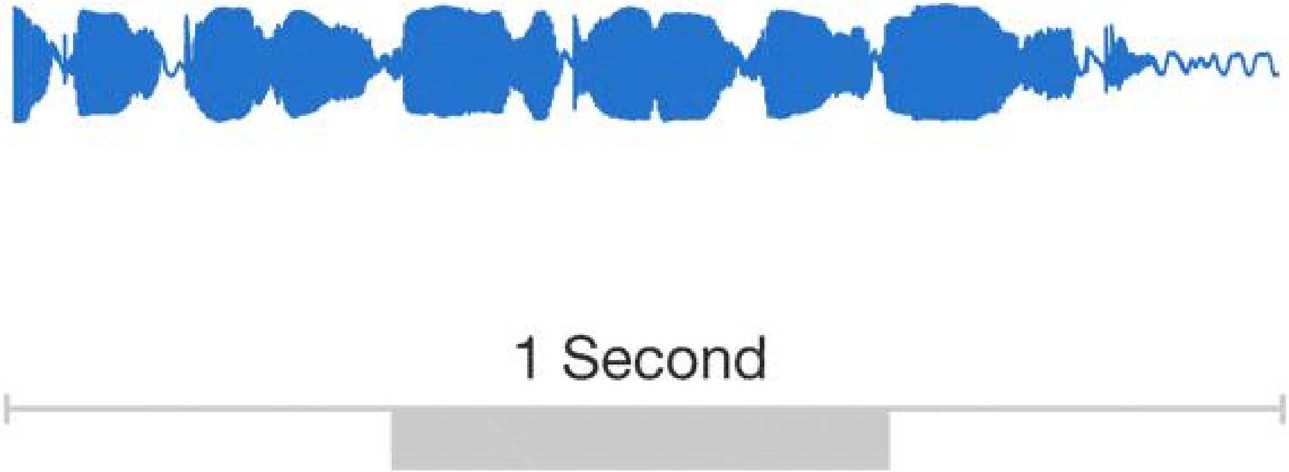
How We Factorize Speech

Dong Wang

2020.09.14

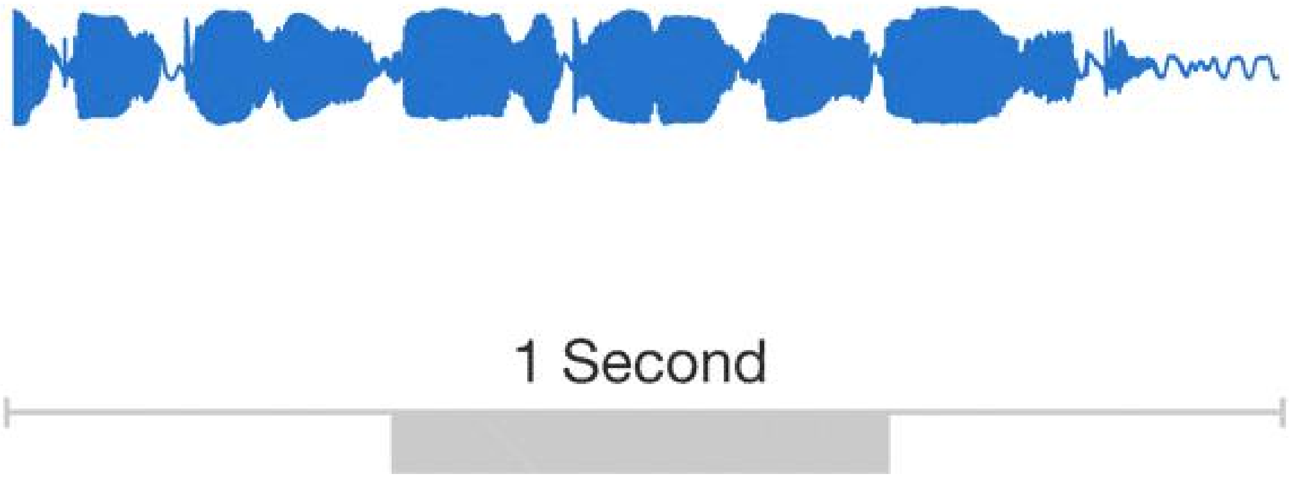
Speech signal



Q: What are special?

https://deepmind.com/blog/wavenet-generative-model-raw-audio/

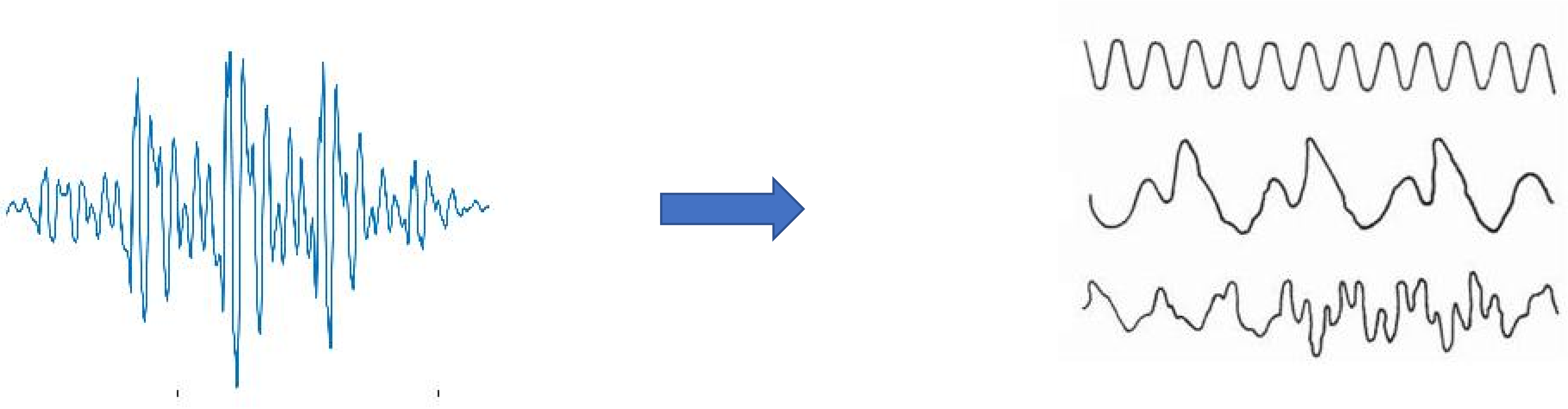
Speech signal



- *Random,periodic,constrained…*
- *Short-time stationary (in statistics)*
- *Long-time dependency (articulatory, linguistics, logical…)*

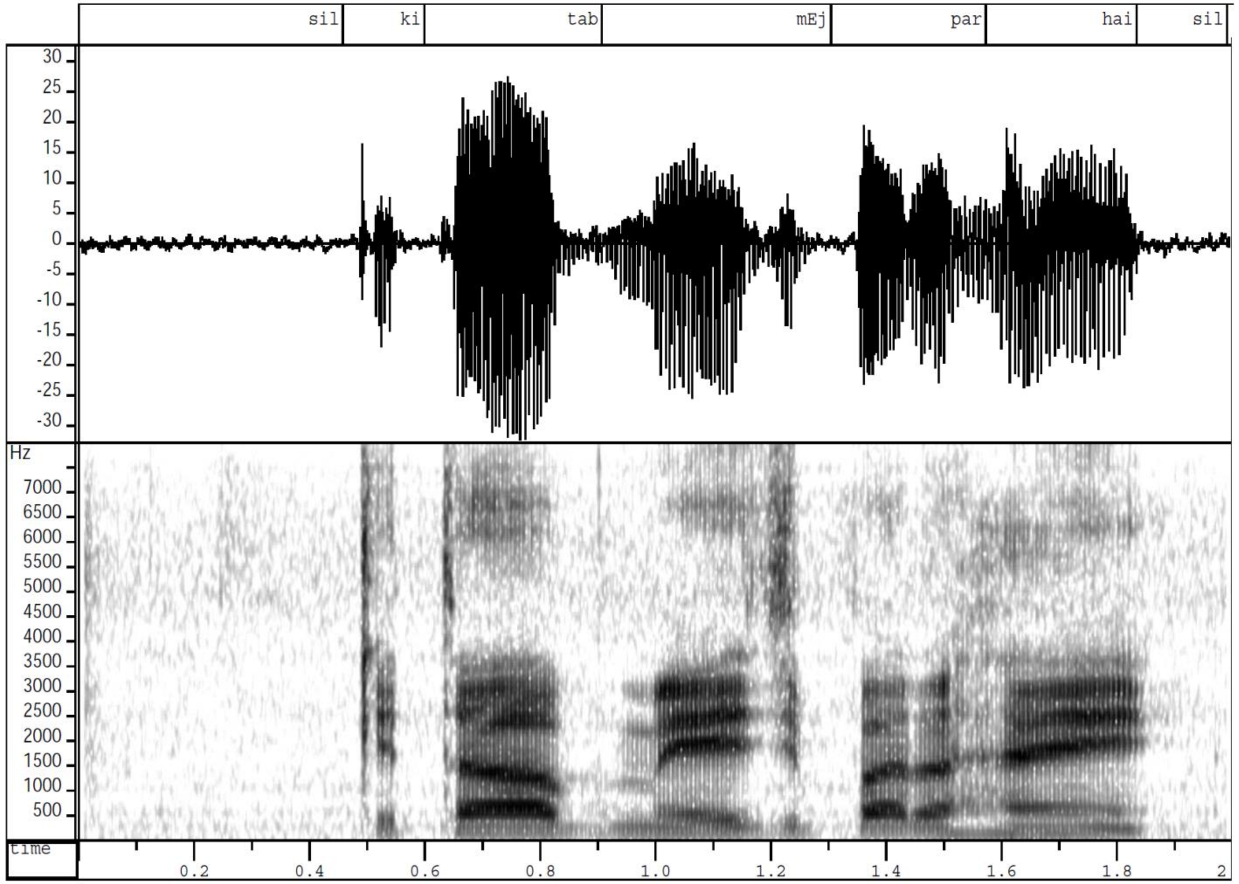
https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Speech analysis

