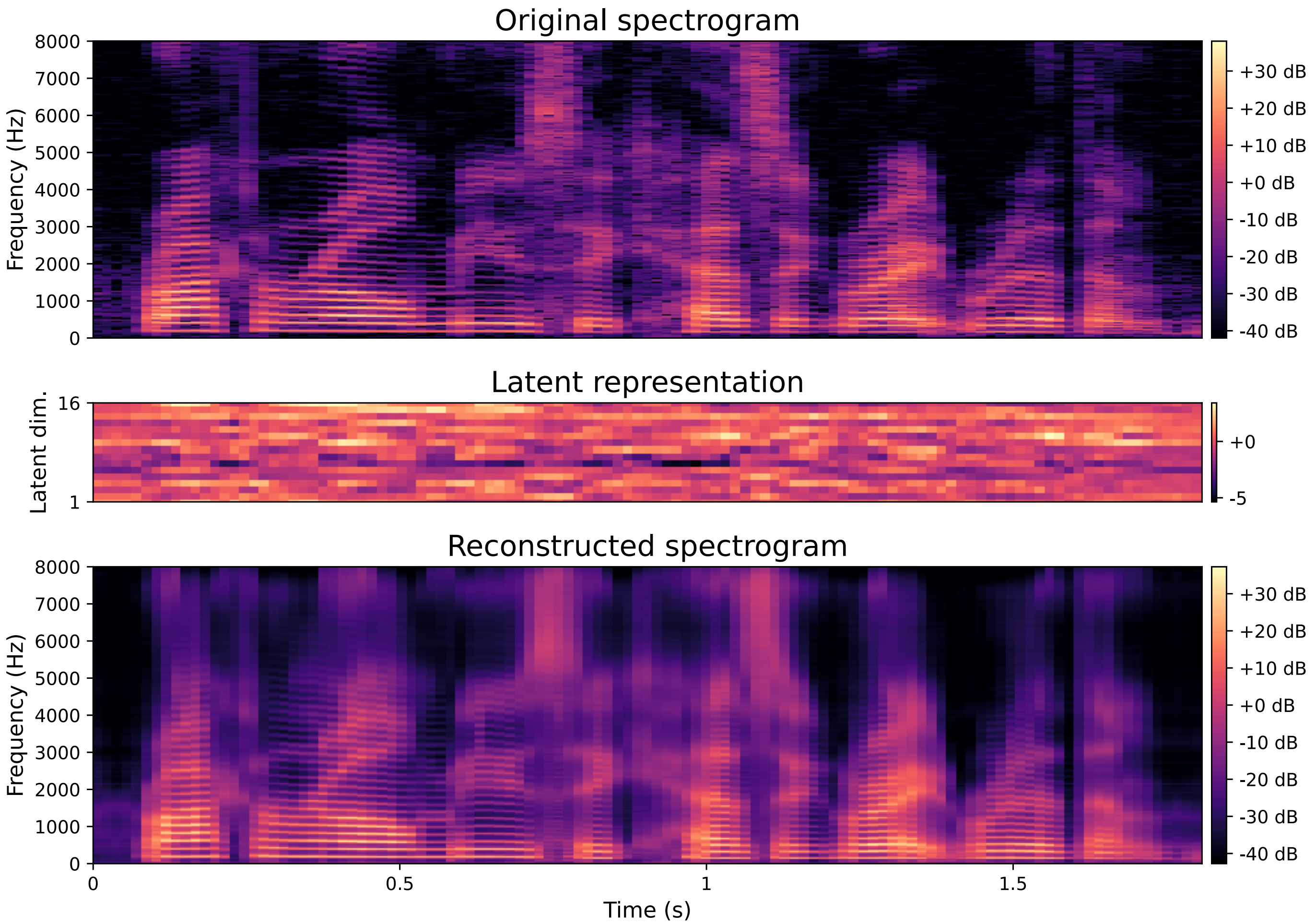
Complete VAE model

Prior	Generative model	Inference model
$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$	$p_\theta(\mathbf{s}\|\mathbf{z}) = \mathcal{N}_c\left(\mathbf{s}; \mathbf{0}, \operatorname{diag}\{\mathbf{v}_\theta(\mathbf{z})\}\right)$	$q_\phi(\mathbf{z}\|\mathbf{s}) = \mathcal{N}\left(\mathbf{z}; \boldsymbol{\mu}_\phi(\mathbf{s}), \operatorname{diag}\{\mathbf{v}_\phi(\mathbf{s})\}\right)$



We trained a vanilla VAE on about 25 hours of unlabeled speech signals at 16 kHz.

Analysis-resynthesis by encoding-decoding



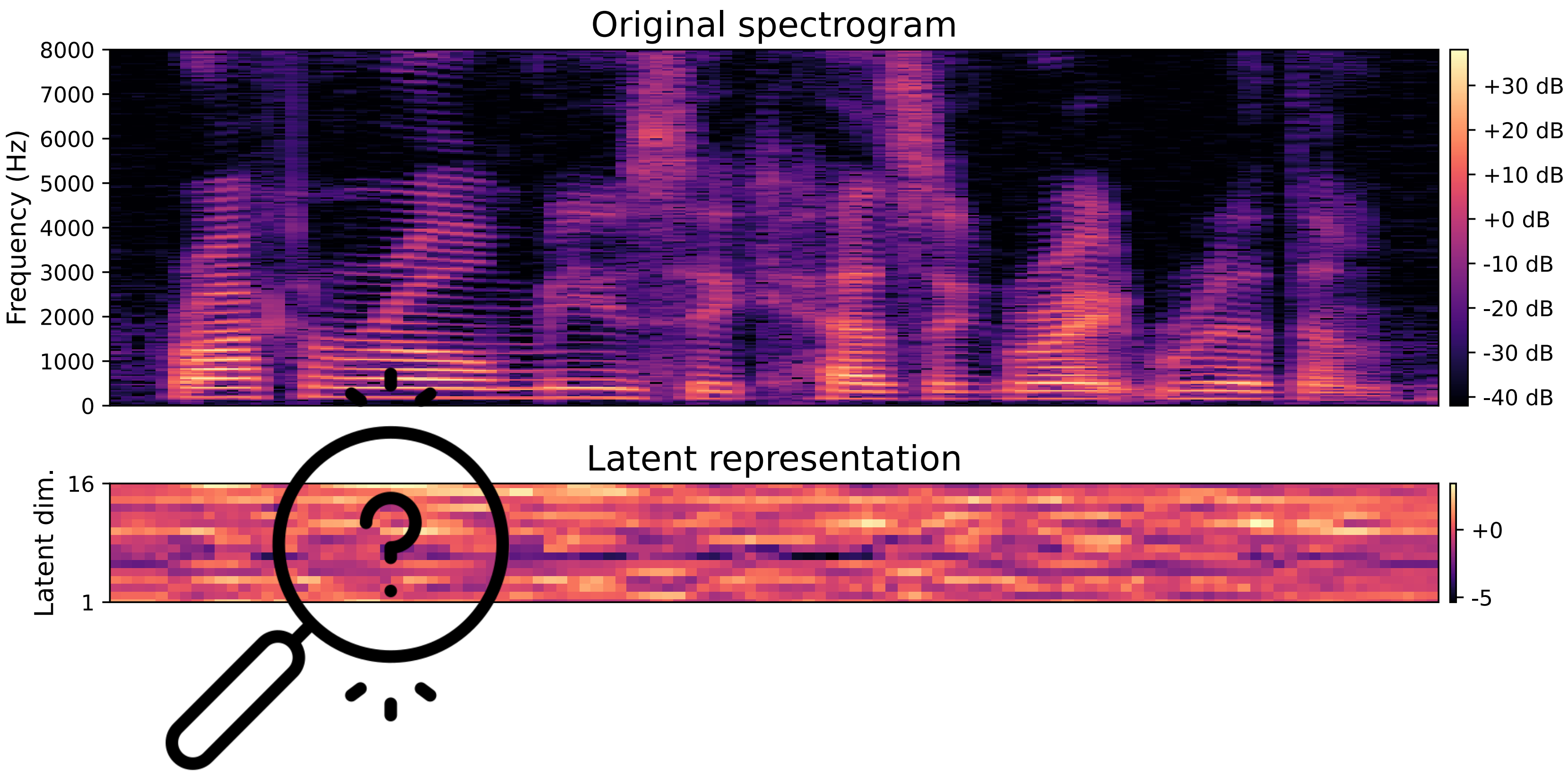
Original signal

Reconstruction with

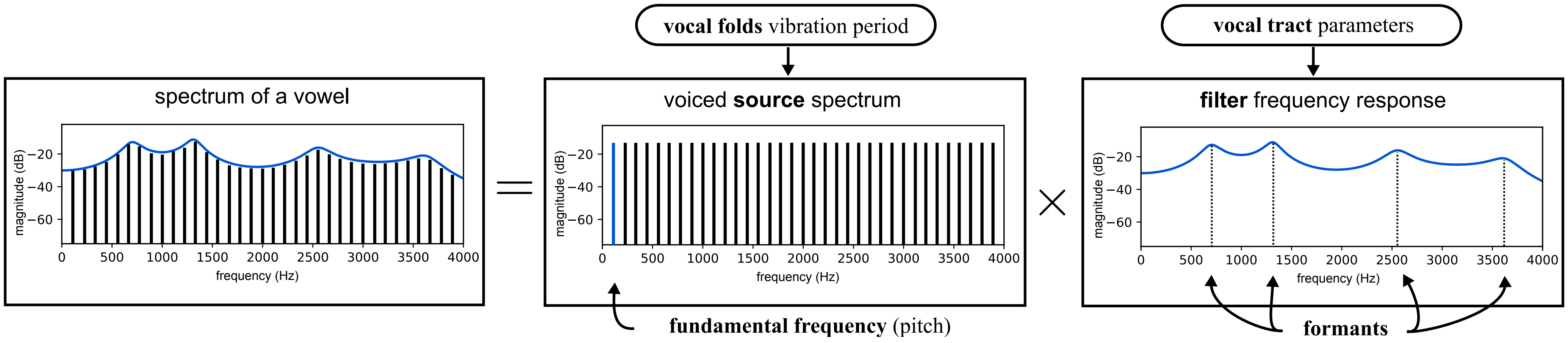
1. oracle phase
2. Griffin-Lim (Griffin and Lim, 1984)
3. WaveGlow (Prenger et al., 2019)

D. Griffin and J.S. Lim, Signal estimation from modified short-time Fourier transform, IEEE TASSP, 1984.
R. Prenger et al., Waveglow: A flow-based generative network for speech synthesis, IEEE ICASSP, 2019.

