UNCERTAINTY AND THE MEDICAL INTERVIEW

TOWARDS SELF-ASSESSMENT IN MACHINE LEARNING MODELS

Jakob Drachmann Havtorn

DTU Compute
Department of Applied Mathematics and Computer Science

UNCERTAINTY AND THE MEDICAL INTERVIEW
TOWARDS SELF-ASSESSMENT IN MACHINE LEARNING MODELS
Jakob Drachmann Havtorn

- Welcome to my PhD defense.
- Thank you to the moderator and the assessment committee for taking part today.
- I will present my work on uncertainty estimation in AI systems for medical domains.
- I will start with an overview of the thesis followed by a brief introduction.
- Then I will present a selection of the research chapters.
- Finally, I will discuss the broader implications of the work.

Thesis

1. The thesis is structured into 10 chapters.
2. The first three chapters are introductory.
3. The next six chapters are research chapters.
4. The final chapter is a discussion and conclusion.

OVERVIEW
Thesis

corti DTU

1. The thesis is structured into 10 chapters.
2. The first three chapters are introductory.
3. The next six chapters are research chapters.
4. The final chapter is a discussion and conclusion.

Background

Industrial PhD project with Corti AI and DTU Compute.

- **2020-2023**
- **Collaboration** between academia and industry partially funded by InnovationFund Denmark.
- **Corti**: Using machine learning to augment communication in the healthcare sector.
- **Project goal**: Pursue research in machine learning at the interface between academic and company interests.

Background

Industrial PhD project with Corti AI and DTU Compute.

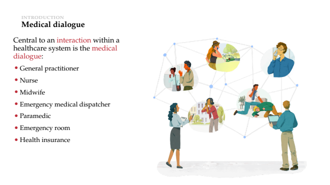
- 2020-2023
- Collaboration between academia and industry partially funded by InnovationFund Denmark.
- Corti: Using machine learning to augment communication in the healthcare sector.
- Project goal: Pursue research in machine learning at the interface between academic and company interests.

1. Industrial project, Corti + DTU Compute, InnovationFund Denmark, 2020-2023.
2. Augment communication in healthcare sector.
3. Research in machine learning at the interface of shared interests.

Medical dialogue

Central to an interaction within a healthcare system is the medical dialogue:

- General practitioner
- Nurse
- Midwife
- Emergency medical dispatcher
- Paramedic
- Emergency room
- Health insurance



1. We focus on medical communication.
2. Involves many different parties.
3. Different contexts and purposes.
4. Busy emergency room, calls to emergency medical dispatchers, visits at general practitioners, etc.

INTRODUCTION

Errors in medical dialogue

- Communication is everywhere in healthcare.
- It is complex, involving multiple participants, different contexts, and different purposes.

1. So yes communication can be complex and noisy, but what are the consequences?
2. Failure of communication is a leading cause of adverse events.
3. Adverse events are episodes of medical error that result in harm to the patient.
4. Luckily, many of these are preventable, although exact numbers vary.
5. Better communication could help reduce these numbers.

INTRODUCTION

Errors in medical dialogue

- Communication is everywhere in healthcare.
- It is complex, involving multiple participants, different contexts, and different purposes.

- Adverse events: Failure of communication contributes to two out of three adverse events [48].
- Preventability: Many adverse outcomes are preventable [10].

Errors in medical dialogue
- Communication is everywhere in healthcare.
- It is complex, involving multiple participants, different contexts, and different purposes.
- Adverse events: Failure of communication contributes to two out of three adverse events [48].
- Preventability: Many adverse outcomes are preventable [10].

1. So yes communication can be complex and noisy, but what are the consequences?
2. Failure of communication is a leading cause of adverse events.
3. Adverse events are episodes of medical error that result in harm to the patient.
4. Luckily, many of these are preventable, although exact numbers vary.
5. Better communication could help reduce these numbers.

Documenting medical encounters

- Documentation is a central part of healthcare.
- E.g. patient records, insurance claims, billing, research, training, legal purposes.


INTRODUCTION
Documenting medical encounters
• Documentation is a central part of healthcare.
• E.g. patient records, insurance claims, billing, research, training, legal purposes.

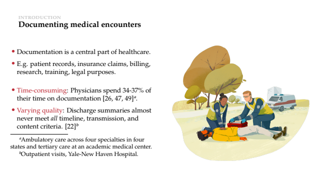
1. Another central aspect of medical communication is documentation.
2. Essential for a number of purposes
3. But, it is time-consuming and of varying quality.
4. (Ambulatory ≡ outpatient care, Tertiary ≡ specialized care)

Documenting medical encounters

- Documentation is a central part of healthcare.
- E.g. patient records, insurance claims, billing, research, training, legal purposes.

- Time-consuming: Physicians spend 34-37% of their time on documentation [26, 47, 49][a].
- Varying quality: Discharge summaries almost never meet *all* timeline, transmission, and content criteria. [22][b]

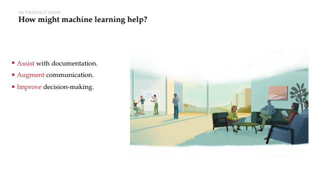
[a]Ambulatory care across four specialties in four states and tertiary care at an academic medical center.
[b]Outpatient visits, Yale-New Haven Hospital.



1. Another central aspect of medical communication is documentation.
2. Essential for a number of purposes
3. But, it is time-consuming and of varying quality.
4. (Ambulatory ≡ outpatient care, Tertiary ≡ specialized care)

How might machine learning help?