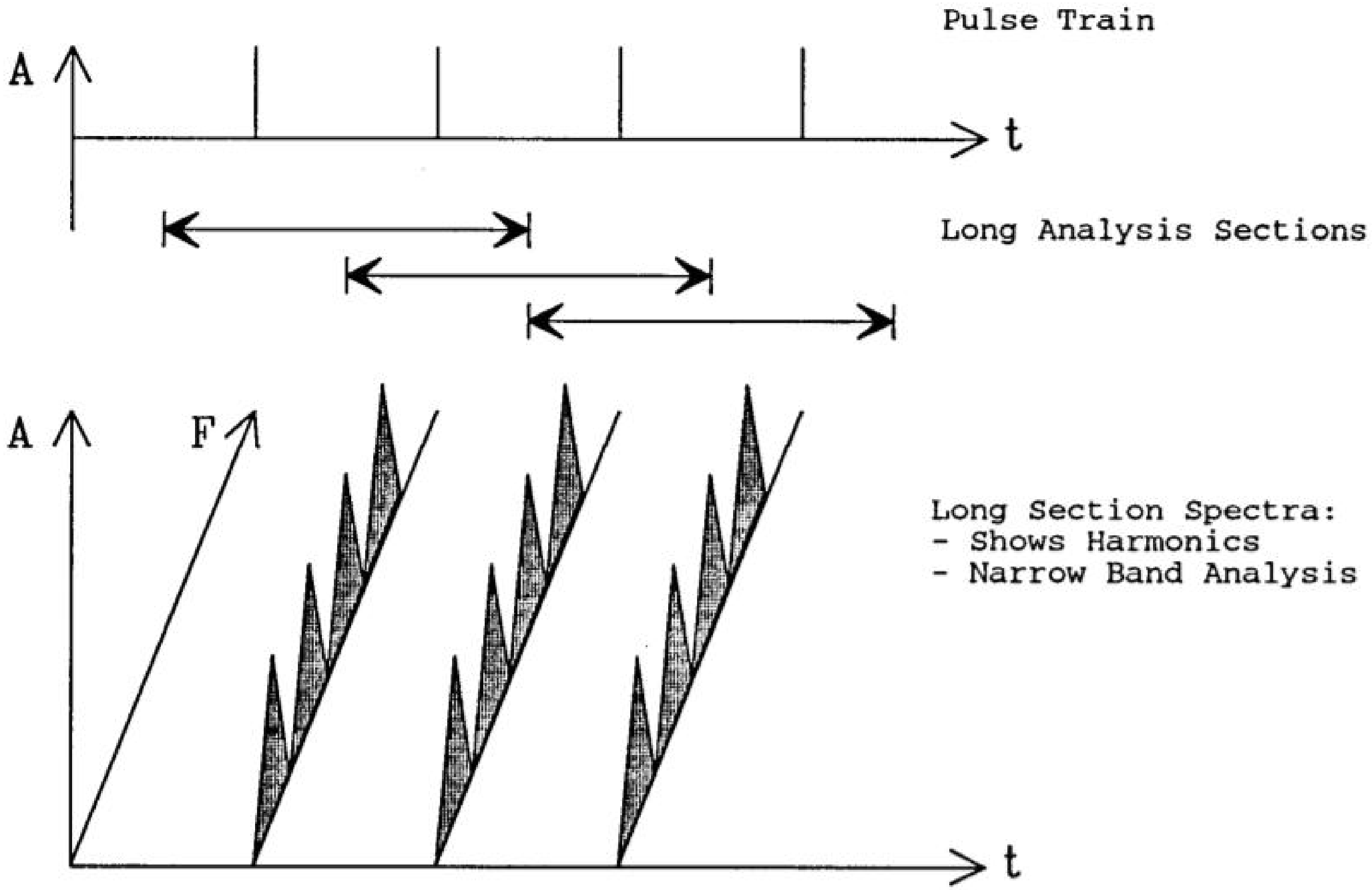


Decompose speech to simple components

Trivial decomposition by Fourier transform

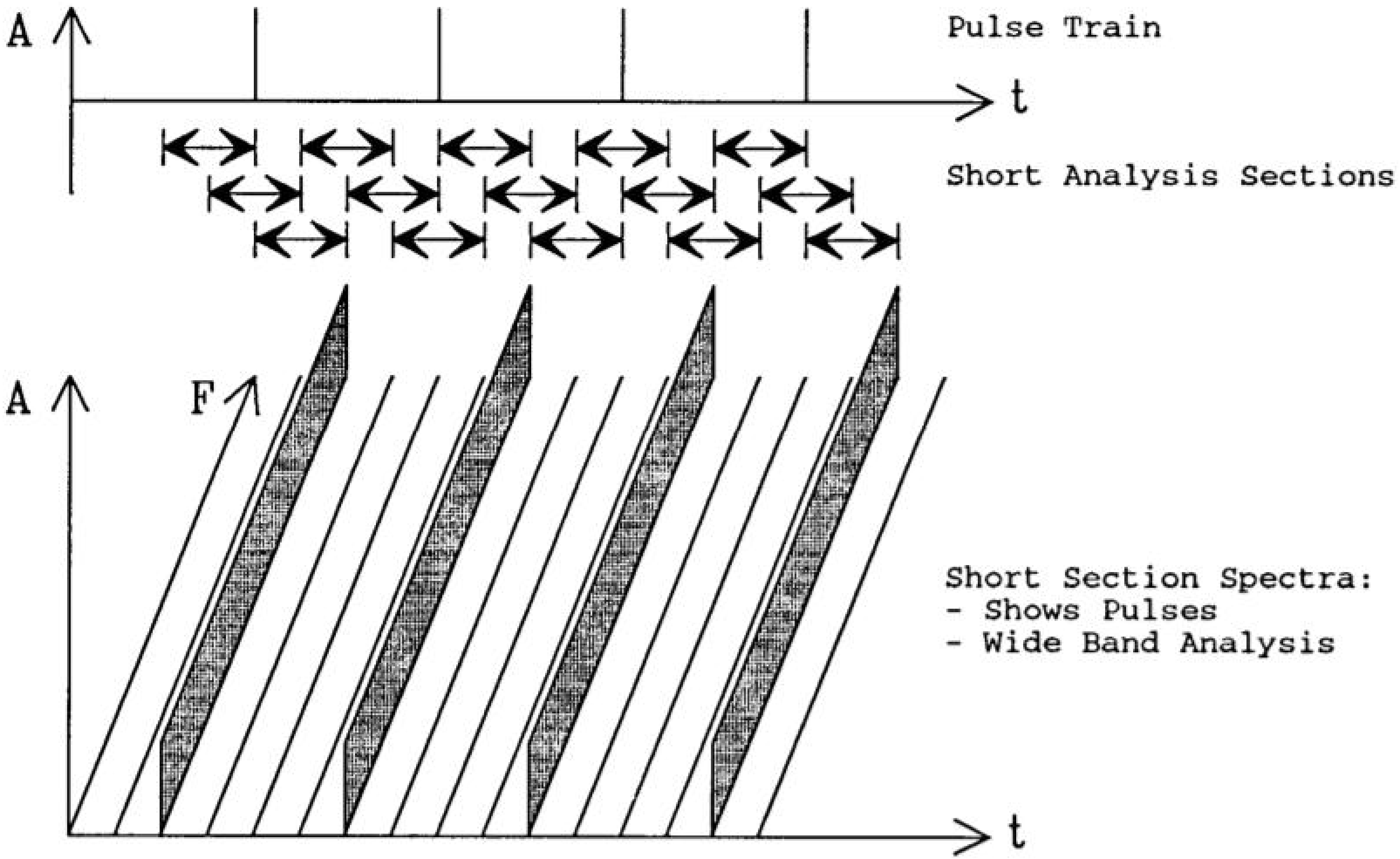


Narrow-band analysis



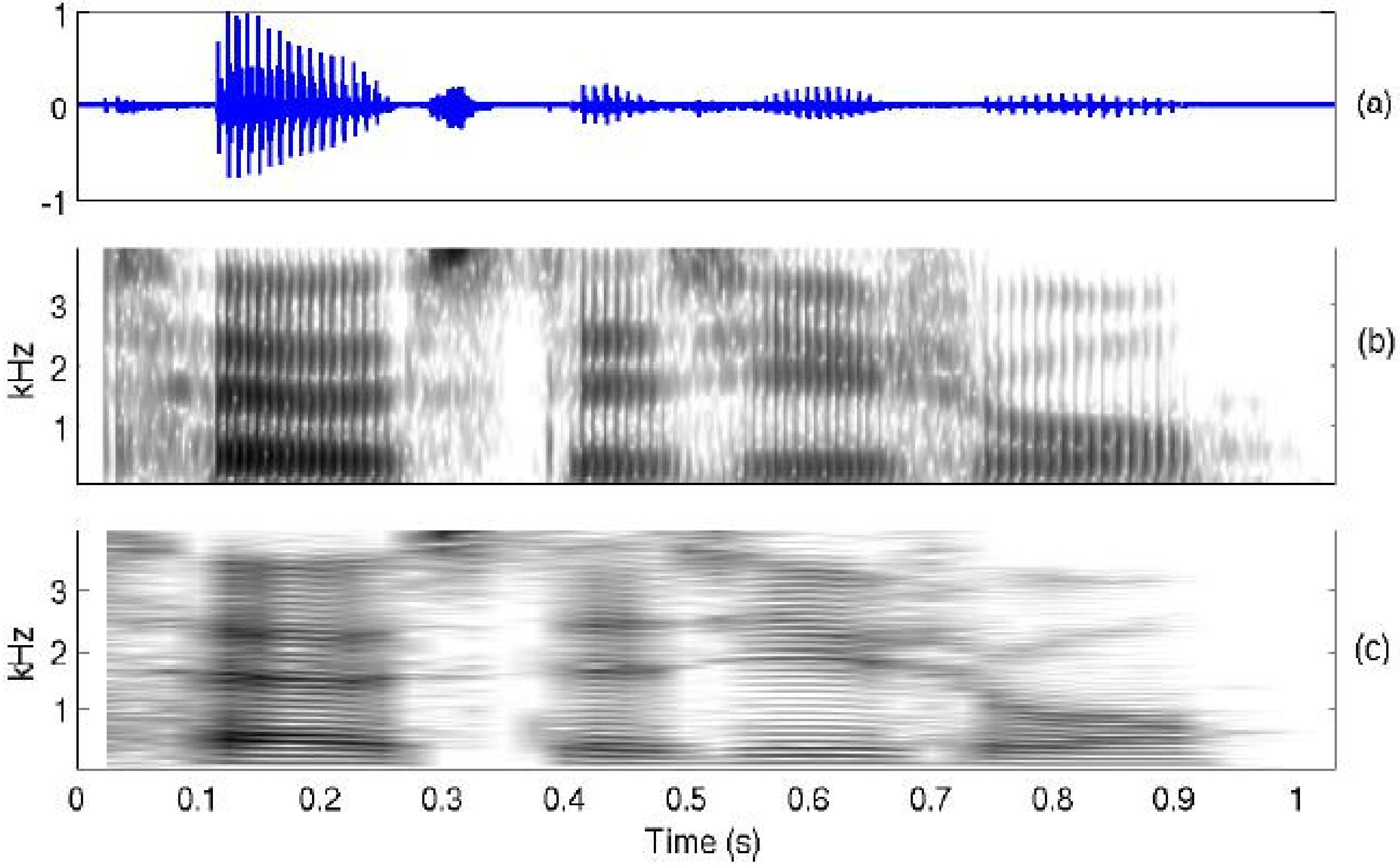