Understanding the structure of the latent space using natural speech signals is difficult, let's "open the black box" with **simpler speech signals**.



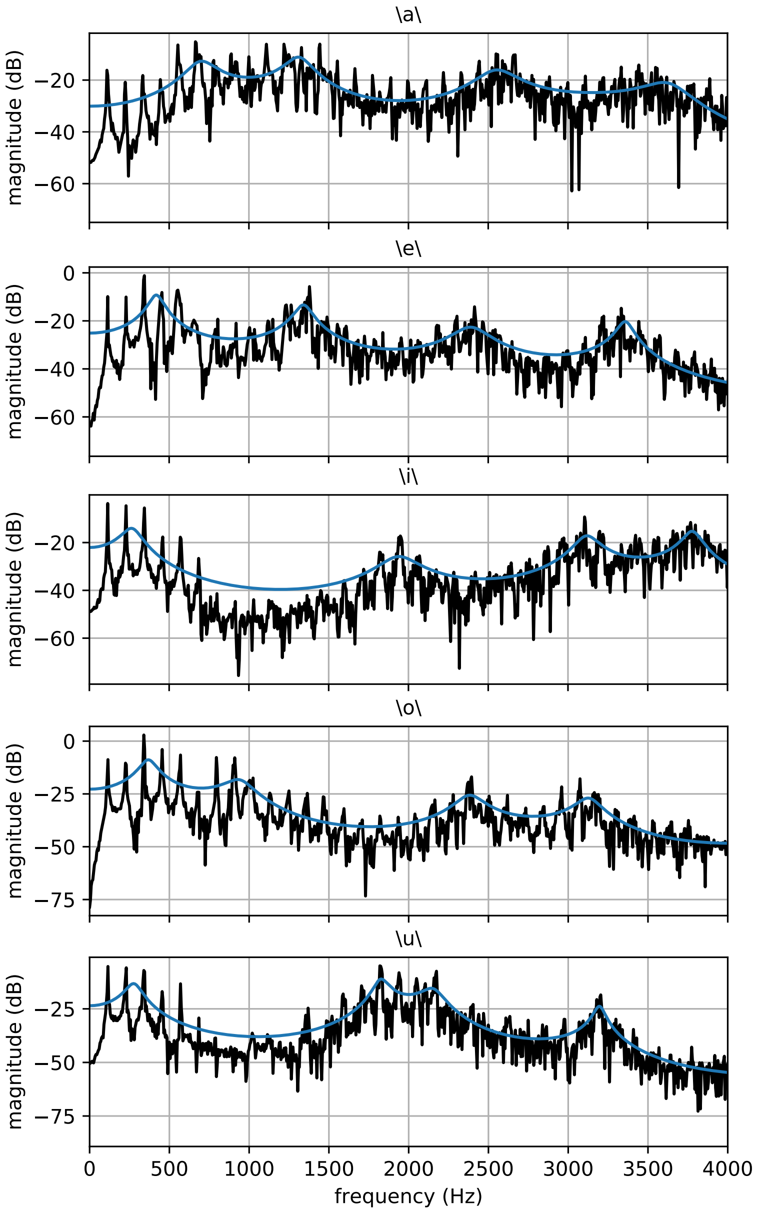
Source-filter model of (voiced) speech production



The source-filter model proposed by (Fant, 1970) considers that the production of speech results from the interaction of a **source signal** with a **linear filter**.

- In voiced speech, the source originates from the vibration of the vocal folds. This vibration is characterized by the fundamental frequency, loosely referred to as the pitch.
- The source signal is modified by the vocal tract, which is assumed to act as a linear filter. The cavities of the vocal tract give rise to resonances, which are called the formants.

G. Fant, Acoustic theory of speech production (No. 2), Walter de Gruyter, 1970.



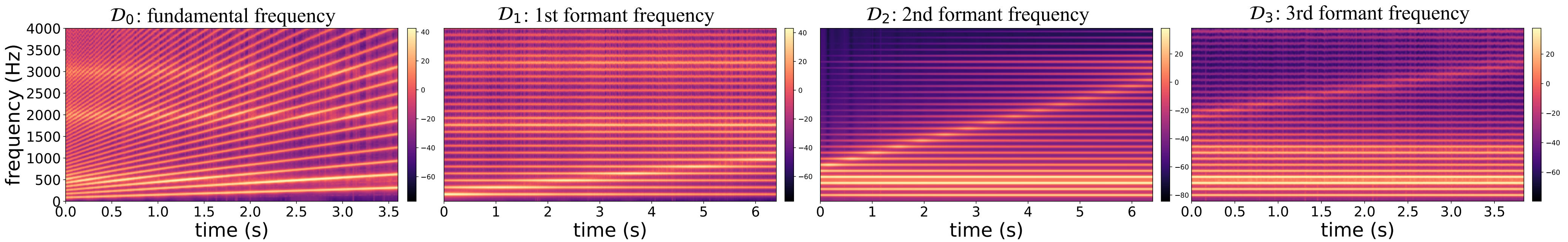
- By moving the speech articulators (tongue, lips, jaw), humans modify the shape of their vocal tract, which results in a change of the formant frequencies.

- The source-filter model tells us that we can control the source (f_0) independently of the filter (the formants) (Fant, 1970).

- The first formant frequencies $\{f_i\}_{i\geq 1}$ can also be controled independently of each other (MacDonald et al., 2011).

E. N. MacDonald, Probing the independence of formant control using altered auditory feedback, JASA, 2011.

Automatically-labeled artificial speech trajectories



- We generate datasets $\{\mathcal{D}_i\}_{i=0}^{3}$ containing a few seconds of vowel-like speech power spectra where only one factor f_i varies, all other factors $\{f_j\}_{j\neq i}$, being arbitrarily fixed.
- We used Soundgen (Anikin, 2019), an artificial speech synthesizer based on the source-filter model.
- All examples in $\mathcal{D}_i$ are automatically-labeled with f_i (this is an input of soundgen).

> We are going to investigate the VAE latent representation associated with these trajectories.

A. Anikin, Soundgen: An open-source tool for synthesizing nonverbal vocalizations, Behavior Research Methods, 2019.

Aggregated posterior

- Let $\hat{p}^{(i)}(\mathbf{s}) = \frac{1}{\#\mathcal{D}_i} \sum_{\mathbf{s}_n \in \mathcal{D}_i} \delta(\mathbf{s} - \mathbf{s}_n)$ denote the empirical distribution associated with $\mathcal{D}_i$.

- The aggregated posterior is a marginal distribution over $\mathbf{z}$ defined by "aggregating, or averaging, the VAE approximate posterior $q_\phi(\mathbf{z}|\mathbf{s})$ over $\hat{p}^{(i)}(\mathbf{s})$":

$$\hat{q}_\phi^{(i)}(\mathbf{z}) = \mathbb{E}_{p^{(i)}(\mathbf{s})}[q_\phi(\mathbf{z}|\mathbf{s})] = \int q_\phi(\mathbf{z}|\mathbf{s})\hat{p}^{(i)}(\mathbf{s})d\mathbf{s} = \frac{1}{\#\mathcal{D}_i} \sum_{\mathbf{s}_n \in \mathcal{D}_i} q_\phi(\mathbf{z}|\mathbf{s}_n).$$

- For instance, we have

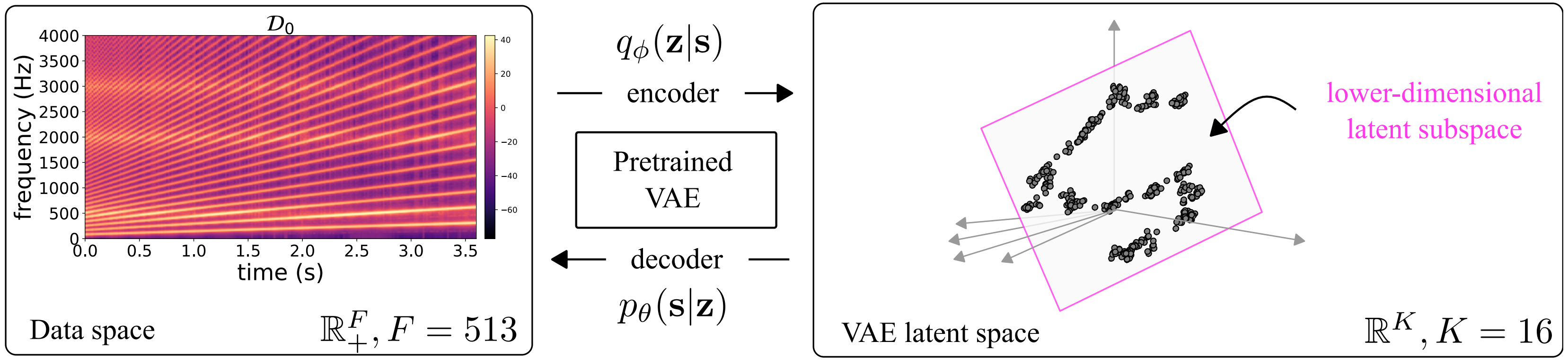
$$\boldsymbol{\mu}_\phi(\mathcal{D}_i) = \mathbb{E}_{\hat{q}_\phi^{(i)}(\mathbf{z})}[\mathbf{z}] = \frac{1}{\#\mathcal{D}_i} \sum_{\mathbf{s}_n \in \mathcal{D}_i} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{s}_n)}[\mathbf{z}] = \frac{1}{\#\mathcal{D}_i} \sum_{\mathbf{s}_n \in \mathcal{D}_i} \boldsymbol{\mu}_\phi(\mathbf{s}_n).$$

- In the following, without loss of generality, we assume centered latent vectors:

$$\mathbf{z} \leftarrow \mathbf{z} - \boldsymbol{\mu}_\phi(\mathcal{D}_i).$$

Source-filter latent subspace learning

- **Intuition**: Because one single factor f_i varies in $\mathcal{D}_i$, we expect the corresponding latent vectors to live in a lower-dimensional manifold of the original latent space $\mathbb{R}^K$.