- Assist with documentation.
- Augment communication.
- Improve decision-making.

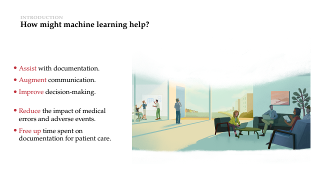


1. Machine learning can help in many ways.
2. The main goal of augmenting communication is:
   - Reduce the impact.
   - Free up time.

How might machine learning help?

- Assist with documentation.
- Augment communication.
- Improve decision-making.

- Reduce the impact of medical errors and adverse events.
- Free up time spent on documentation for patient care.



1. Machine learning can help in many ways.
2. The main goal of augmenting communication is:
   - Reduce the impact.
   - Free up time.

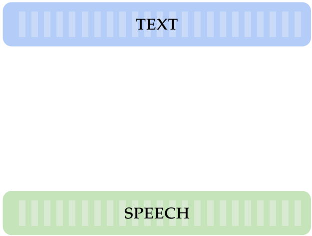
Building a decision-support system

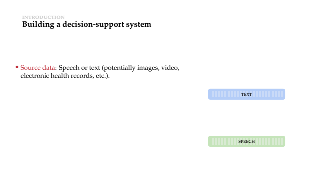
Building a decision-support system

1. We will take a modular approach to building a decision-support system.
2. First we need source data
3. Then we need to convert it into representations useful for downstream tasks.
4. Then we can perform the downstream tasks.
5. Finally, we need to estimate the reliability of our data, representations, and predictions.

Building a decision-support system

- Source data: Speech or text (potentially images, video, electronic health records, etc.).

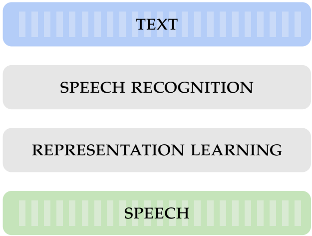


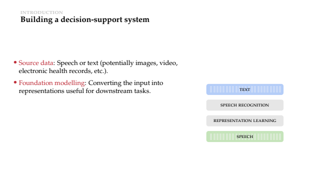


1. We will take a modular approach to building a decision-support system.
2. First we need source data
3. Then we need to convert it into representations useful for downstream tasks.
4. Then we can perform the downstream tasks.
5. Finally, we need to estimate the reliability of our data, representations, and predictions.

Building a decision-support system

- Source data: Speech or text (potentially images, video, electronic health records, etc.).
- Foundation modelling: Converting the input into representations useful for downstream tasks.

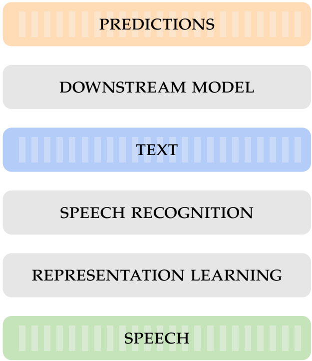


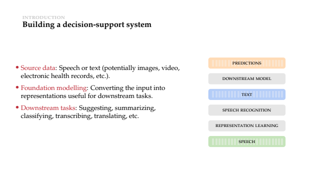


1. We will take a modular approach to building a decision-support system.
2. First we need source data
3. Then we need to convert it into representations useful for downstream tasks.
4. Then we can perform the downstream tasks.
5. Finally, we need to estimate the reliability of our data, representations, and predictions.

Building a decision-support system

- Source data: Speech or text (potentially images, video, electronic health records, etc.).
- Foundation modelling: Converting the input into representations useful for downstream tasks.
- Downstream tasks: Suggesting, summarizing, classifying, transcribing, translating, etc.

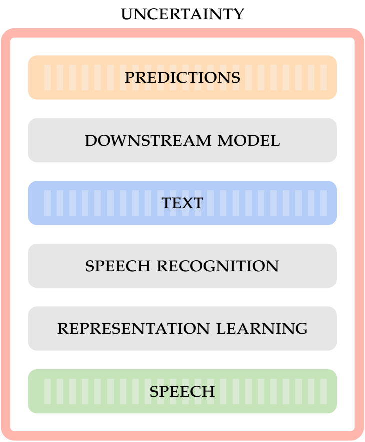


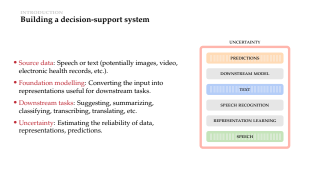


1. We will take a modular approach to building a decision-support system.
2. First we need source data
3. Then we need to convert it into representations useful for downstream tasks.
4. Then we can perform the downstream tasks.
5. Finally, we need to estimate the reliability of our data, representations, and predictions.

Building a decision-support system

- Source data: Speech or text (potentially images, video, electronic health records, etc.).
- Foundation modelling: Converting the input into representations useful for downstream tasks.
- Downstream tasks: Suggesting, summarizing, classifying, transcribing, translating, etc.
- Uncertainty: Estimating the reliability of data, representations, predictions.