https://www.phon.ucl.ac.uk/courses/spsci/acoustics/week1-10.pdf

Wide-band analysis

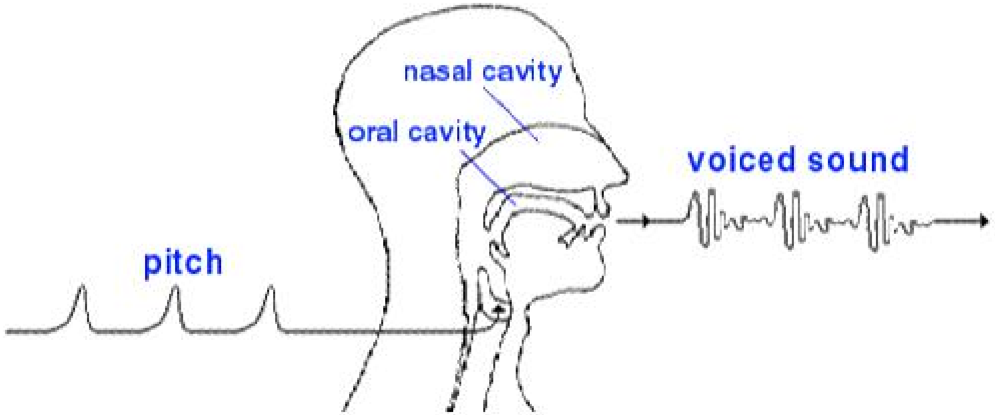


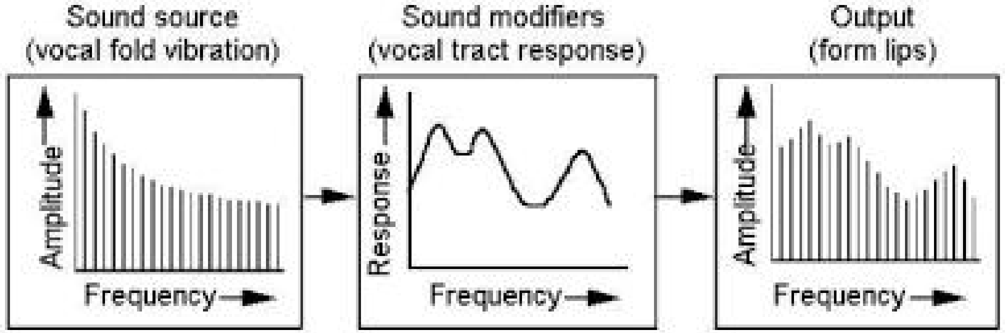
https://www.phon.ucl.ac.uk/courses/spsci/acoustics/week1-10.pdf