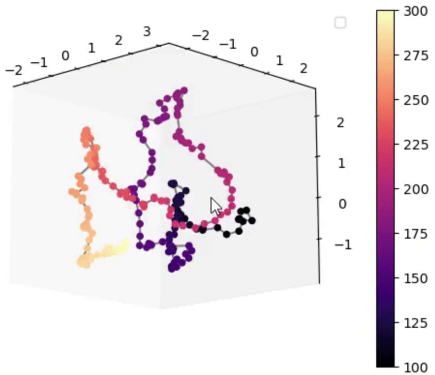
- We assume this manifold to be a linear subspace characterized by its semi-orthogonal basis matrix $\mathbf{U}_i \in \mathbb{R}^{K\times M_i}$, $M_i < K$, computed by solving

$$\min_{\mathbf{U}\in\mathbb{R}^{K\times M_i}} \mathbb{E}_{\hat{q}_\phi^{(i)}(\mathbf{z})}\left[\| \mathbf{z} - \mathbf{U}\mathbf{U}^\top \mathbf{z} \|_2^2\right], \qquad s.t.\ \mathbf{U}^\top\mathbf{U} = \mathbf{I}.$$

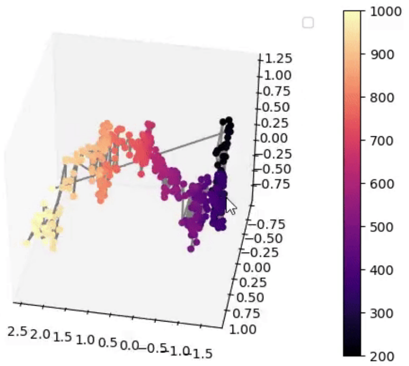
- As in principal component analysis (PCA), a closed-form solution is obtained by an eigendecomposition of a symmetric positive semi-definite matrix.

Trajectories in the learned latent subspaces

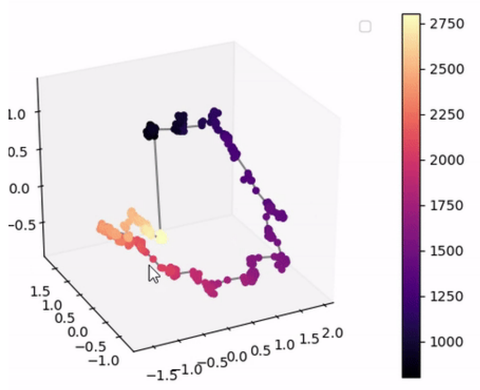
- For each element $\mathbf{s} \in \mathcal{D}_i$, we plot $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{s})}[\mathbf{U}_i^\top \mathbf{z}] = \mathbf{U}_i^\top \boldsymbol{\mu}_\phi(\mathbf{s}) \in \mathbb{R}^{M_i}$ $(M_i = 3)$.



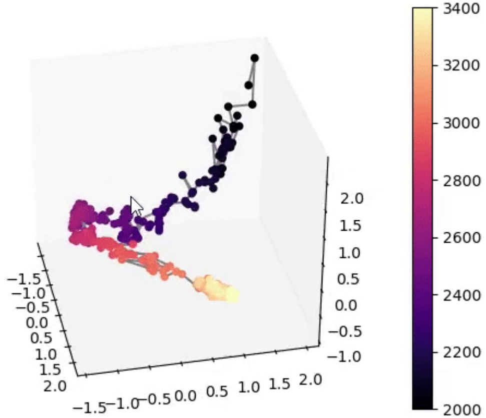
f_0 *latent trajectory*



f_1 *latent trajectory*



f_2 *latent trajectory*



f_3 *latent trajectory*

- Two speech spectra with close values for the factor f_i have latent representations that are also close in the learned subspaces.

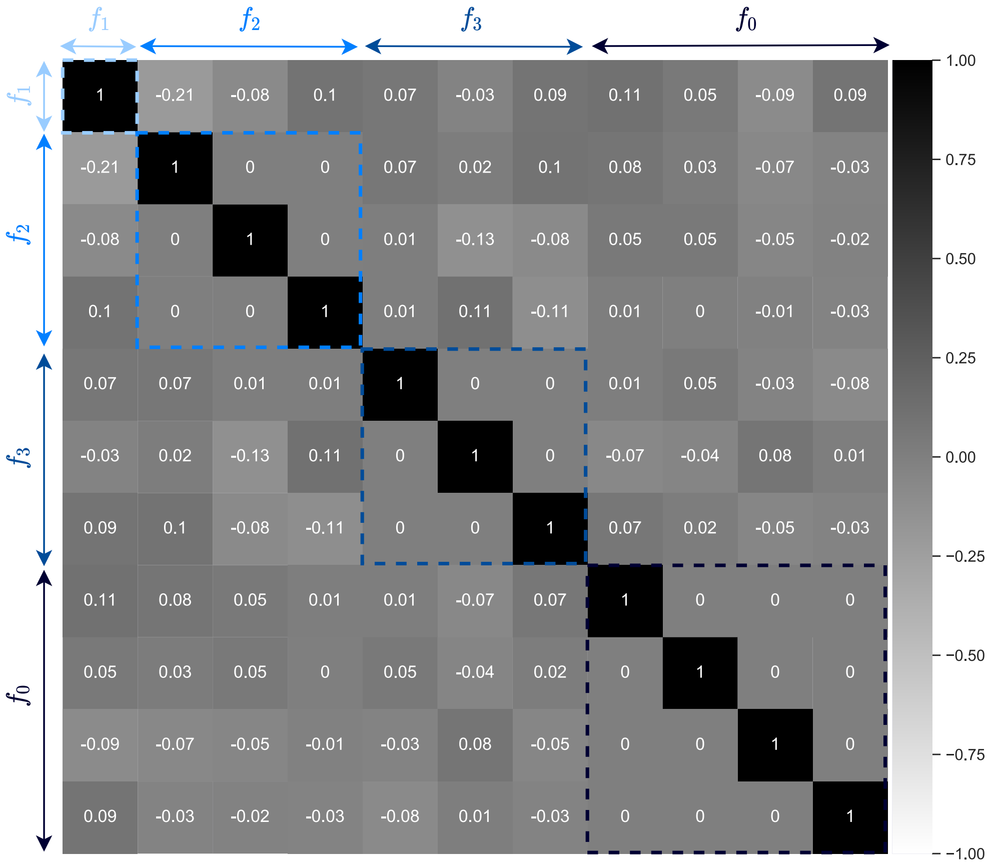
The latent representation learned by the VAE preserves the notion of proximity in terms of fundamental and formant frequencies.

Disentanglement analysis

The proposed approach offers a natural and straightforward way to **quantitatively measure** if the VAE managed to learn a **disentangled representation** of the source-filter characteristics of speech.

- By looking at the eigenvalues associated with the columns of $\mathbf{U}_i \in \mathbb{R}^{K \times M_i}$, we can measure the amount of variance that is retained by the projection $\mathbf{U}_i \mathbf{U}_i^\top$.
- If a small number of components M_i represents most of the variance, it indicates that only a few intrinsic dimensions of the latent space are dedicated to the factor f_i.
- If for two different factors f_i and f_j, the columns of $\mathbf{U}_i$ are orthogonal to those of $\mathbf{U}_j$, the two factors are encoded in orthogonal subspaces and therefore disentangled (Higgins et al., 2018).

I. Higgins et al., Towards a definition of disentangled representations. arXiv preprint arXiv:1812.02230, 2018.



- We choose M_i so as to retain 80% of the data variance after projection onto the latent subspaces. It gives

$$M_0 = 4, M_1 = 1, M_2 = 3, M_3 = 3.$$

- We compute the dot product between all pairs of unit vectors in the matrices $\{\mathbf{U}_i \in \mathbb{R}^{K \times M_i}\}_{i=0}^{3}$.

- Except for a correlation value of -0.21 between f_1 and the 1st component of f_2, all values are below 0.13 (in absolute value).

This analysis confirms the orthogonality of the source-filter latent subspaces and the disentanglement of the corresponding factors in the VAE latent space.

Conclusion

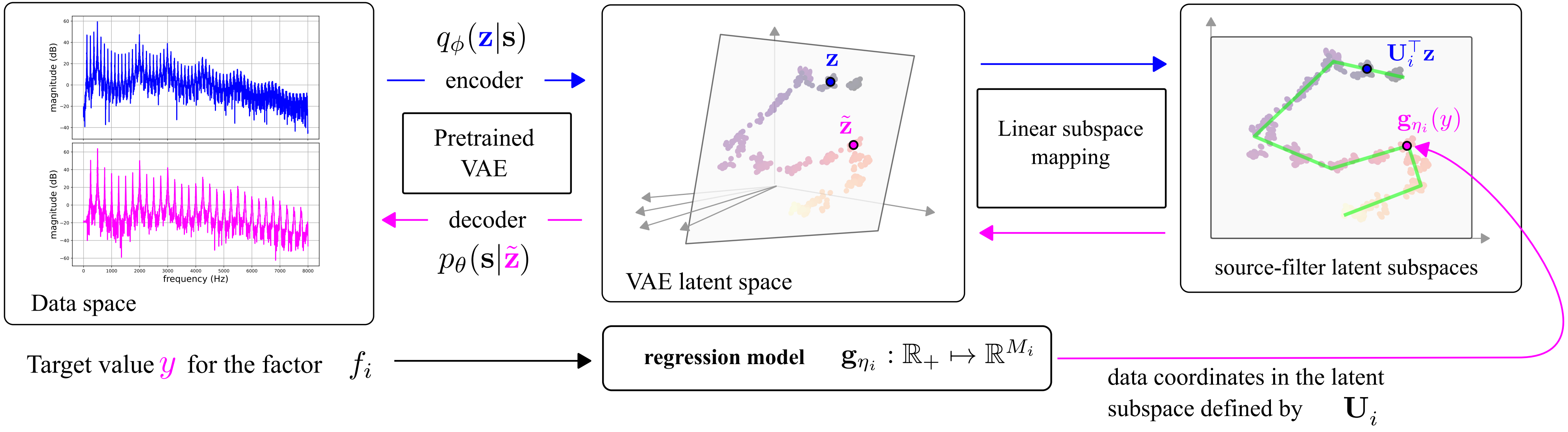
- Using only a few seconds of artificially generated speech, we put in evidence that a VAE trained in an unsupervised manner learns a latent representation that is consistent with the source-filter model of speech production.

  Indeed, the fundamental frequency and first formant frequencies are encoded in orthogonal subspaces of the original VAE latent space.

- It suggests that we could manipulate one factor in its latent subspace without affecting the others, similarly as how humans produce speech according to the source-filter model.