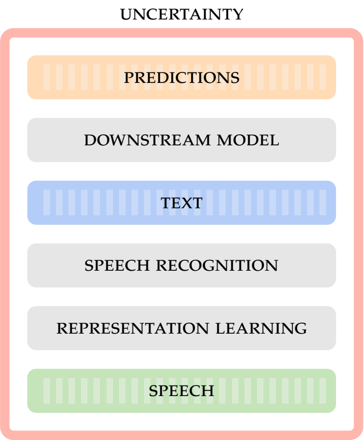


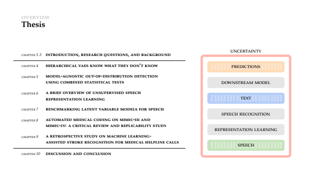


1. We will take a modular approach to building a decision-support system.
2. First we need source data
3. Then we need to convert it into representations useful for downstream tasks.
4. Then we can perform the downstream tasks.
5. Finally, we need to estimate the reliability of our data, representations, and predictions.

OVERVIEW

Thesis





Chp. 4 Examine how a certain class of generative models can be used to detect data that is different from training data.

Chp. 5 Generalize this to be model-agnostic and phrased as a statistical test.

Chp. 6 Step back and provide an overview of modern methods for unsupervised speech representation learning.

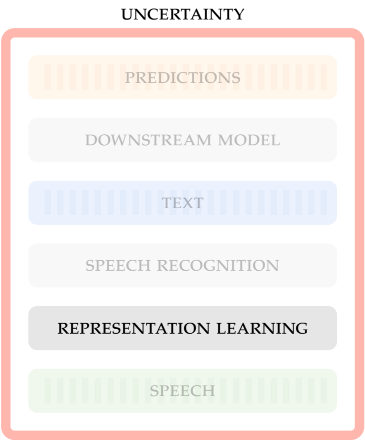
Chp. 7 Benchmark some of these methods, generative latent variable models, on likelihood and automatic speech recognition.

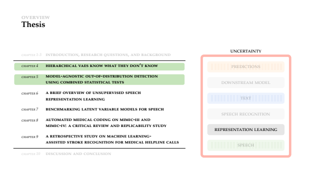
Chp. 8 Move to a different domain, medical coding, and provide a critical review of recent works.

Chp. 9 Provide a retrospective study on a machine learning-assisted stroke recognition system for medical helpline calls.

1. So how does the thesis tackle this problem of building such a system?

Thesis





Chp. 4 Examine how a certain class of generative models can be used to detect data that is different from training data.

Chp. 5 Generalize this to be model-agnostic and phrased as a statistical test.

Chp. 6 Step back and provide an overview of modern methods for unsupervised speech representation learning.

Chp. 7 Benchmark some of these methods, generative latent variable models, on likelihood and automatic speech recognition.

Chp. 8 Move to a different domain, medical coding, and provide a critical review of recent works.

Chp. 9 Provide a retrospective study on a machine learning-assisted stroke recognition system for medical helpline calls.

1. So how does the thesis tackle this problem of building such a system?

OVERVIEW

Thesis

CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

CHAPTER 5 MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS

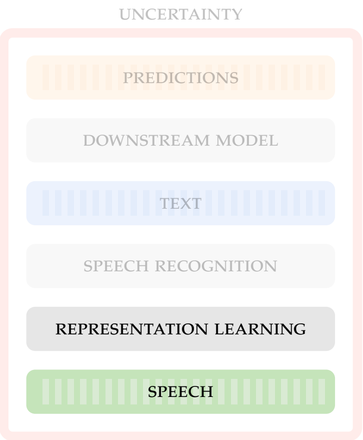
CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

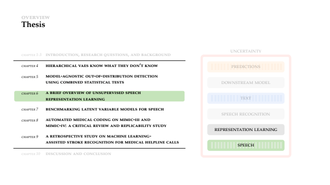
CHAPTER 7 BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH

CHAPTER 8 AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY

CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

CHAPTER 10 DISCUSSION AND CONCLUSION





Chp. 4 Examine how a certain class of generative models can be used to detect data that is different from training data.

Chp. 5 Generalize this to be model-agnostic and phrased as a statistical test.

Chp. 6 Step back and provide an overview of modern methods for unsupervised speech representation learning.

Chp. 7 Benchmark some of these methods, generative latent variable models, on likelihood and automatic speech recognition.

Chp. 8 Move to a different domain, medical coding, and provide a critical review of recent works.

Chp. 9 Provide a retrospective study on a machine learning-assisted stroke recognition system for medical helpline calls.

1. So how does the thesis tackle this problem of building such a system?

Thesis

CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

CHAPTER 5 MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS

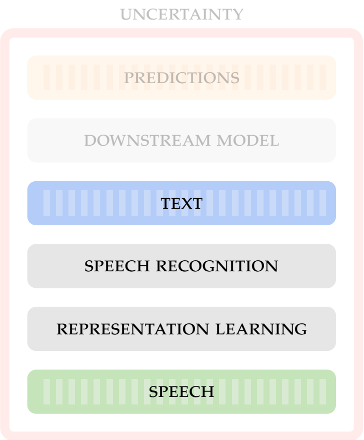
CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

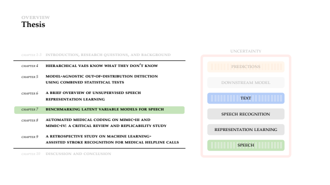
CHAPTER 7 BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH

CHAPTER 8 AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY

CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

CHAPTER 10 DISCUSSION AND CONCLUSION





Chp. 4 Examine how a certain class of generative models can be used to detect data that is different from training data.

Chp. 5 Generalize this to be model-agnostic and phrased as a statistical test.

Chp. 6 Step back and provide an overview of modern methods for unsupervised speech representation learning.

Chp. 7 Benchmark some of these methods, generative latent variable models, on likelihood and automatic speech recognition.