Balance between frequency and time



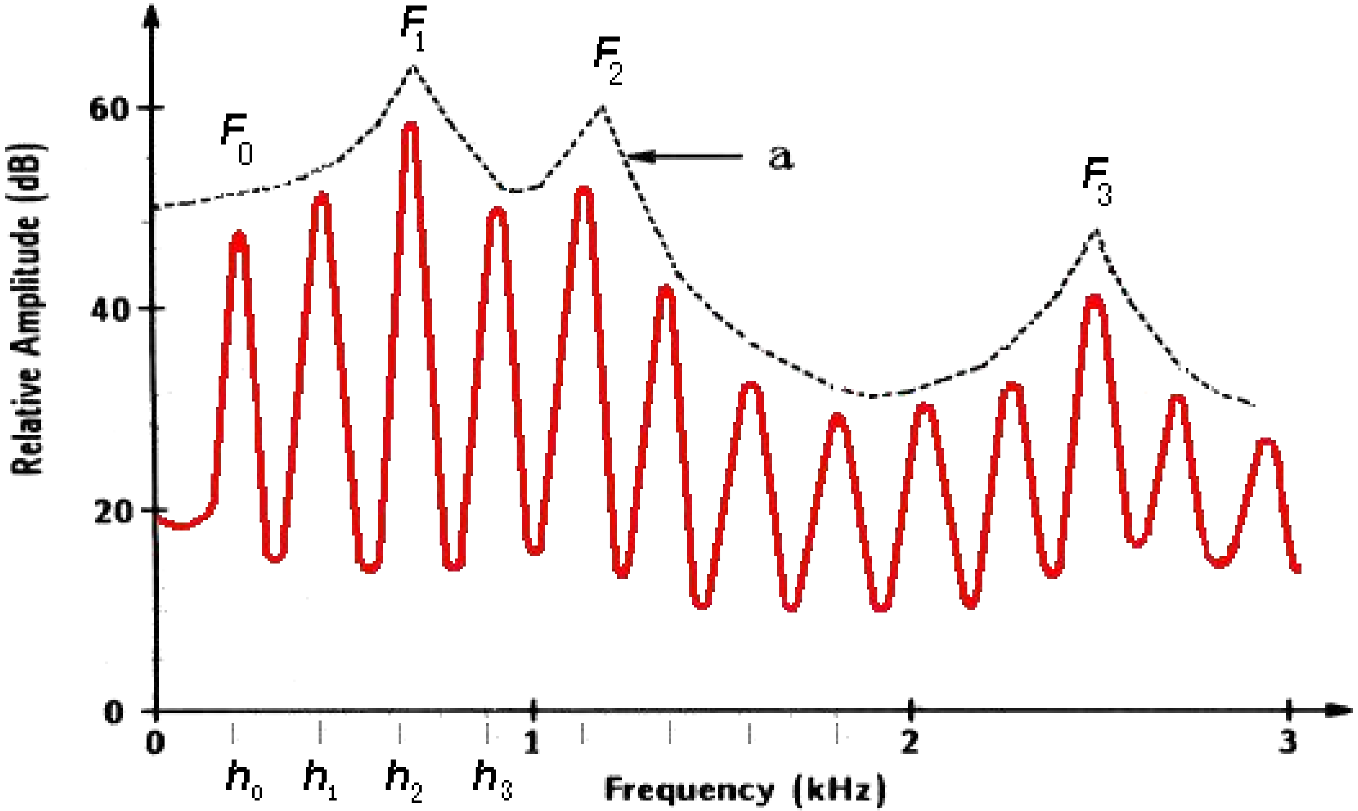
Physiological decomposition





http://147.162.36.50/cochlea/cochleapages/theory/sndproc/sndcomm.htm

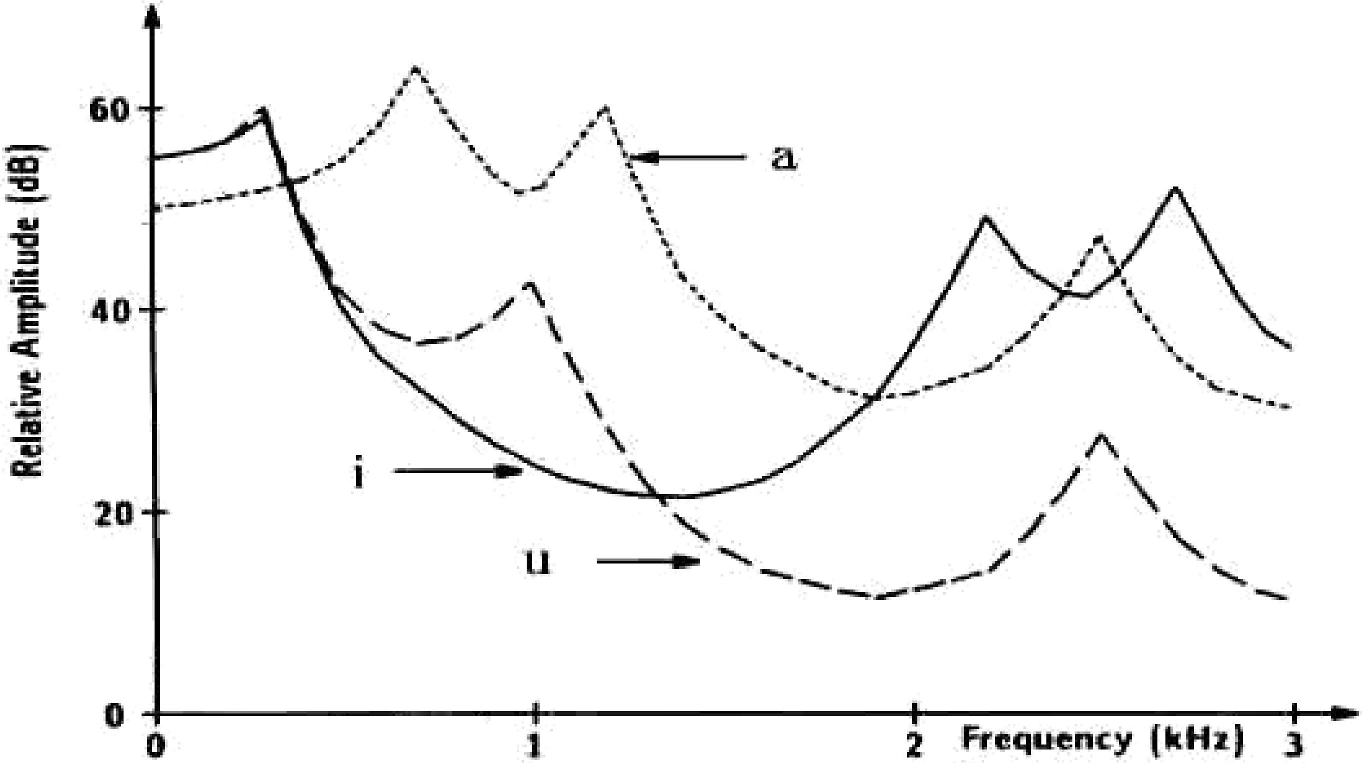
F0, harmonics, and formants



This figure illustrates the spectral structure of the harmonics under the formant profile of vowel ***a***. In the course of a speech utterance, these profiles undergo abrupt changes that can perceived by the auditory apparatus only if appropriately filtered.

http://147.162.36.50/cochlea/cochleapages/theory/sndproc/sndcomm.htm

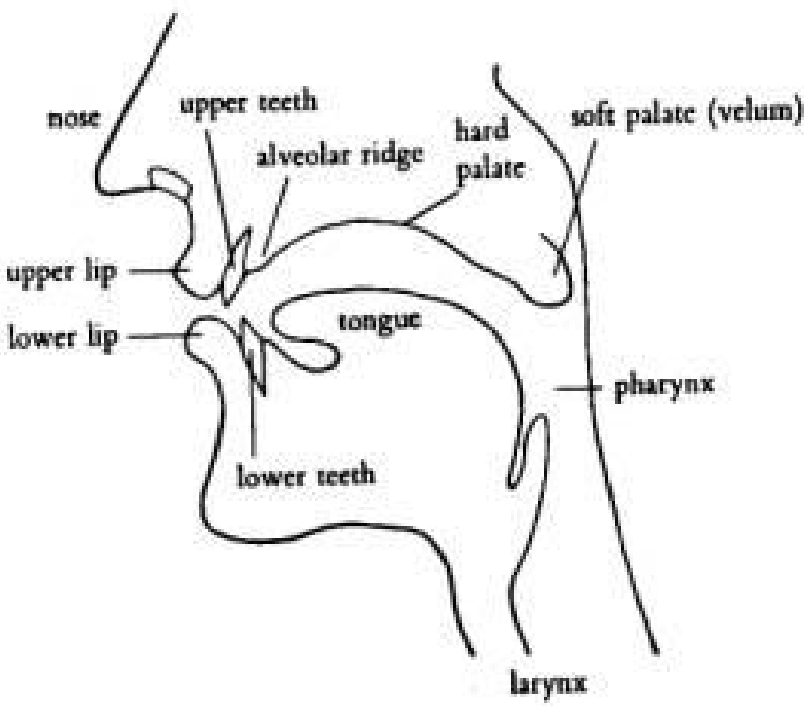
Acoustics and phonetics

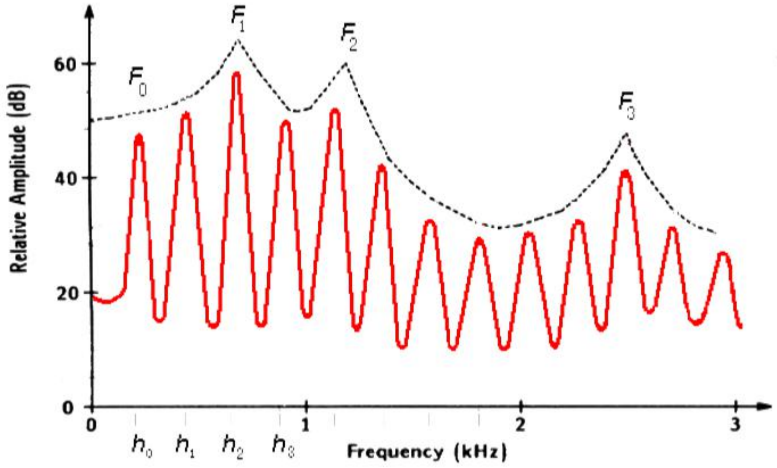


Cross sections of spectra from the middle of English vowels of a male speaker, showing formants as spectral peaks.

From D.O'Shaughnessy (1990) - Speech Communication, Addison-Wesley Pub.Com.