Moving in the source-filter latent subspaces

Disentangled speech manipulation in the VAE latent space

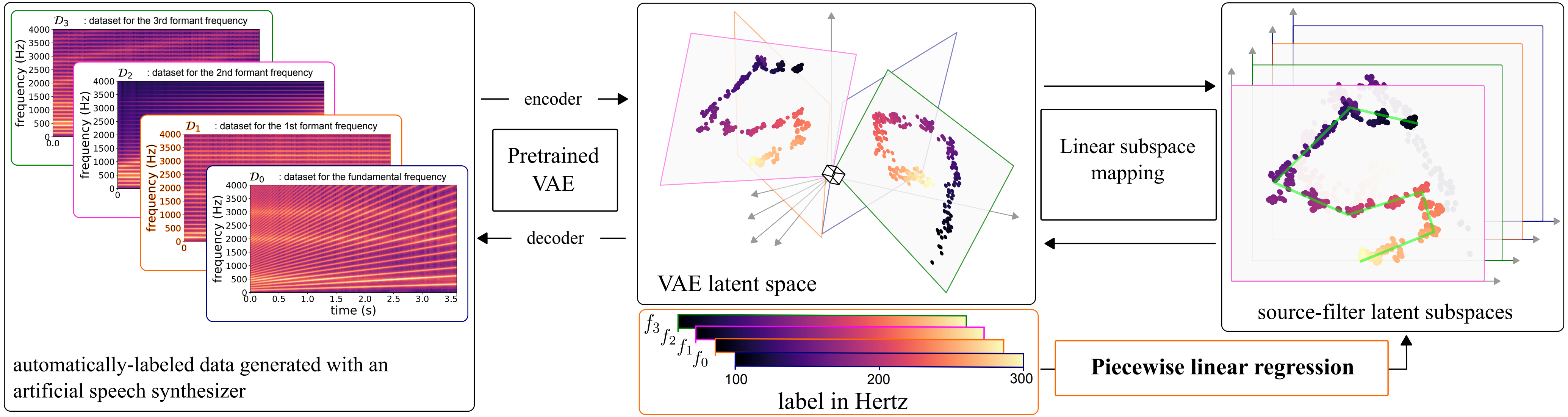


We can transform a speech spectrum by analyzing it with the VAE encoder, applying the following affine transformation, and resynthesizing with the VAE decoder:

$$\tilde{\mathbf{z}} = \mathbf{z} - \mathbf{U}_i \mathbf{U}_i^\top \mathbf{z} + \mathbf{U}_i \mathbf{g}_{\eta_i}(y).$$

This transformation allows us to **move only in the subspace associated with f_i**, leaving other source-filter factors unchanged thanks to the orthogonality property.

Weakly-supervised piecewise linear regression learning

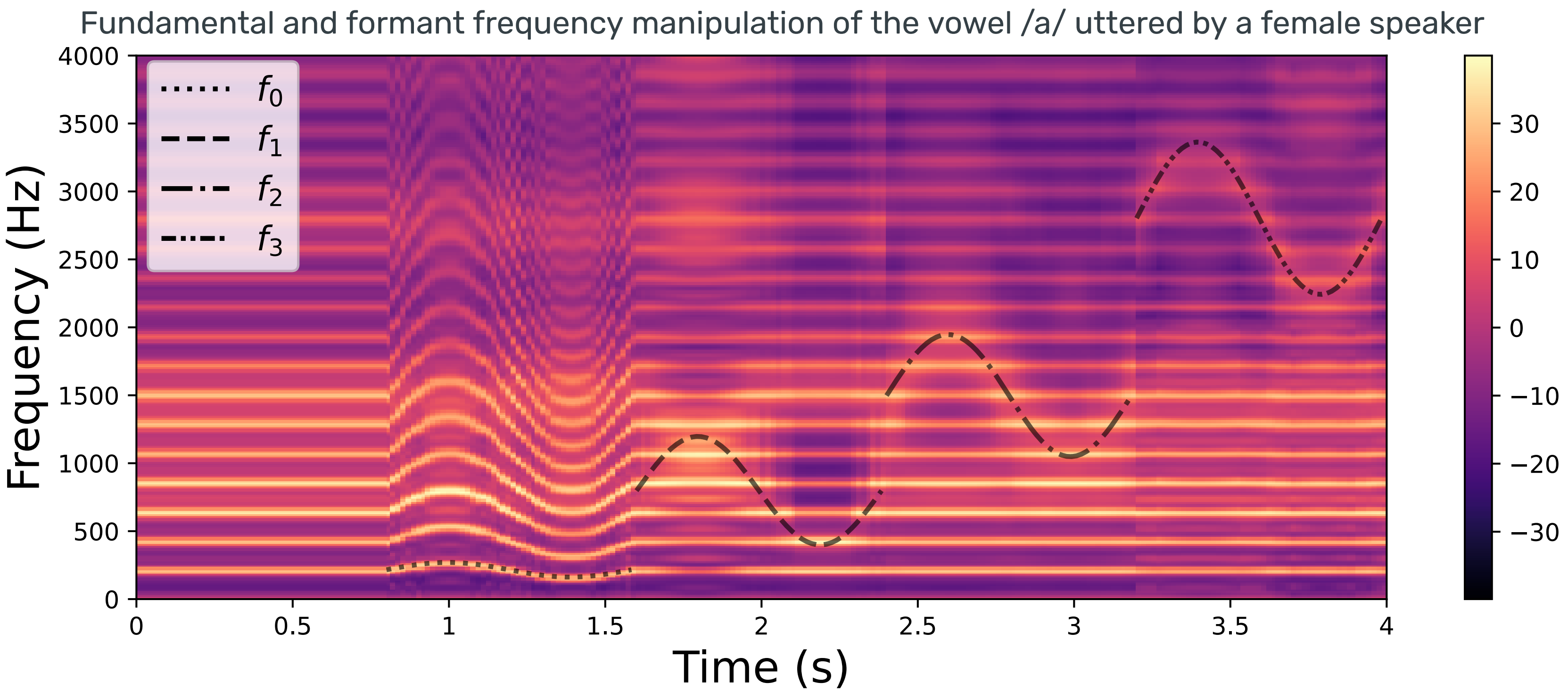


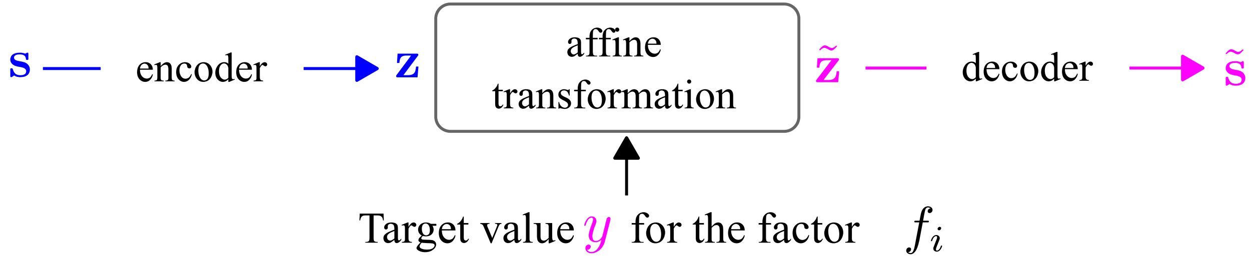
Making now use of the labels in $\mathcal{D}_i$, we learn a piecewise-linear regression model $\mathbf{g}_{\eta_i} : \mathbb{R}_+ \mapsto \mathbb{R}^{M_i}$ from the value $y \in \mathbb{R}_+$ of the factor f_i to the data coordinates $\mathbf{U}_i^\top \mathbf{z}$ in the latent subspace:

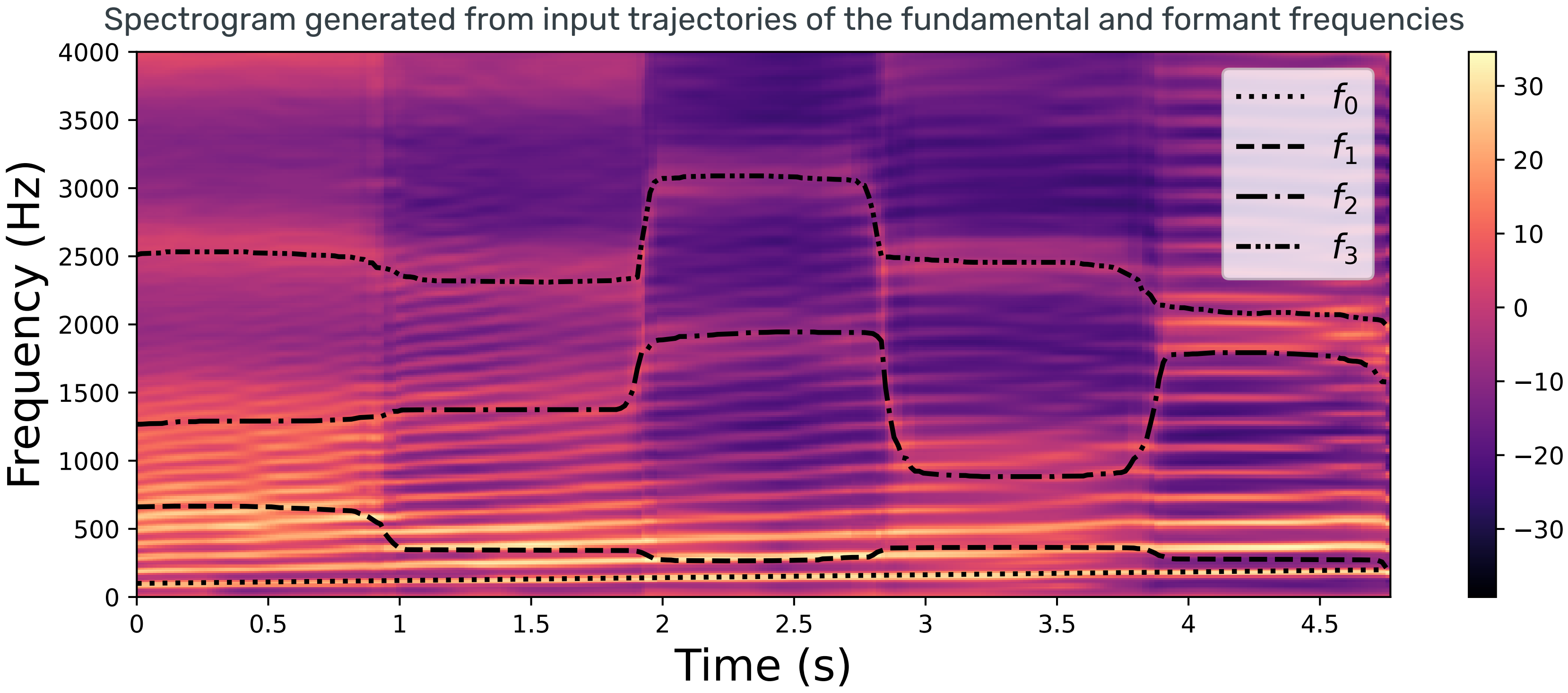
$$\eta_i = \underset{\eta}{\arg\min}\ \mathbb{E}_{\hat{q}_\phi^{(i)}(\mathbf{z},y)}\left[\|\mathbf{g}_\eta(y) - \mathbf{U}_i^\top \mathbf{z}\|_2^2\right],$$

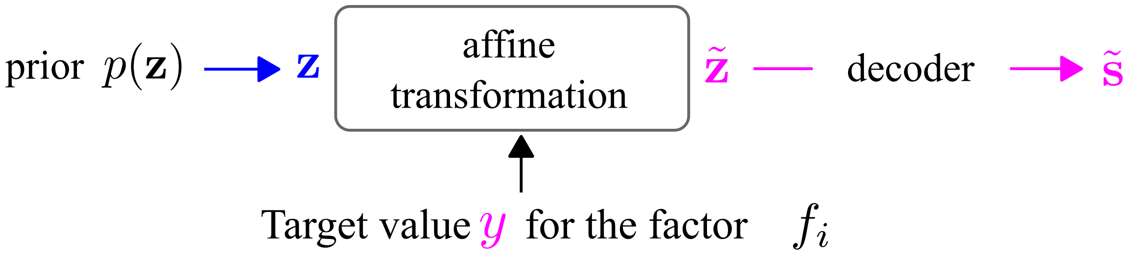
where $\hat{q}_\phi^{(i)}(\mathbf{z}, y) = \int q_\phi(\mathbf{z}|\mathbf{s})\hat{p}^{(i)}(\mathbf{s}, y)d\mathbf{s}$ and $\hat{p}^{(i)}(\mathbf{s}, y)$ is the empirical distribution of $\mathcal{D}_i = \{(\mathbf{s}_n, y_n)\}_n$.