Chp. 8 Move to a different domain, medical coding, and provide a critical review of recent works.

Chp. 9 Provide a retrospective study on a machine learning-assisted stroke recognition system for medical helpline calls.

1. So how does the thesis tackle this problem of building such a system?

OVERVIEW
Thesis

CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

CHAPTER 5 MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS

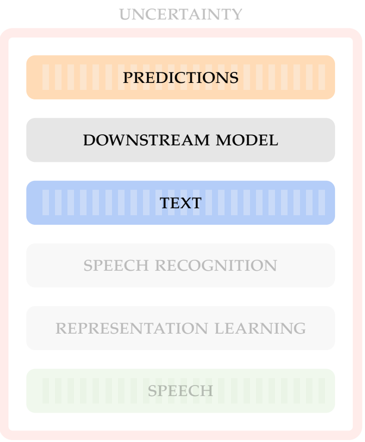
CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

CHAPTER 7 BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH

CHAPTER 8 AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY

CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

CHAPTER 10 DISCUSSION AND CONCLUSION



Chp. 4 Examine how a certain class of generative models can be used to detect data that is different from training data.

Chp. 5 Generalize this to be model-agnostic and phrased as a statistical test.

Chp. 6 Step back and provide an overview of modern methods for unsupervised speech representation learning.

Chp. 7 Benchmark some of these methods, generative latent variable models, on likelihood and automatic speech recognition.

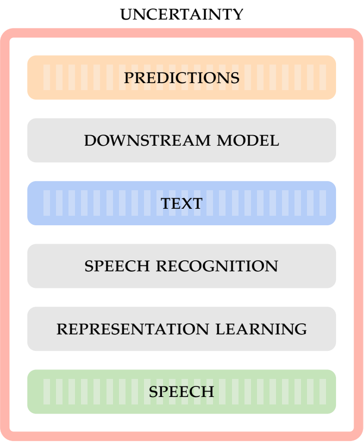
Chp. 8 Move to a different domain, medical coding, and provide a critical review of recent works.

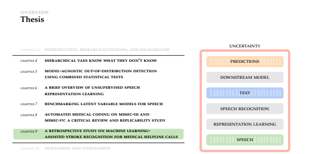
Chp. 9 Provide a retrospective study on a machine learning-assisted stroke recognition system for medical helpline calls.

1. So how does the thesis tackle this problem of building such a system?

Thesis

CHAPTER 1-3	INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND
CHAPTER 4	HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW
CHAPTER 5	MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS
CHAPTER 6	A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING
CHAPTER 7	BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH
CHAPTER 8	AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY
CHAPTER 9	A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS
CHAPTER 10	DISCUSSION AND CONCLUSION





Chp. 4 Examine how a certain class of generative models can be used to detect data that is different from training data.

Chp. 5 Generalize this to be model-agnostic and phrased as a statistical test.

Chp. 6 Step back and provide an overview of modern methods for unsupervised speech representation learning.

Chp. 7 Benchmark some of these methods, generative latent variable models, on likelihood and automatic speech recognition.

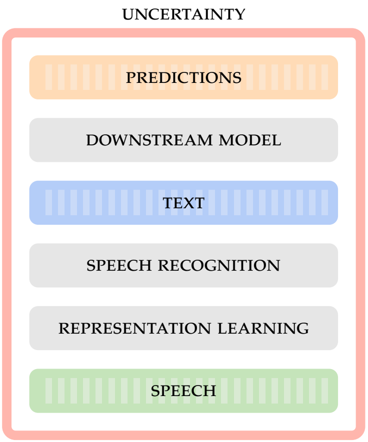
Chp. 8 Move to a different domain, medical coding, and provide a critical review of recent works.

Chp. 9 Provide a retrospective study on a machine learning-assisted stroke recognition system for medical helpline calls.

1. So how does the thesis tackle this problem of building such a system?

Presentation

CHAPTER 1-3	**INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND**
CHAPTER 4	**HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW**
CHAPTER 5	**MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS**
CHAPTER 6	**A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING**
CHAPTER 7	**BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH**
CHAPTER 8	**AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY**
CHAPTER 9	**A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS**
CHAPTER 10	**DISCUSSION AND CONCLUSION**