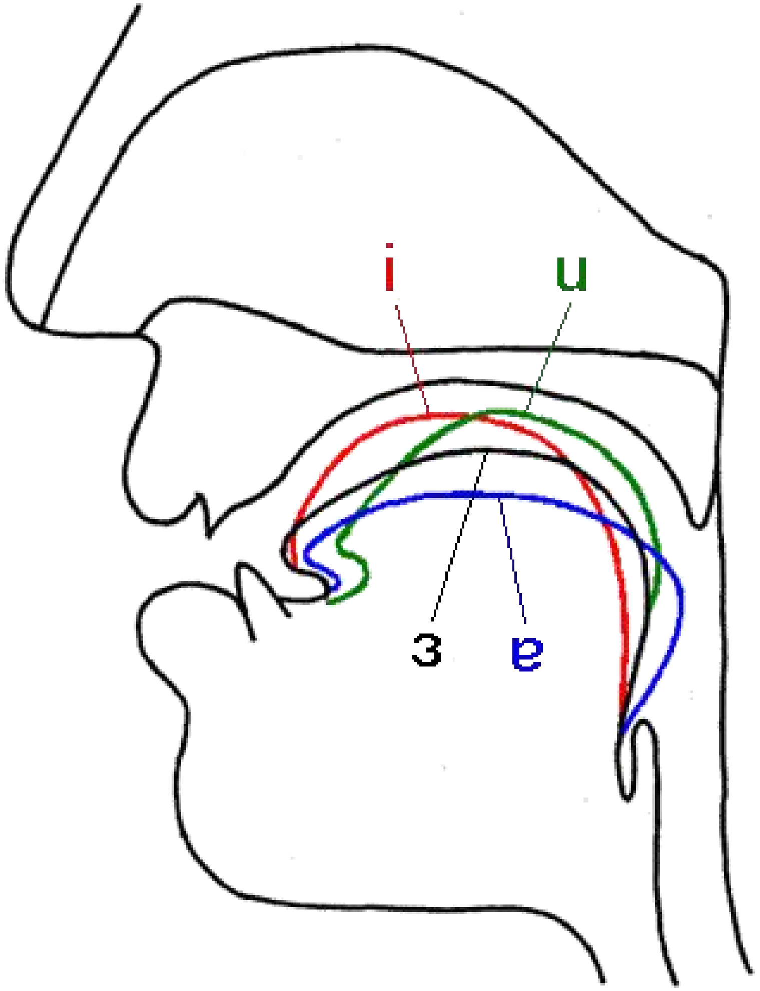
Acoustics and articulatory



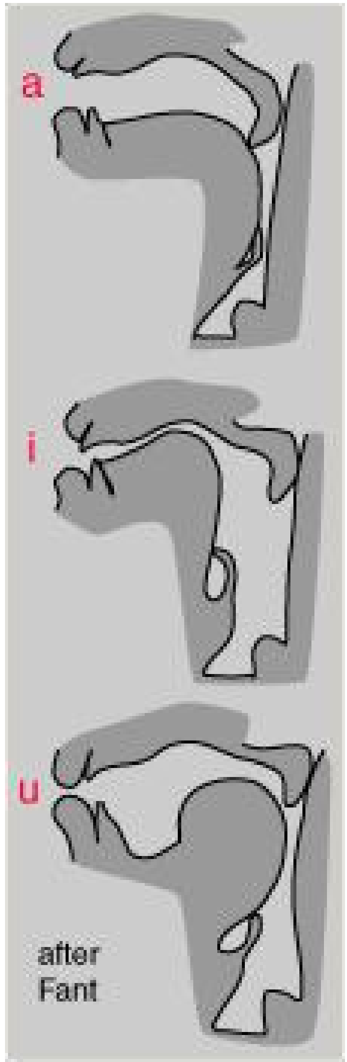


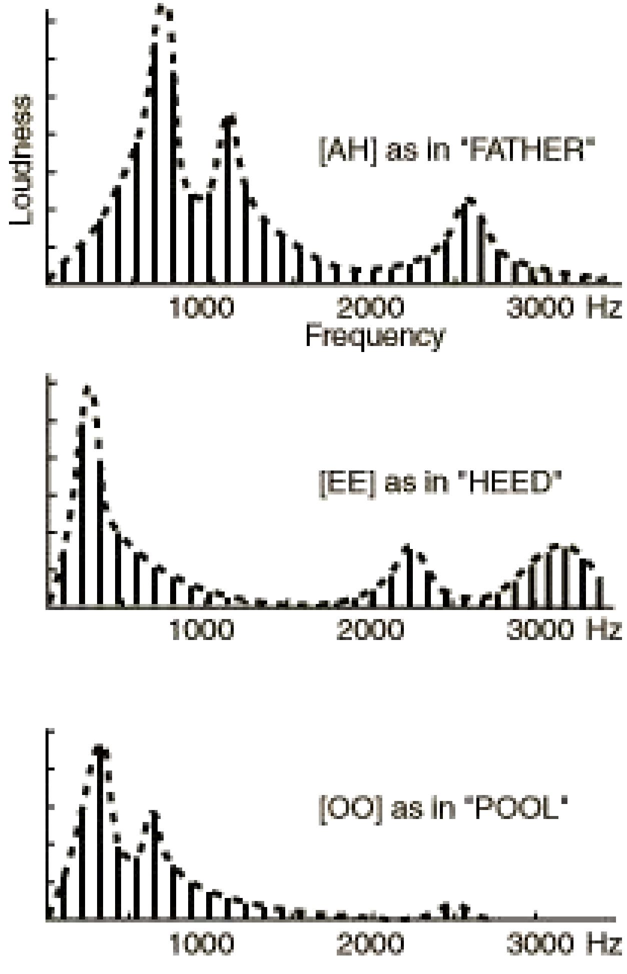
- F1 : pharynx
- F2 : oral cavity
- F3 : nasal cavity (nasal vowels, in french for instance)
- F4 : sinuses (singing formant)

Articulatory and phonetics



Articulatory, acoustics, and phonetics





after Benade

This decomposition is in sufficient

- To make the production model computable, it must be simple, thus cannot handle too much practical situations, e.g., the pseudo periodic source, the nonlinear effect of the vocal tract…
- It does not use labelled data, purely rely on assumption.
- The decomposition is nothing to do with information retrieval tasks, e.g., speech recognition and speaker recognition.
- It resembles to MFCC, which is perfect well designed, but not necessarily the best feature for a particular task.

We hope a factorization model

- Informational: The factors correspond to some desired information, i.e., they should be useful.
- Orthogonal: Flexible enough to derive uncorrelated factors if these factors are truly independent.
- Complete: Information preservation

Shallow factorization

- GMM-UBM: phone and speaker factorization

  $x = m_0^q + m_s^q + e^q$

- JFA: phone, speaker and session factorization

  $x = m_0^q + T_q m_s + V_q n_c + e^q$

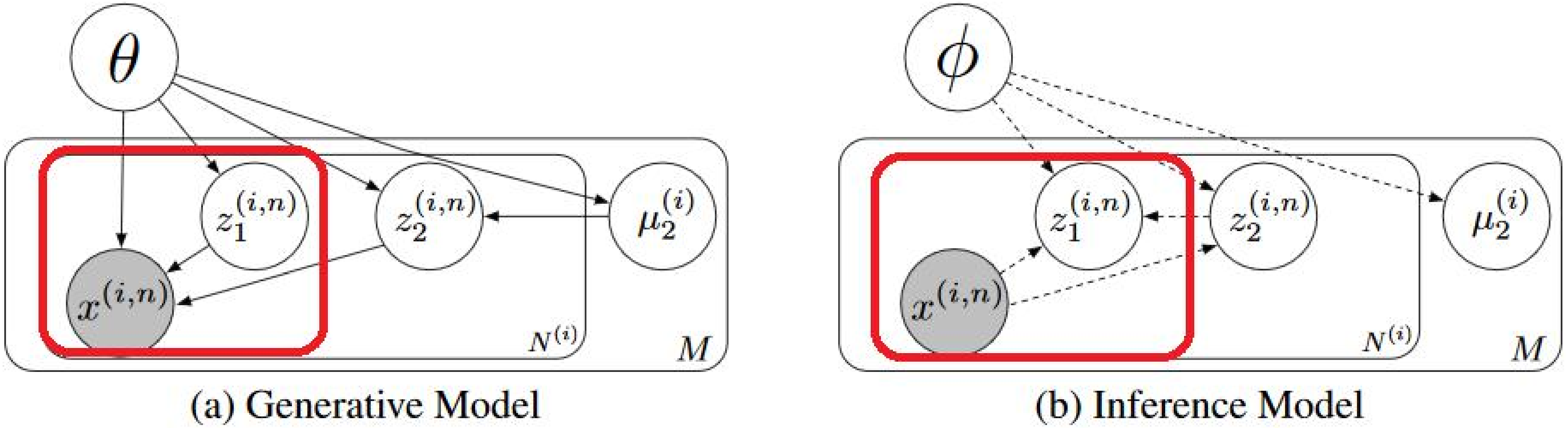
- i-vector: phone and session factorization

  $x = m_0^q + V_q n_c + e^q$

Keynote: short-long bias

Deep factorization

- Place short/long bias



(a) Generative Model

(b) Inference Model

Figure 2: Graphical illustration of the proposed generative model and inference model. Grey nodes denote the observed variables, and white nodes are the hidden variables.

Hsu et al., Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data, NIPS 2017.

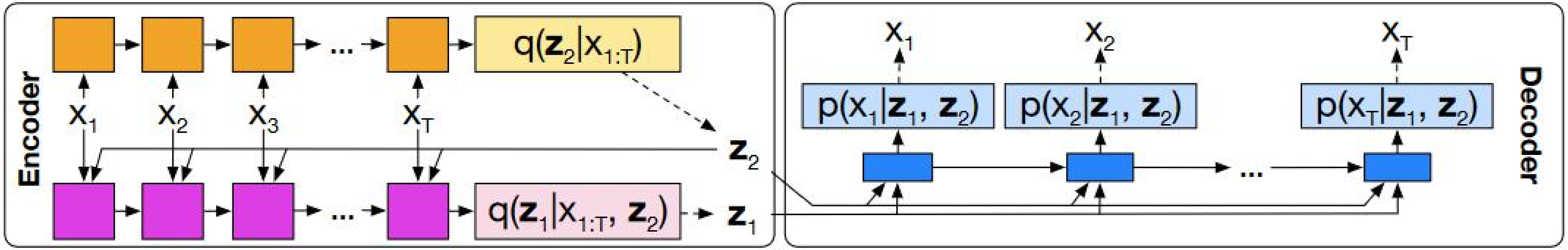


Figure 3: Sequence-to-sequence factorized hierarchical variational autoencoder. Dashed lines indicate the sampling process using the reparameterization trick [23]. The encoders for z_1 and z_2 are pink and amber, respectively, while the decoder for x is blue. Darker colors denote the recurrent neural networks, while lighter colors denote the fully-connected layers predicting the mean and log variance.

Hsu et al., Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data, NIPS 2017.