Qualitative results

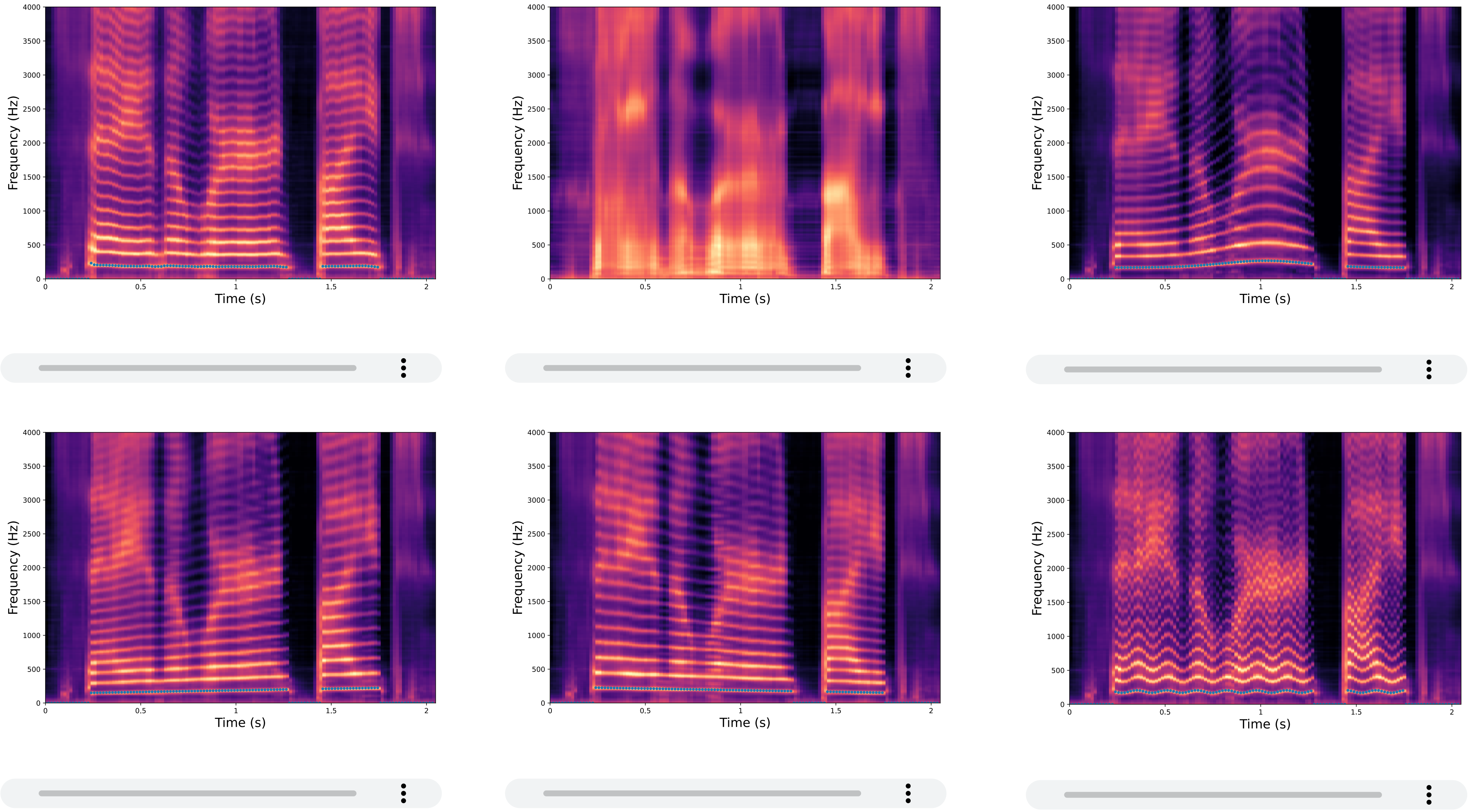

Fundamental and formant frequency manipulation of the vowel /a/ uttered by a female speaker
Frequency (Hz)
Time (s)
f_0
f_1
f_2
f_3
4000
3500
3000
2500
2000
1500
1000
500
0
0.5
1
1.5
2
2.5
3
3.5
4
30
20
10
0
−10
−20
−30


$\mathbf{s}$
encoder
$\mathbf{z}$
affine transformation
$\tilde{\mathbf{z}}$
decoder
$\tilde{\mathbf{s}}$
Target value y for the factor f_i





We have defined a deep generative model of speech spectrograms that is **conditioned on interpretable trajectories** of the fundamental and formant frequencies.



(top left) reconstructed w/o modification, (top middle) whispered spectrogram obtained with $\tilde{\mathbf{z}} = \mathbf{z} - \mathbf{U}_0 \mathbf{U}_0^\top \mathbf{z}$, (other) various f_0 transformations. Waveforms are obtained from the spectrograms using WaveGlow (Prenger et al., 2019).

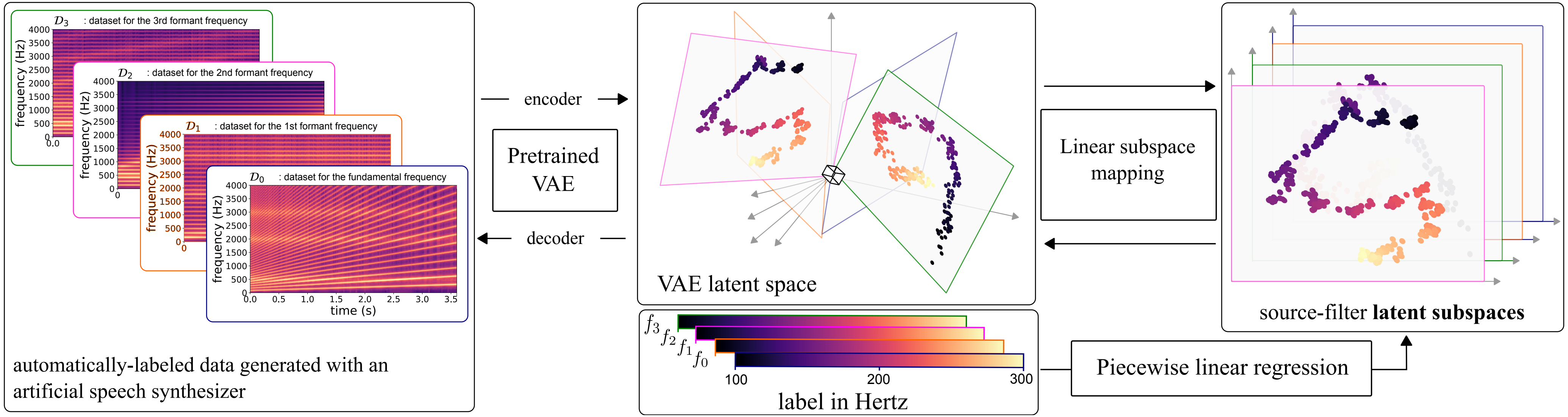
Quantitative results

We refer you to the paper, or you can ask for the backup slides.

In summary, a quantitative analysis using datasets of English vowels and speech utterances confirms that

- source-filter factors can be manipulated accurately, especially f_0;
- varying one factor (e.g., f_0) has little effect on the others (e.g., the formants).

Conclusion



In this work, given a VAE trained on hours of unlabeled speech data and a few seconds of automatically-labeled data generated with an artificial speech synthesizer,

- we put in evidence that the latent representation learned by a VAE is consistent with the source-filter model of speech production (Fant, 1970);
- we proposed a weakly-supervised method to learn how to move in the VAE latent space, so as to perform disentangled speech manipulations.

Future work

- Take the non-linear nature of the manifolds into account;
- Address the phase reconstruction issue, with better neural vocoders or working directly in the time domain (Caillon and Esling, 2021);
- Extend the approach to multi-microphone and reverberant signals, to learn both spectro-temporal and spatial representations of speech;
- Exploit the invariance of the projected representations to perform analysis (e.g., f_0 estimation);
- Leverage the proposed conditional deep generative speech model to guide VAE-based speech enhancement methods with the pitch information.

A. Caillon and P. Esling, RAVE: A variational autoencoder for fast and high-quality neural audio synthesis, arXiv preprint arXiv:2111.05011, 2021.

Thank you

Code and audio examples available online

https://samsad35.github.io/site-sfvae/

Quantitative results

Dataset

12 English vowels $\times$ 50 male and 50 female speakers, labeled with fundamental and formant frequencies.

	Min (Hz)	Max (Hz)	Step (Hz)
f_0	100	300	1
f_1	300	900	10
f_2	1100	2700	20
f_3	2200	3200	20

Task

We transform each vowel by varying one single factor f_i at a time.

Metrics

- **Accuracy and disentanglement** (lower is better)

  We compute the relative absolute error $\delta f_i = |\hat{y} - y|/y \times 100\%$, where y is the target value for f_i and $\hat{y}$ its estimation on the output transformed signal.

- **Speech naturalness** (higher is better)

  We use NISQA (Mittag and Möller, 2020), an objective metric developed in the context of speech transformation algorithms to be highly correlated with subjective mean opinion scores.