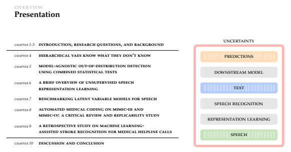


2024-03-04

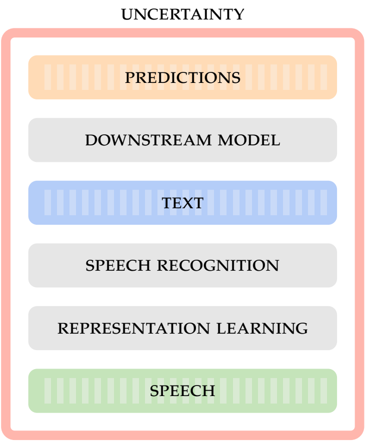
UNCERTAINTY AND THE MEDICAL INTERVIEW

└overview

└Presentation

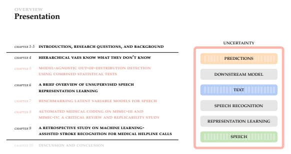


Presentation

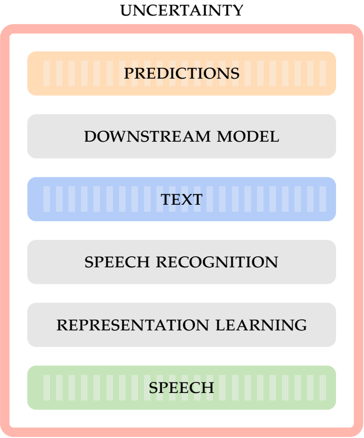


2024-03-04

UNCERTAINTY AND THE MEDICAL INTERVIEW
└overview
└Presentation

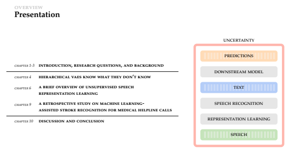


Presentation

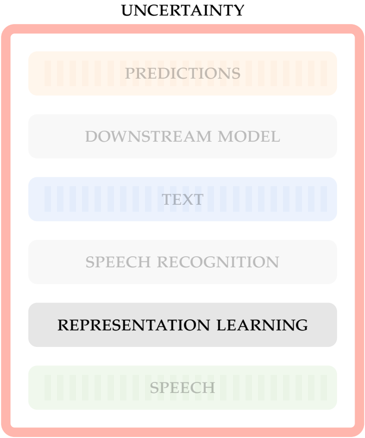


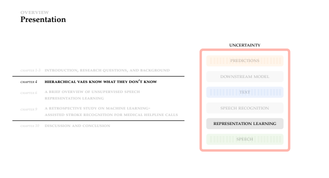
UNCERTAINTY AND THE MEDICAL INTERVIEW
└overview
└Presentation

2024-03-04



Presentation



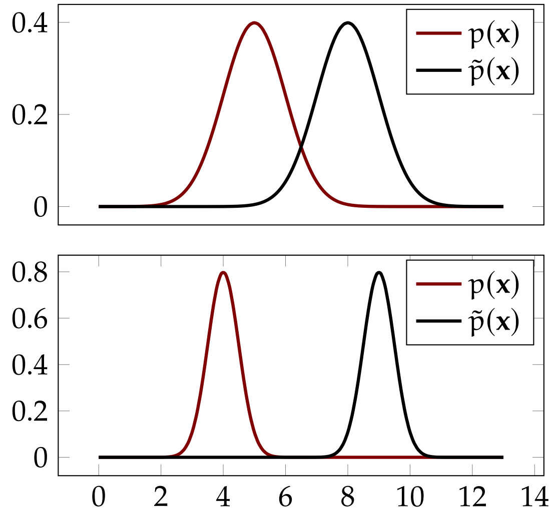


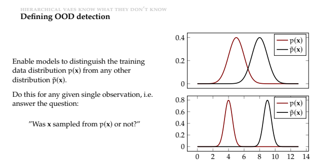
Defining OOD detection

Enable models to distinguish the training data distribution $p(\mathbf{x})$ from any other distribution $\tilde{p}(\mathbf{x})$.

Do this for any given single observation, i.e. answer the question:

"Was $\mathbf{x}$ sampled from $p(\mathbf{x})$ or not?"

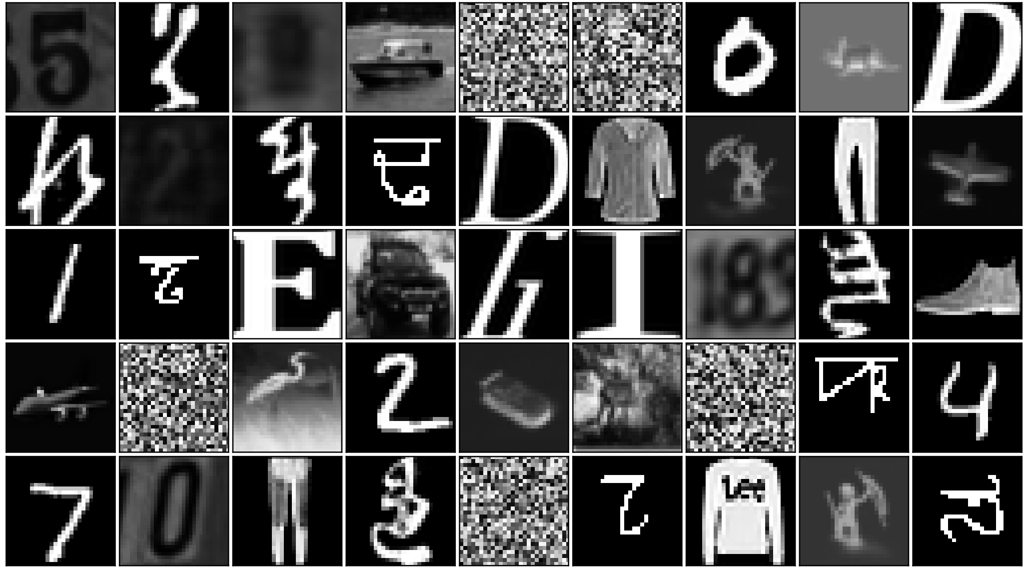




1. Define OOD
2. In 1D it is easy to see when this is possible and when it is not.
3. In higher dimensions it is not as easy to see.