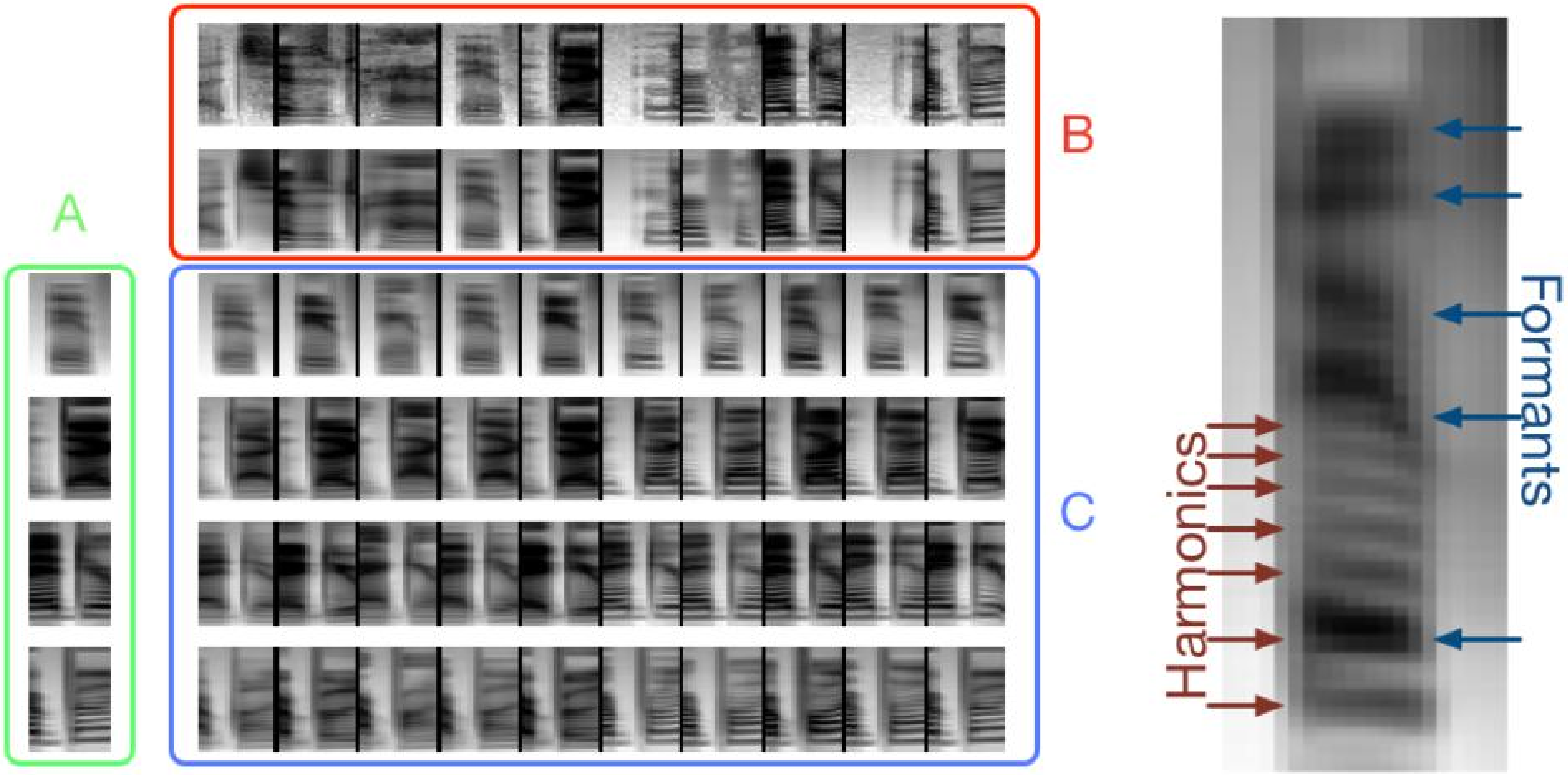
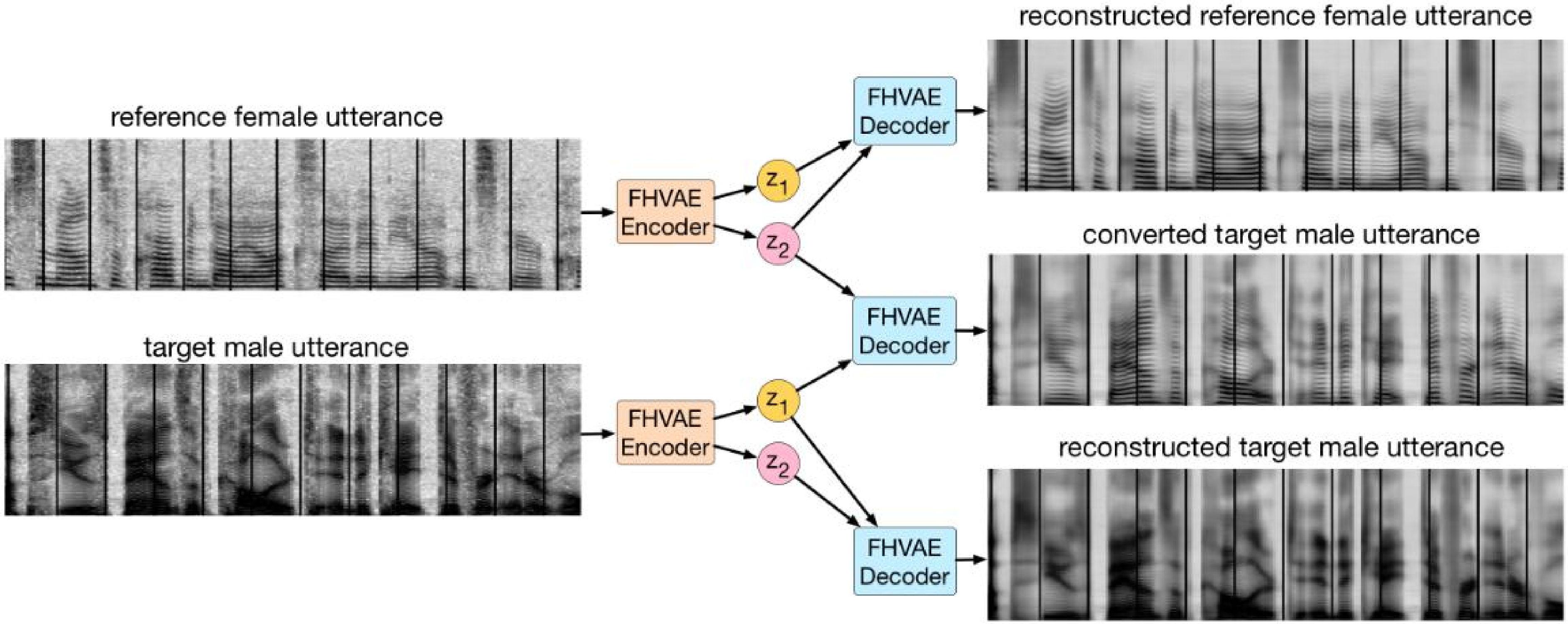


Figure 4: (left) Examples generated by varying different latent variables. (right) An illustration of harmonics and formants in filter bank images. The green block ‘A’ contains four reconstructed examples. The red block ‘B’ contains ten original sequences on the first row with the corresponding reconstructed examples on the second row. The entry on the i-th row and the j-th column in the blue block ‘C’ is the reconstructed example using the latent segment variable z_1 of the i-th row from block ‘A’ and the latent sequence variable z_2 of the j-th column from block ‘B’.

Hsu et al., Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data, NIPS 2017.


reference female utterance
target male utterance
FHVAE Encoder
FHVAE Encoder
z_1
z_2
z_1
z_2
FHVAE Decoder
FHVAE Decoder
FHVAE Decoder
reconstructed reference female utterance
converted target male utterance
reconstructed target male utterance

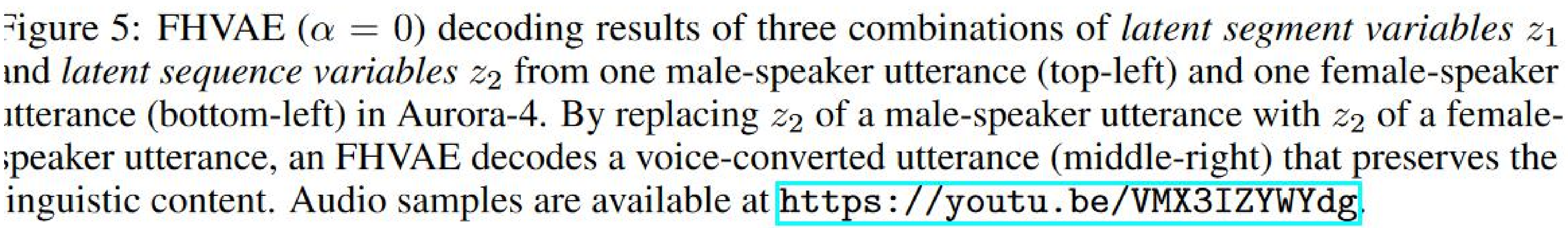
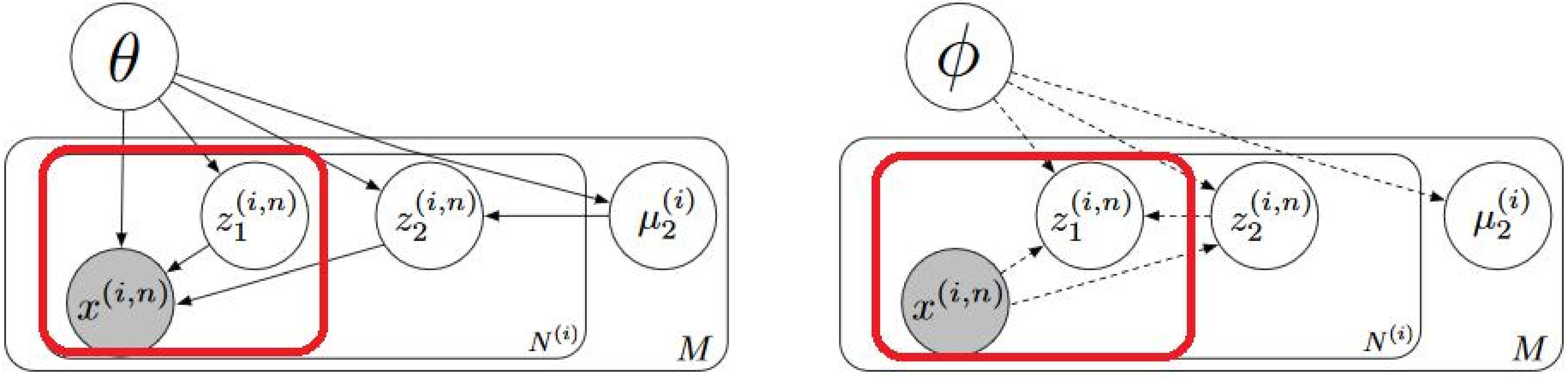


Figure 5: FHVAE ($\alpha = 0$) decoding results of three combinations of *latent segment variables* z_1 and *latent sequence variables* z_2 from one male-speaker utterance (top-left) and one female-speaker utterance (bottom-left) in Aurora-4. By replacing z_2 of a male-speaker utterance with z_2 of a female-speaker utterance, an FHVAE decodes a voice-converted utterance (middle-right) that preserves the linguistic content. Audio samples are available at `https://youtu.be/VMX3IZYWYdg`.