G. Mittag and S. Möller, Deep learning based assessment of synthetic speech naturalness, Interspeech, 2020.

Methods

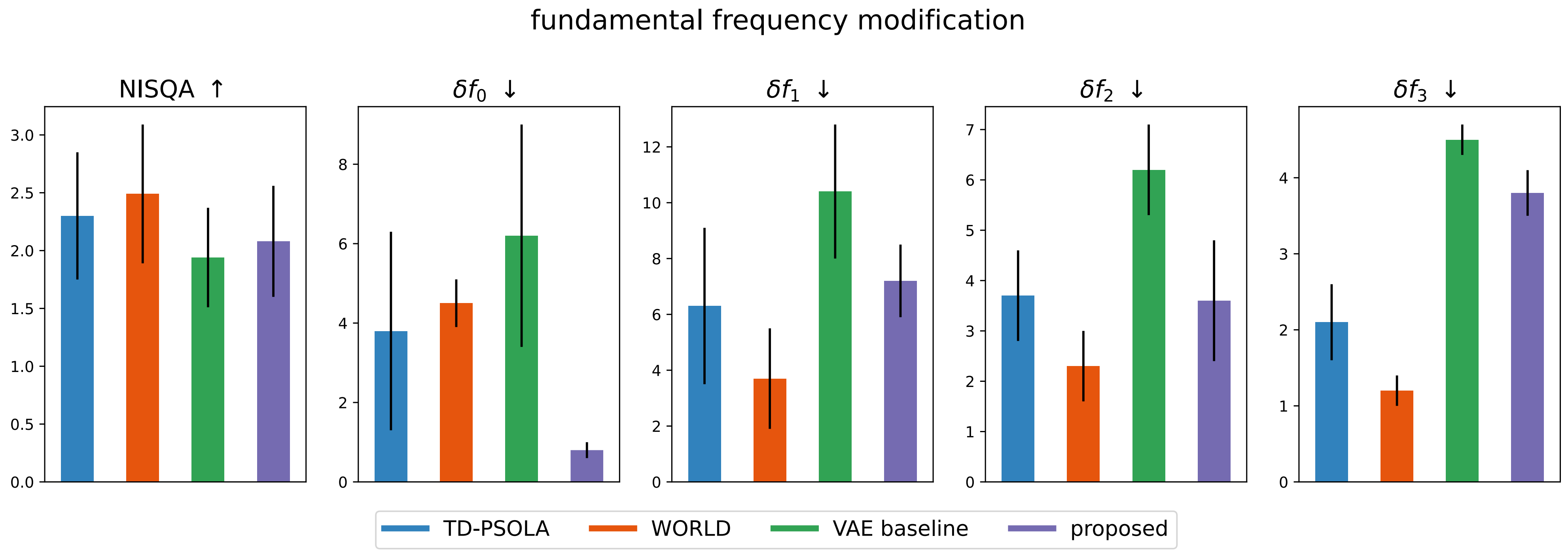
- **TD-PSOLA** (Moulines and Charpentier, 1990) performs f_0 modification through a decomposition of the signal into pitch-synchronized overlapping frames.
- **WORLD** (Morise et al., 2016) is a vocoder also used for f_0 modification. It decomposes the signal into three components characterizing f_0, the aperiodicity, and the spectral envelope.
- The **VAE baseline** (Hsu et al. 2017) consists in applying translations directly in the VAE latent space:

$$\tilde{\mathbf{z}} = \mathbf{z} - \boldsymbol{\mu}_{\text{src}} + \boldsymbol{\mu}_{\text{trgt}},$$

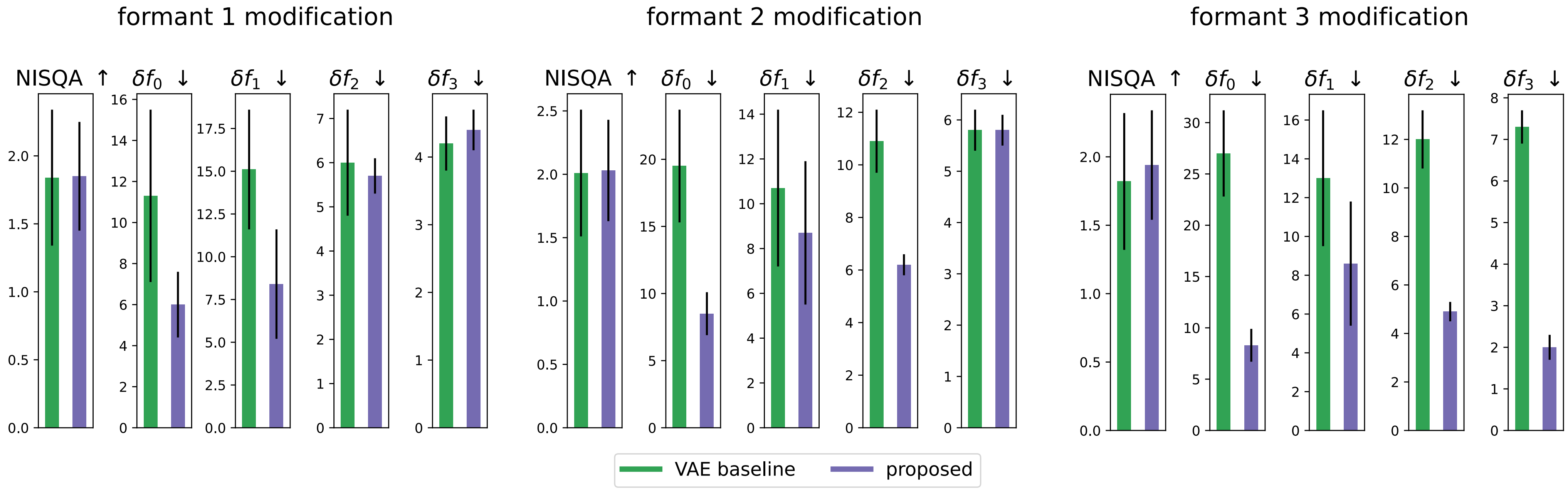
  where $\boldsymbol{\mu}_{\text{src}}$ and $\boldsymbol{\mu}_{\text{trgt}}$ are predefined latent attribute representations associated with the source and target values of the factor to be modified, respectively.

  Computing $\boldsymbol{\mu}_{\text{src}}$ requires analyzing the input speech signal (e.g., to estimate f_0), which is not the case of the proposed method that only relies on a projection of $\mathbf{z}$.

E. Moulines and F. Charpentier, Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, Speech Communication, 1990.
M. Morise et al., World: a vocoder-based high-quality speech synthesis system for real-time applications, IEICE TIS, 2016.
W.-N. Hsu et al., Learning latent representations for speech generation and transformation, Interspeech, 2017.



- The proposed method always outperforms the baseline.
- δf_0 is lower than 1 % for the proposed method $\rightarrow$ very good precision in f_0 manipulation.
- WORLD obtains the best performance in terms of disentanglement $(\delta f_i, i > 0)$ because the source and filter contributions are decoupled in the architecture of the vocoder.
- Traditional signal processing methods obtain the best performance in terms of speech naturalness (NISQA) probably because they directly operate in the time domain (no phase reconstruction issue).



- In terms of accuracy, the proposed method always outperforms the baseline (by 7%, 5% and 5% for f_1, f_2 and f_3, respectively.)
- In terms of disentanglement, the pitch is much less affected by formant manipulations with the proposed method.

- A similar analysis on a dataset of short speech utterances (TIMIT) leads to similar conclusion.
- This dataset is phonemically richer than the isolated vowels dataset.
- However, it is not labeled with the fundamental and formant frequencies, so the groud truth required to measure disentanglement is estimated on the original speech signals, which makes the evaluation less reliable.

- The objective of this study is not to compete with traditional signal processing methods such as TD-PSOLA and WORLD for pitch shifting.
- It is rather to advance on the understanding of deep generative modeling of speech signals and to compare honestly with highly-specialized traditional systems.
- TD-PSOLA and WORLD exploit signal models that are specifically designed for the task at hand, while the proposed method is data-driven and the exact same methodology applies for modifying f_0 or the formant frequencies.
- TD-PSOLA is still a strong baseline that is difficult to outperform with deep learning techniques, see e.g. controllable LPCNet (Morrison et al., 2020).

M. Morrison et al., Controllable Neural Prosody Synthesis, Interspeech, 2020.

VAE model training

Parameters estimation

- Direct maximization of the marginal likelihood is intractable due to non-linearities.
- For any distribution $q_\phi(\mathbf{z}|\mathbf{x})$, we have (Neal and Hinton, 1999; Jordan et al. 1999)

$$\ln p(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\phi,\theta) + D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x})),$$

where $\mathcal{L}(\mathbf{x};\phi,\theta)$ is the evidence lower bound (ELBO) defined by

$$\mathcal{L}(\mathbf{x};\phi,\theta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p(\mathbf{x},\mathbf{z};\theta) - \ln q_\phi(\mathbf{z}|\mathbf{x})].$$

R.M. Neal and G.E. Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants, in M. I. Jordan (Ed.), *Learning in graphical models*, 1999.
M.I. Jordan et al., An introduction to variational methods for graphical models, Machine Learning, 1999.

Parameters estimation

- Direct maximization of the marginal likelihood is intractable due to non-linearities.
- For any distribution $q_\phi(\mathbf{z}|\mathbf{x})$, we have (Neal and Hinton, 1999; Jordan et al. 1999)

$$\ln p(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\phi,\theta) + D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x})),$$

where $\mathcal{L}(\mathbf{x};\phi,\theta)$ is the evidence lower bound (ELBO) defined by

$$\mathcal{L}(\mathbf{x};\phi,\theta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p(\mathbf{x},\mathbf{z};\theta) - \ln q_\phi(\mathbf{z}|\mathbf{x})].$$

Problem #1

$$\max_\theta \mathcal{L}(\mathbf{x};\phi,\theta),$$

where $\mathcal{L}(\mathbf{x};\phi,\theta) \leq \ln p(\mathbf{x};\theta)$

R.M. Neal and G.E. Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants, in M. I. Jordan (Ed.), *Learning in graphical models*, 1999.
M.I. Jordan et al., An introduction to variational methods for graphical models, Machine Learning, 1999.