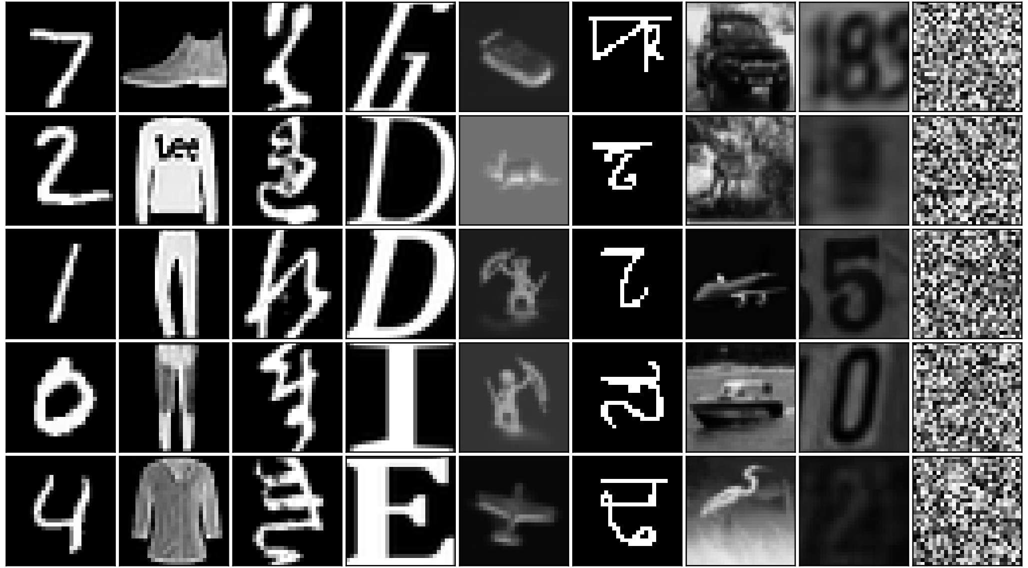
In distribution?

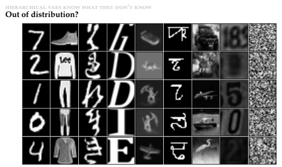
corti DTU



1. Datasets can overlap quite a bit in their raw data space.
2. What we usually care about is a more semantic notion of similarity.

Out of distribution?

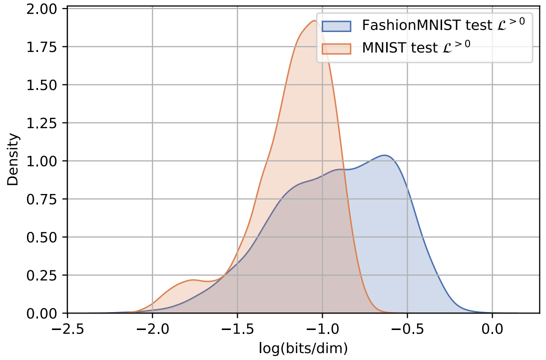




1. Datasets can overlap quite a bit in their raw data space.
2. What we usually care about is a more semantic notion of similarity.

Out-of-distribution detection with generative models

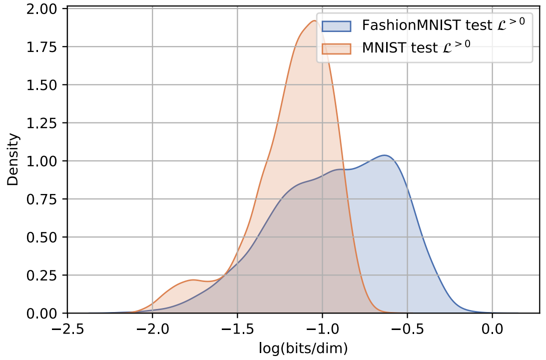
- Generative models learn to approximate the **data distribution** p(**x**).
- The likelihood of the model given a sample **x** is a measure of how well the model **explains the data**.
- **Model likelihood** has long been thought of as useful for OOD detection [5].



Out-of-distribution detection with generative models

- Generative models learn to approximate the **data distribution** $p(\mathbf{x})$.
- The likelihood of the model given a sample $\mathbf{x}$ is a measure of how well the model **explains the data**.
- **Model likelihood** has long been thought of as useful for OOD detection [5].

HVAE trained on FashionMNIST evaluated on FashionMNIST and MNIST.



Hierarchical VAE

corti DTU

We choose the hierarchical VAE as our model [30, 44].

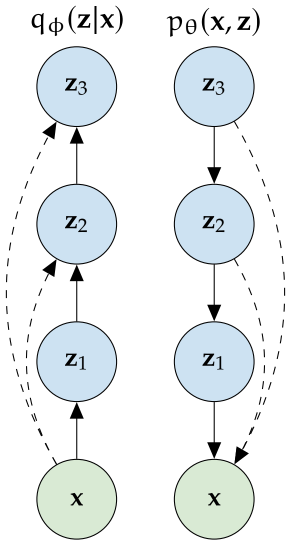
$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int p_\theta(\mathbf{x}|\mathbf{z}) p_\theta(\mathbf{z}) d\mathbf{z}$$

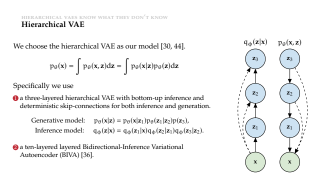
Specifically we use

1. a three-layered hierarchical VAE with bottom-up inference and deterministic skip-connections for both inference and generation.

$$\begin{aligned} \text{Generative model:} \quad & p_\theta(\mathbf{x}|\mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z}_1) p_\theta(\mathbf{z}_1|\mathbf{z}_2) p(\mathbf{z}_3), \\ \text{Inference model:} \quad & q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{z}_1|\mathbf{x}) q_\phi(\mathbf{z}_2|\mathbf{z}_1) q_\phi(\mathbf{z}_3|\mathbf{z}_2). \end{aligned}$$

2. a ten-layered layered Bidirectional-Inference Variational Autoencoder (BIVA) [36].