Hsu et al., Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data, NIPS 2017.

Some insights

- The deep factorization is not much different from i-vector. All of them place the factorization in the way of short/long bias, and all of them is based on maximum likelihood.



(a) Generative Model

(b) Inference Model

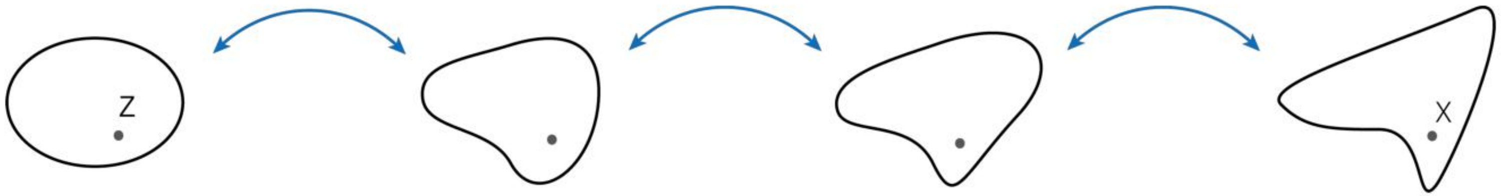
Figure 2: Graphical illustration of the proposed generative model and inference model. Grey nodes denote the observed variables, and white nodes are the hidden variables.

Key shortage

- Our goal is a powerful factorization model, for that goal we should use as much resource as possible, rather than rely on the simple structure bias.
- This bias is not much different from the production model, to some extent.
- The likelihood is not accurate with VAE.

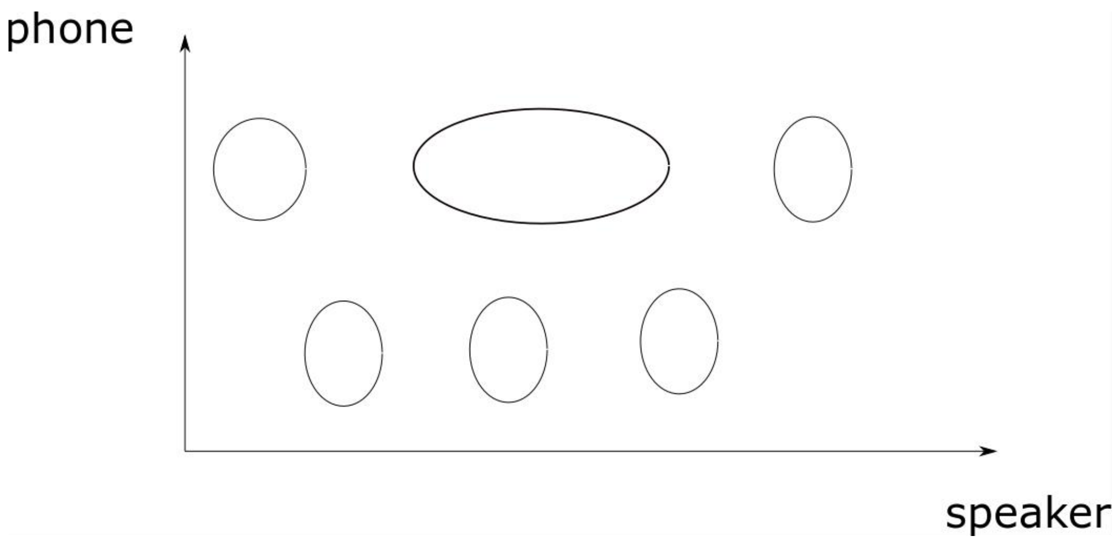
We can use subspace flow to solve the problem

- A purely information preservation model
- Purely supervised learning
- Flexible enough to learn independent factors



Subspace flow

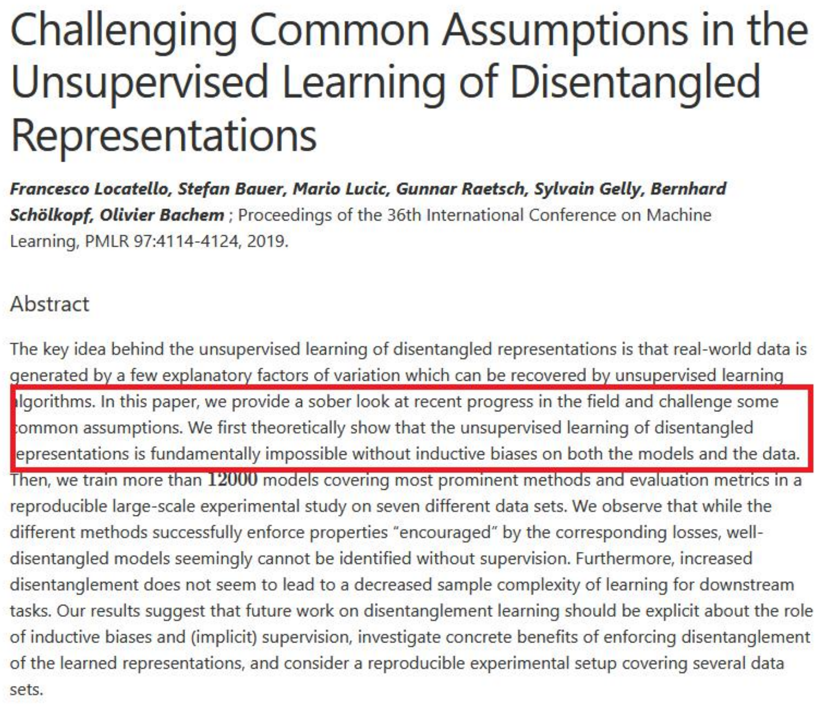
- Divide variation to two groups of dimensions, each one correspond to a factor
- If the label for one factor miss, it returns back to the general NF if the prior is assumed to be Gaussian; otherwise, it is treated as a training example for the GMM model, each component corresponding to a particular value of the discrete factor (e.g., phone).
- For the training data, the two factors can be made independent, but for test set, there would be residual dependency that should be solved.



How to deal with sequential data?

- Essentially, the speaker label has provided more than the sequential data provides.
- However, we didn’t use it for inference. A possible way is to infer the true mean vector, and then compute the likelihood ratio for the testing data, where the prior is set to be the speaker dependent and speaker independent.

Revisit the importance of induction bias



Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations

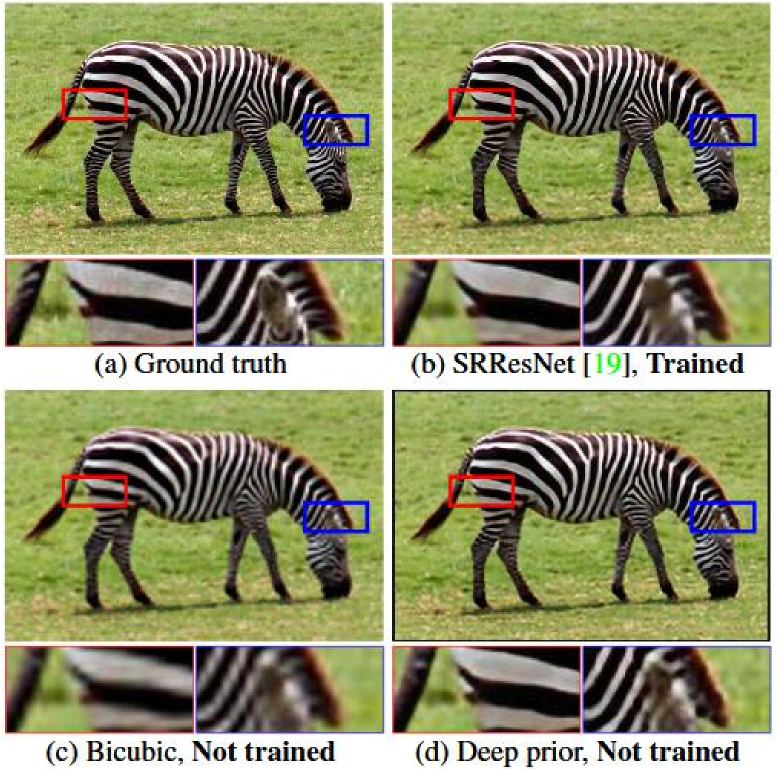
Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, Olivier Bachem ; Proceedings of the 36th International Conference on Machine Learning, PMLR 97:4114-4124, 2019.