Parameters estimation

- Direct maximization of the marginal likelihood is intractable due to non-linearities.
- For any distribution $q_\phi(\mathbf{z}|\mathbf{x})$, we have (Neal and Hinton, 1999; Jordan et al. 1999)

$$\ln p(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\phi,\theta) + D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x})),$$

where $\mathcal{L}(\mathbf{x};\phi,\theta)$ is the evidence lower bound (ELBO) defined by

$$\mathcal{L}(\mathbf{x};\phi,\theta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p(\mathbf{x},\mathbf{z};\theta) - \ln q_\phi(\mathbf{z}|\mathbf{x})].$$

Problem #1

$$\max_\theta \mathcal{L}(\mathbf{x};\phi,\theta),$$

where $\mathcal{L}(\mathbf{x};\phi,\theta) \leq \ln p(\mathbf{x};\theta)$

Problem #2

$$\max_\phi \mathcal{L}(\mathbf{x};\phi,\theta)$$

$$\Leftrightarrow \min_\phi D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x}))$$

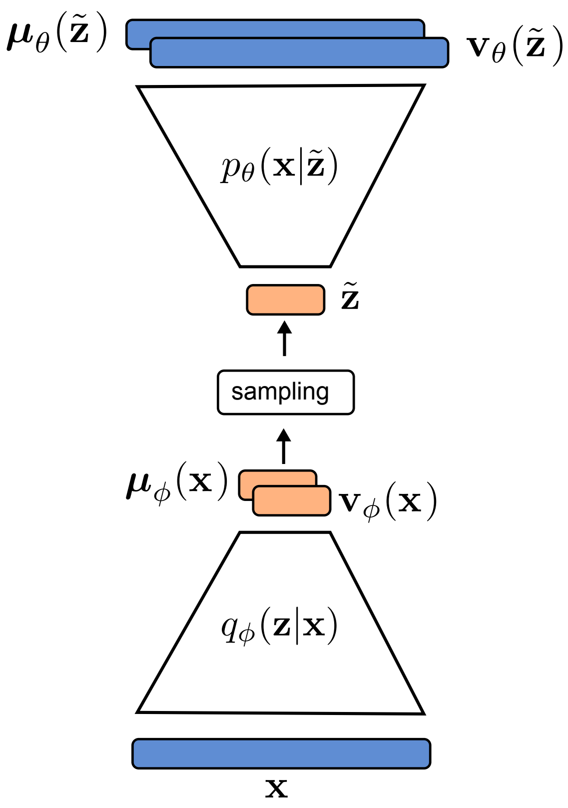
R.M. Neal and G.E. Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants, in M. I. Jordan (Ed.), *Learning in graphical models*, 1999.
M.I. Jordan et al., An introduction to variational methods for graphical models, Machine Learning, 1999.

ELBO

The ELBO is now fully defined:

$$\begin{aligned}\mathcal{L}(\mathbf{x};\phi,\theta) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p(\mathbf{x},\mathbf{z};\theta) - \ln q_\phi(\mathbf{z}|\mathbf{x})] \\ &= \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{reconstruction accuracy}} - \underbrace{D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z}))}_{\text{regularization}}.\end{aligned}$$

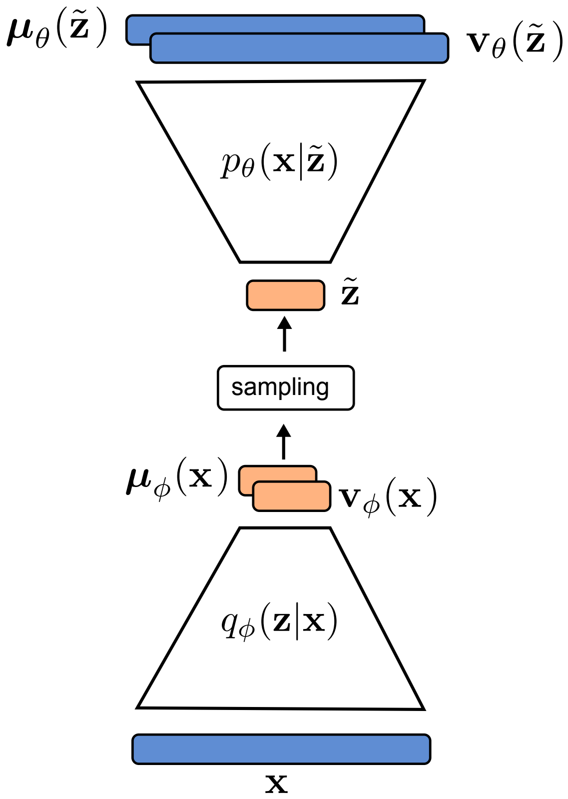
- prior: $p(\mathbf{z}) = \mathcal{N}(\mathbf{z};\mathbf{0},\mathbf{I})$
- likelihood model: $p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}\left(\mathbf{x};\boldsymbol{\mu}_\theta(\mathbf{z}), \operatorname{diag}\left\{\mathbf{v}_\theta(\mathbf{z})\right\}\right)$
- inference model: $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z};\boldsymbol{\mu}_\phi(\mathbf{x}), \operatorname{diag}\left\{\mathbf{v}_\phi(\mathbf{x})\right\}\right)$



ELBO

The ELBO is now fully defined:

$$\begin{aligned}\mathcal{L}(\mathbf{x};\phi,\theta) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p(\mathbf{x},\mathbf{z};\theta) - \ln q_\phi(\mathbf{z}|\mathbf{x})] \\ &= \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{reconstruction accuracy}} - \underbrace{D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))}_{\text{regularization}}.\end{aligned}$$



- prior: $p(\mathbf{z}) = \mathcal{N}(\mathbf{z};\mathbf{0},\mathbf{I})$
- likelihood model: $p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}\left(\mathbf{x};\boldsymbol{\mu}_\theta(\mathbf{z}),\operatorname{diag}\{\mathbf{v}_\theta(\mathbf{z})\}\right)$
- inference model: $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z};\boldsymbol{\mu}_\phi(\mathbf{x}),\operatorname{diag}\{\mathbf{v}_\phi(\mathbf{x})\}\right)$

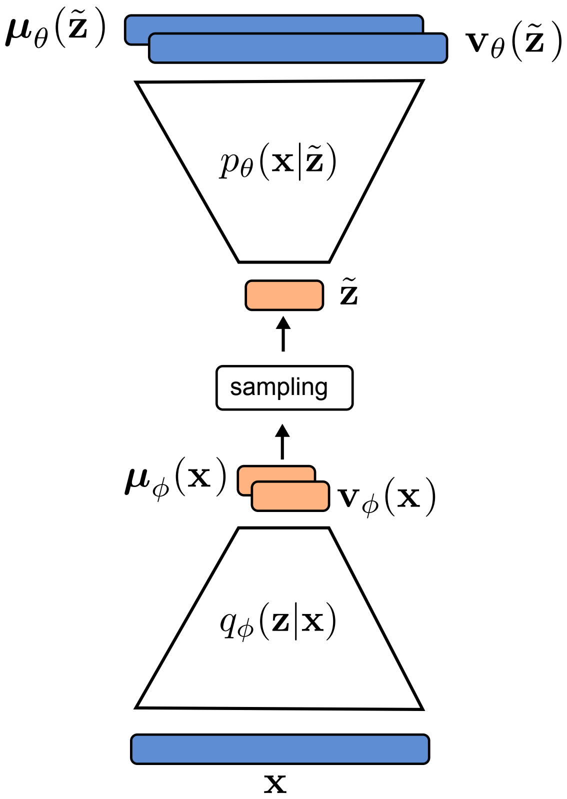
The reconstruction accuracy term is approximated with a Monte Carlo estimate:

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})] \approx \frac{1}{R}\sum_{r=1}^{R} \ln p_\theta(\mathbf{x}|\tilde{\mathbf{z}}_r), \qquad \text{with} \quad \tilde{\mathbf{z}}_r \sim q_\phi(\mathbf{z}|\mathbf{x}).$$

ELBO

The ELBO is now fully defined:

$$\begin{aligned}\mathcal{L}(\mathbf{x};\phi,\theta) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p(\mathbf{x},\mathbf{z};\theta) - \ln q_\phi(\mathbf{z}|\mathbf{x})] \\ &= \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{reconstruction accuracy}} - \underbrace{D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))}_{\text{regularization}}.\end{aligned}$$



- prior: $p(\mathbf{z}) = \mathcal{N}(\mathbf{z};\mathbf{0},\mathbf{I})$
- likelihood model: $p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}\left(\mathbf{x};\boldsymbol{\mu}_\theta(\mathbf{z}), \operatorname{diag}\{\mathbf{v}_\theta(\mathbf{z})\}\right)$
- inference model: $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z};\boldsymbol{\mu}_\phi(\mathbf{x}), \operatorname{diag}\{\mathbf{v}_\phi(\mathbf{x})\}\right)$

The reconstruction accuracy term is approximated with a Monte Carlo estimate, using the so-called reparametrization trick, to make the (sampled version of the) ELBO derivable w.r.t. ϕ:

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})] \approx \frac{1}{R}\sum_{r=1}^{R} \ln p_\theta(\mathbf{x}|\tilde{\mathbf{z}}_r), \qquad \begin{cases} \boldsymbol{\epsilon}_r & \sim \mathcal{N}(\mathbf{0},\mathbf{I}) \\ \tilde{\mathbf{z}}_r & = \boldsymbol{\mu}_\phi(\mathbf{x}) + \operatorname{diag}\{\mathbf{v}_\phi(\mathbf{x})\}^{\frac{1}{2}}\, \boldsymbol{\epsilon}_r \end{cases}.$$

Training procedure

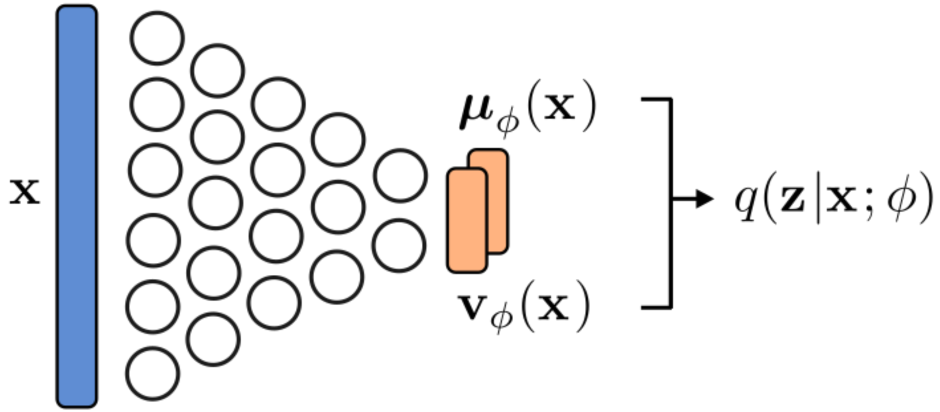
Step 1: Pick an example in the dataset

$$\mathcal{L}(\mathbf{x};\phi,\theta) = \ln p_\theta(\mathbf{x}|\tilde{\mathbf{z}}) - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))$$