1. We want to explain why HVAEs fail at OOD detection when using their likelihood.
2. And how to fix them.

What is wrong with the ELBO for OOD detection?

corti DTU

We can split the ELBO into two terms

$$\mathcal{L}(\mathbf{x};\theta,\phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{reconstruction likelihood}} - \underbrace{D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))}_{\text{regularization penalty}} . \quad (1)$$

The first term is high if the data is well-explained by $\mathbf{z}$. The second term we can rewrite as,

$$D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\sum_{i=1}^{L-1}\log\frac{p_\theta(\mathbf{z}_i|\mathbf{z}_{i+1})}{q_\phi(\mathbf{z}_i|\mathbf{z}_{i-1})} + \log\frac{p_\theta(\mathbf{z}_L)}{q_\phi(\mathbf{z}_L|\mathbf{z}_{L-1})}\right] . \quad (2)$$

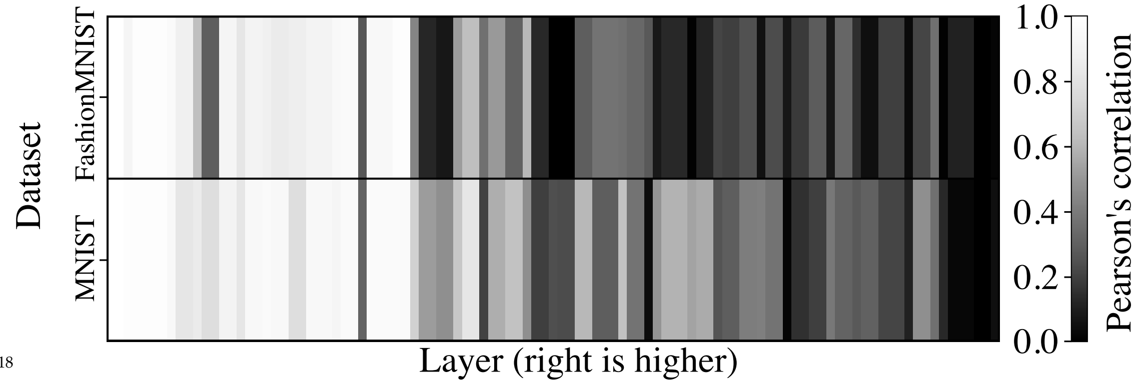
Since the individual terms are computed by summing over the dimensionality of $\mathbf{z}_i$, the absolute log-ratios grow with $\dim(\mathbf{z}_i)$.
Since the lower-most latent variables are usually higher dimensional than top ones, these are weighted higher in the ELBO.

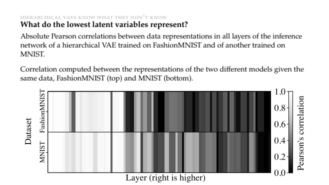
1. There is a tendency for the lower-most latent variables to be weighted higher than the higher ones.

What do the lowest latent variables represent?

Absolute Pearson correlations between data representations in all layers of the inference network of a hierarchical VAE trained on FashionMNIST and of another trained on MNIST.

Correlation computed between the representations of the two different models given the same data, FashionMNIST (top) and MNIST (bottom).





1. Strong evidence that the lowest latent variables are generalizing across datasets.

Likelihood ratios

We suggest to define a likelihood ratio score [9] using the ELBO $\mathcal{L}(\mathbf{x})$ and a relaxed bound $\mathcal{L}^{>k}(\mathbf{x})$.

$$\mathrm{LLR}^{>k}(\mathbf{x}) \equiv \mathcal{L}(\mathbf{x}) - \mathcal{L}^{>k}(\mathbf{x}) \,, \tag{3}$$

where the exact form of the bounds is,

$$\begin{aligned} \mathcal{L} &= \log p_\theta(\mathbf{x}) - D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p_\theta(\mathbf{z}|\mathbf{x})\right) \,, \\ \mathcal{L}^{>k} &= \log p_\theta(\mathbf{x}) - D_{\mathrm{KL}}\left(p_\theta(\mathbf{z}_{\leqslant}|\mathbf{z}_{>k}) q_\phi(\mathbf{z}_{>k}|\mathbf{x}) \,\|\, p_\theta(\mathbf{z}|\mathbf{x})\right) \,. \end{aligned} \tag{4}$$

In the likelihood ratio $\log p_\theta(\mathbf{x})$ cancels out and only the KL-divergences from the approximate to the true posterior remain.

$$\begin{aligned} \mathrm{LLR}^{>k}(\mathbf{x}) = &-D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p_\theta(\mathbf{z}|\mathbf{x})\right) \\ &+ D_{\mathrm{KL}}\left(p_\theta(\mathbf{z}_{\leqslant}|\mathbf{z}_{>k}) q_\phi(\mathbf{z}_{>k}|\mathbf{x}) \,\|\, p_\theta(\mathbf{z}|\mathbf{x})\right) \,. \end{aligned} \tag{5}$$

1. Write likelihood-ratio using the exact form of the bounds including intractable KL-divergence.

Importance sampling the ELBO

The importance weighted autoencoder (IWAE) bound is tight with the true likelihood in the limit of infinite samples, $S \to \infty$ [8],

$$\mathcal{L}_S = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log \frac{1}{N}\sum_{s=1}^{S} \frac{p(\mathbf{x}, \mathbf{z}^{(s)})}{q(\mathbf{z}^{(s)}|\mathbf{x})}\right] \leqslant \log p_\theta(\mathbf{x}) , \tag{6}$$

Consequently, by importance sampling the ELBO, $D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x})\right) \to 0$ and our likelihood ratio reduces to the KL-divergence of $\mathcal{L}^{>k}$.