Abstract

The key idea behind the unsupervised learning of disentangled representations is that real-world data is generated by a few explanatory factors of variation which can be recovered by unsupervised learning lgorithms. In this paper, we provide a sober look at recent progress in the field and challenge some ommon assumptions. We first theoretically show that the unsupervised learning of disentangled epresentations is fundamentally impossible without inductive biases on both the models and the data. Then, we train more than 12000 models covering most prominent methods and evaluation metrics in a reproducible large-scale experimental study on seven different data sets. We observe that while the different methods successfully enforce properties "encouraged" by the corresponding losses, well-disentangled models seemingly cannot be identified without supervision. Furthermore, increased disentanglement does not seem to lead to a decreased sample complexity of learning for downstream tasks. Our results suggest that future work on disentanglement learning should be explicit about the role of inductive biases and (implicit) supervision, investigate concrete benefits of enforcing disentanglement of the learned representations, and consider a reproducible experimental setup covering several data sets.

- Locatello F, Bauer S, Lucic M, et al. Challenging common assumptions in the unsupervised learning of disentangled representations, ICML 2019.

Deep image prior



(a) Ground truth

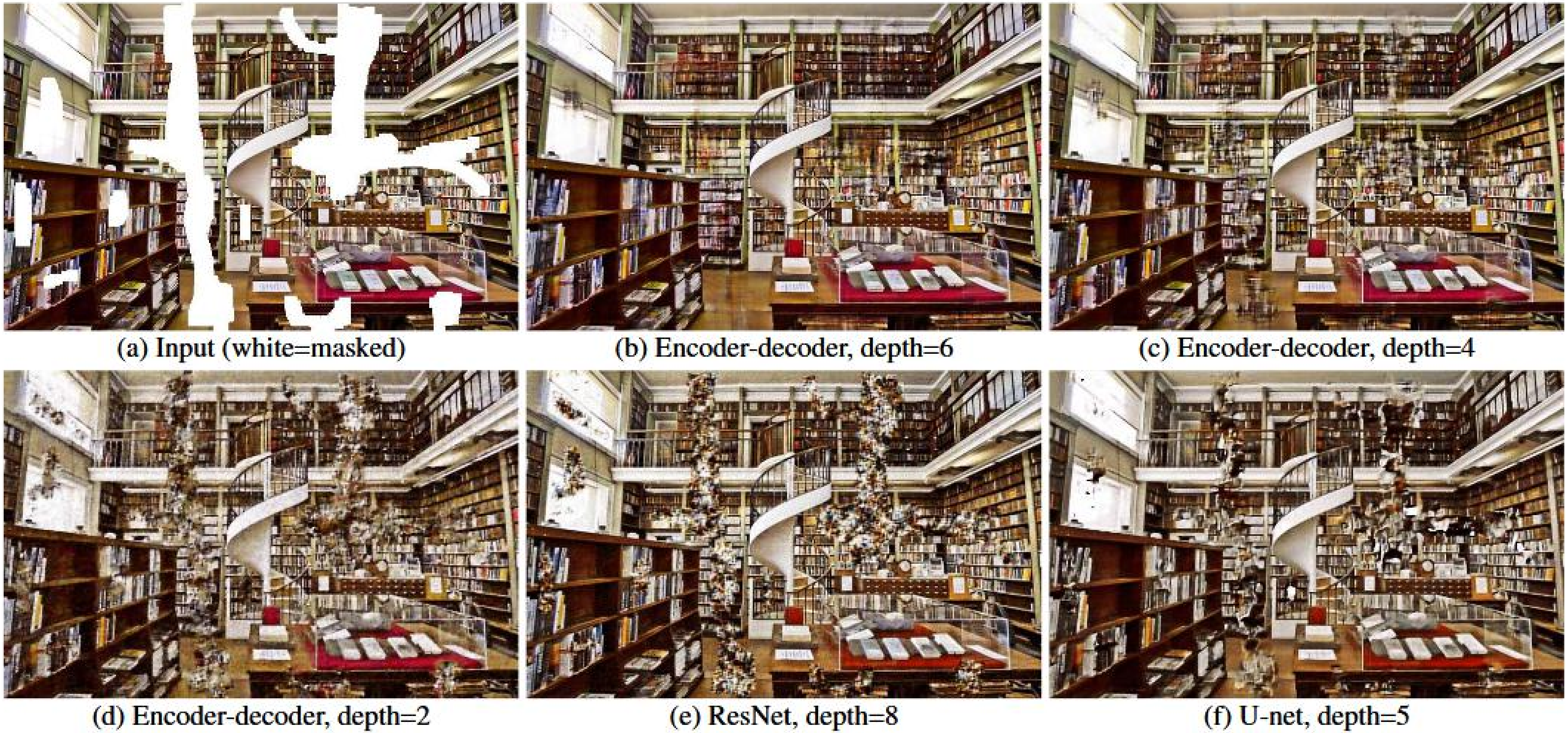
(b) SRResNet [19], **Trained**

(c) Bicubic, **Not trained**

(d) Deep prior, **Not trained**

Figure 1: **Super-resolution using the deep image prior.** Our method uses a randomly-initialized ConvNet to upsample an image, using its structure as an image prior; similar to bicubic upsampling, this method does not require learning, but produces much cleaner results with sharper edges. In fact, our results are quite close to state-of-the-art super-resolution methods that use ConvNets learned from large datasets. The deep image prior works well for all inverse problems we could test.

Ulyanov D, Vedaldi A, Lempitsky V. Deep image prior[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 9446-9454.



(a) Input (white=masked)

(b) Encoder-decoder, depth=6

(c) Encoder-decoder, depth=4

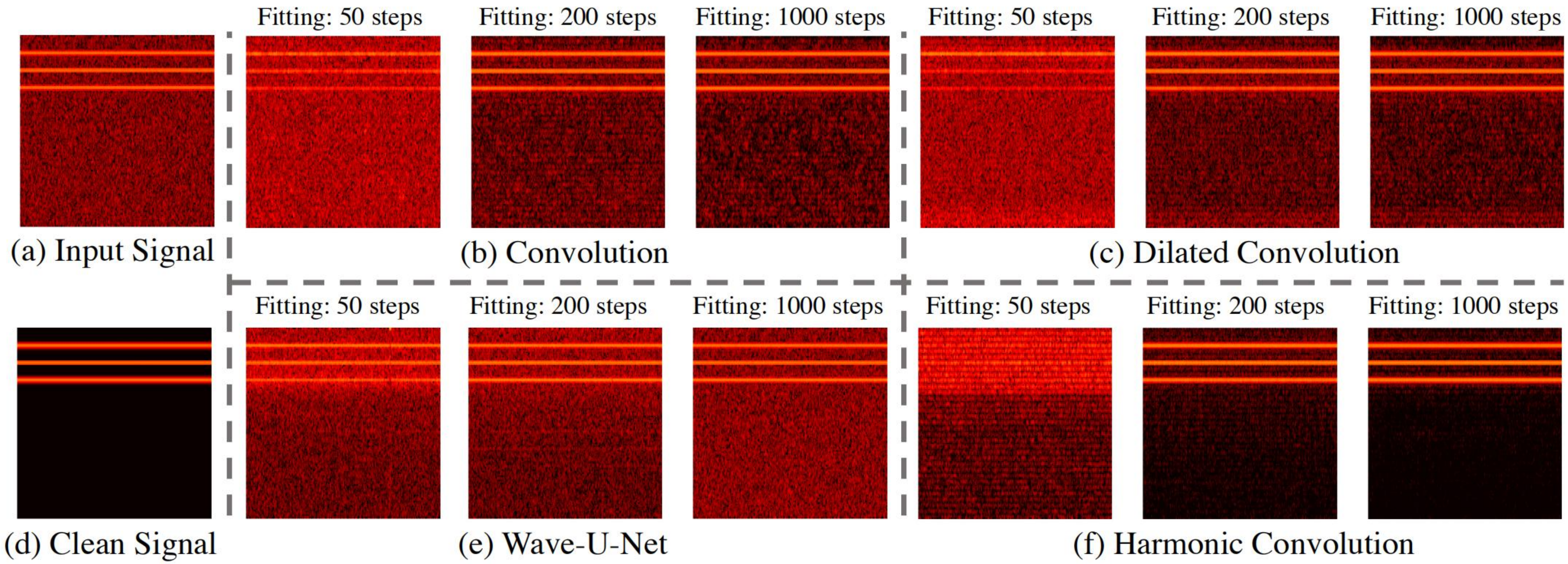
(d) Encoder-decoder, depth=2

(e) ResNet, depth=8

(f) U-net, depth=5

Figure 8: **Inpainting using different depths and architectures.** The figure shows that much better inpainting results can be obtained by using deeper random networks. However, adding skip connections to ResNet in U-Net is highly detrimental.

Deep prior for speech



DEEP AUDIO PRIORS EMERGE FROM HARMONIC CONVOLUTIONAL NETWORKS, ICLR 2020.

More discussion on bias

- Bias is axiom (foundation of the geometry)
- Bias is intuition and belief for models (e.g., production model, do you know it is true?)
- Bias is prior in inference
- Bias is knowledge (low Entropy)
- Bias is belief, intuition, and intention in model design
- No free lunch, no general model, no blind learning
- Bias is good, but it should not be a hindrance, otherwise will be broken.