Training procedure

Step 2: Map through the encoder

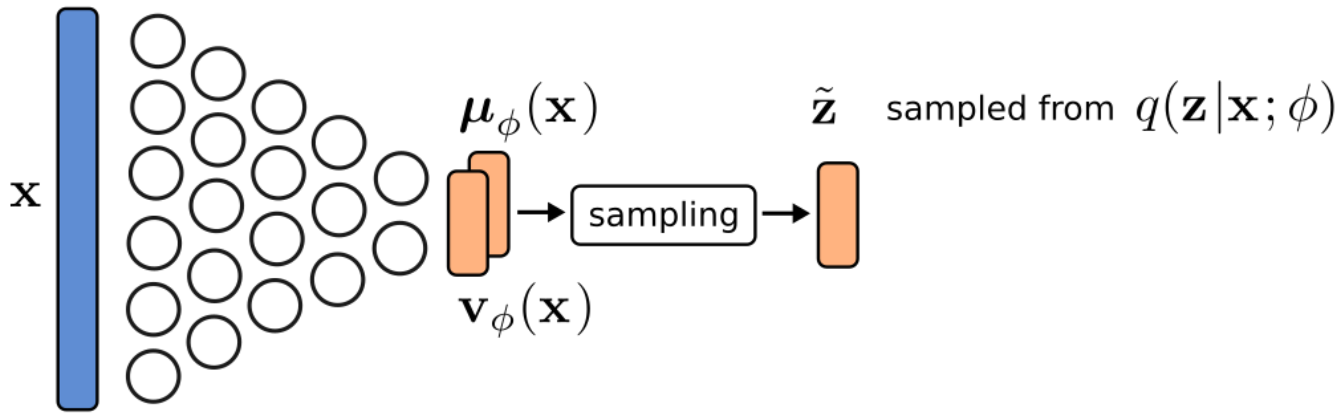
$$\mathcal{L}(\mathbf{x};\phi,\theta) = \ln p_\theta(\mathbf{x}|\tilde{\mathbf{z}}) - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))$$



Training procedure

Step 3: Sample from the inference model

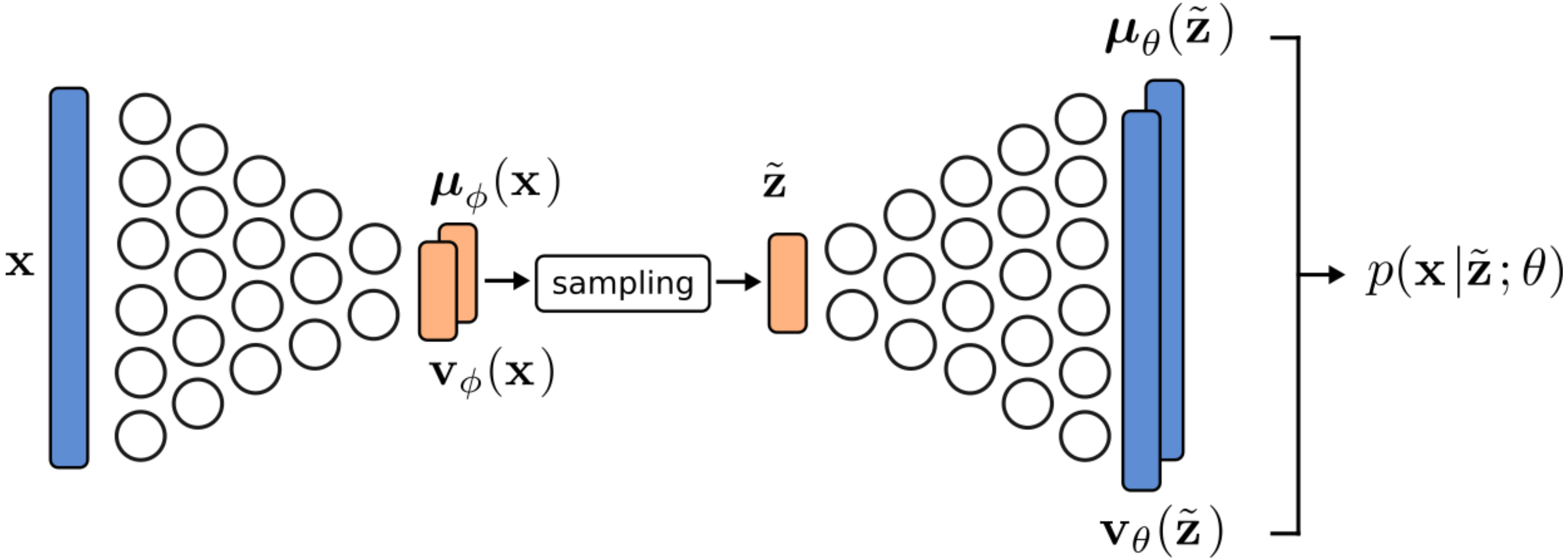
$$\mathcal{L}(\mathbf{x};\phi,\theta) = \ln p_\theta(\mathbf{x}|\tilde{\mathbf{z}}) - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))$$



Training procedure

Step 4: Map through the decoder

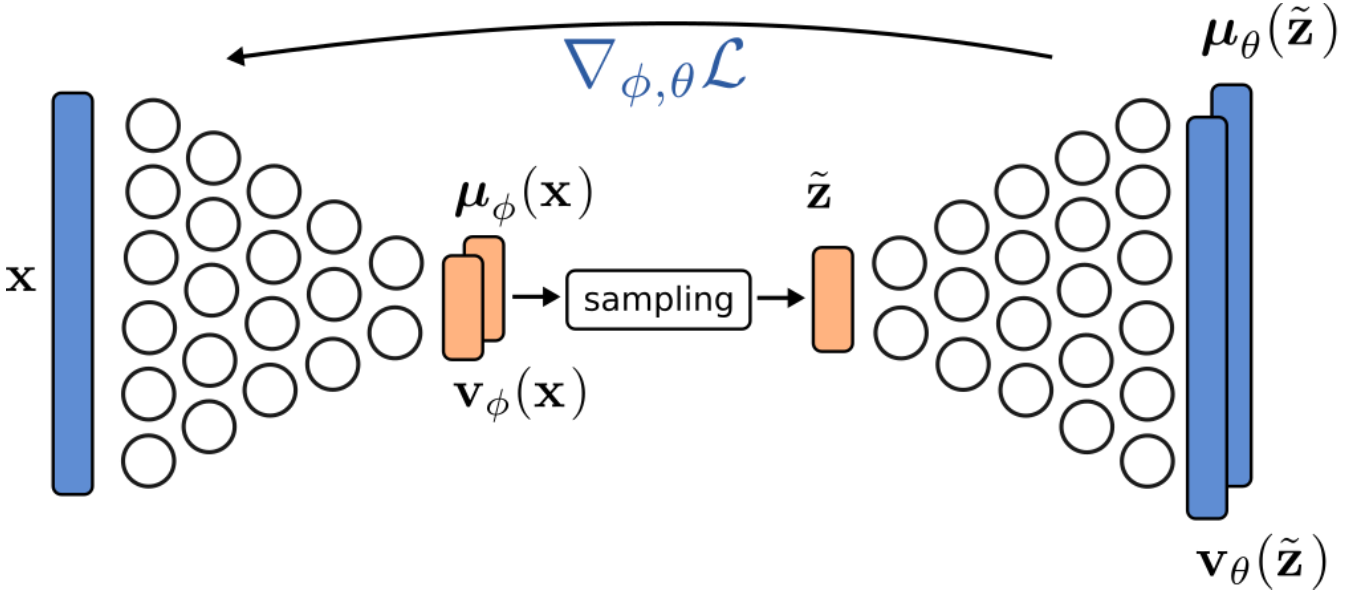
$$\mathcal{L}(\mathbf{x};\phi,\theta) = \ln p_\theta(\mathbf{x}|\tilde{\mathbf{z}}) - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))$$



Training procedure

Step 5: Gradient ascent step on the ELBO

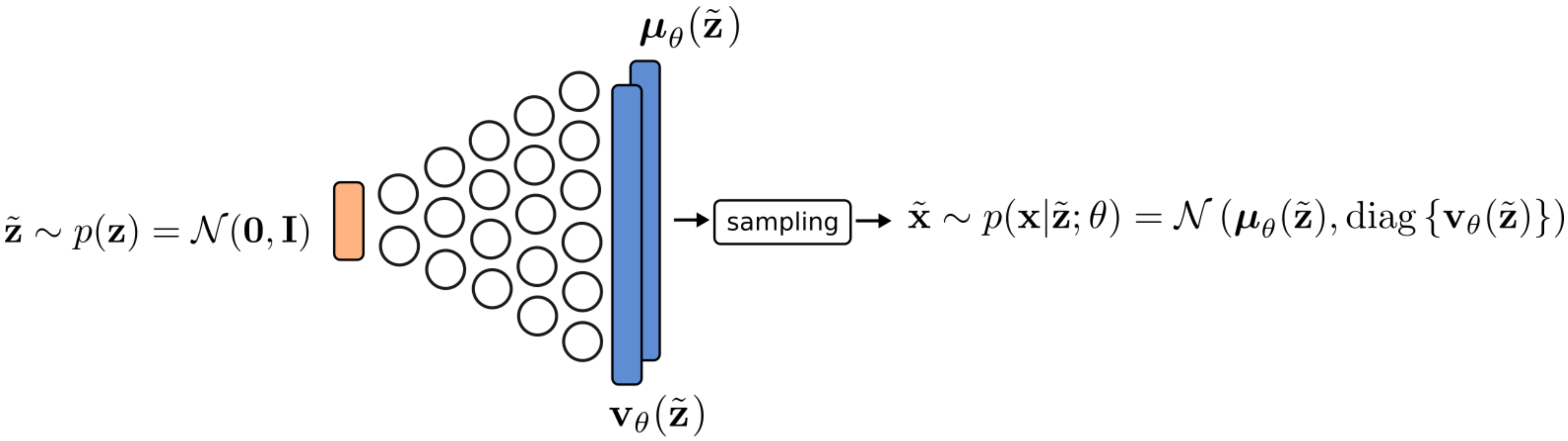
$$\mathcal{L}(\mathbf{x}; \phi, \theta) = \ln p_\theta(\mathbf{x}|\tilde{\mathbf{z}}) - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))$$



Encoder-decoder shape, which correspond to an inference-generation process.

In practice, one averages over mini batches before doing the backpropagation.

At test time



- The encoder was primarily introduced in order to estimate the parameters of the decoder.
- We do not need the encoder for generating new samples.
- But it is useful if we need to do inference.