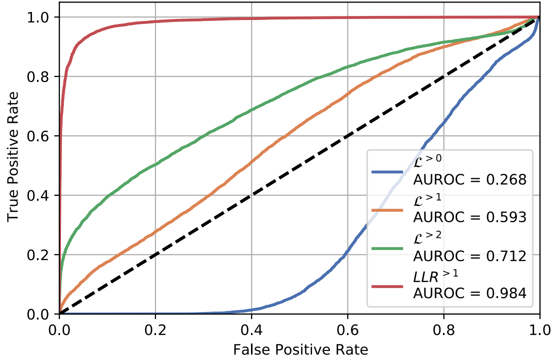
$$\mathrm{LLR}_S^{>k}(\mathbf{x}) \to D_{KL}(p(\mathbf{z}_{\leqslant}|\mathbf{z}_{>k})q(\mathbf{z}_{>k}|\mathbf{x}) \parallel p(\mathbf{z}|\mathbf{x})) . \tag{7}$$

$\mathrm{LLR}_S^{>k}(\mathbf{x})$ performs KL-divergence-based OOD detection using top-most latent variables.

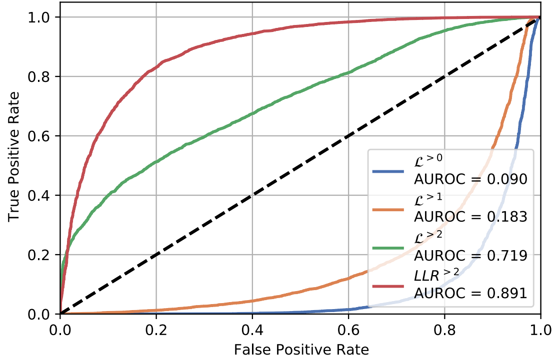
1. We can use the IWAE bound force the KL-divergence of the original ELBO towards zero.
2. Then, the likelihood ratio reduces to the KL-divergence of $\mathcal{L}^{>k}$.
3. We can hope that using the conditional prior for the lowest latent variables will make the KL-divergence more sensitive to OOD data in the top-most latent variables.

ROC curves with $\mathcal{L}^{>k}$ and LLR$^{>k}$

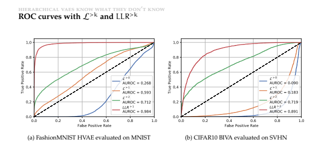
corti DTU



(a) FashionMNIST HVAE evaluated on MNIST



(b) CIFAR10 BIVA evaluated on SVHN



Results on CIFAR10/SVHN

Method	AUROC↑	AUPRC↑	FPR80↓
CIFAR10 (in) / SVHN (out)			
Use prior knowledge of OOD			
Backgr. contrast. LR (PixelCNN) [43]	0.930	0.881	0.066
Backgr. contrast. LR (VAE) [53]	0.265	-	-
Outlier exposure [21]	0.984	-	-
Input complexity (S, Glow) [46]	0.950	-	-
Input complexity (S, PixelCNN++) [46]	0.929	-	-
Input complexity (S, HVAE) (Ours) [46]	0.833	0.855	0.344
Use in-distribution data labels y			
Mahalanobis distance [32]	0.991	-	-
No OOD-specific assumptions			
- Ensembles			
WAIC, 5 models, Glow [11]	1.000	-	-
WAIC, 5 models, PixelCNN [43]	0.628	0.616	0.657
- Not ensembles			
Likelihood regret [53]	0.875	-	-
$\text{LLR}^{>2}$ + HVAE (ours)	0.811	0.837	0.394
$\text{LLR}^{>2}$ + BIVA (ours)	**0.891**	**0.875**	**0.172**

Conclusions

corti DTU

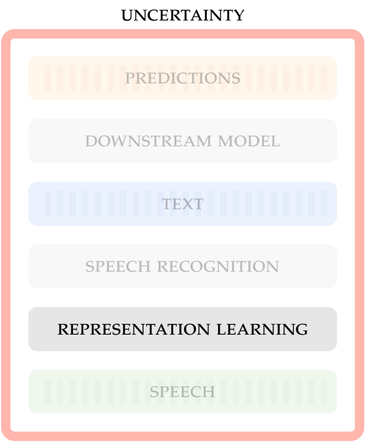
- Key observations:
  - The likelihood of a generative model is not a good score for OOD detection [37].
  - Strong correlations between some latent variables for different datasets.

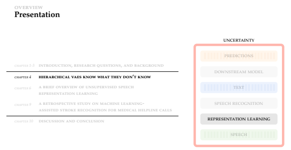
→ Provide explanation for why the likelihood fails fro OODD for HVAEs.

→ Proposed a likelihood-ratio score, $\mathrm{LLR}^{>k}$, that uses the conditional prior for the bottom-most latent variables in the hierarchy and showed its effectiveness.

HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW
Conclusions
- Key observations:
  - The likelihood of a generative model is not a good score for OOD detection [37].
  - Strong correlations between some latent variables for different datasets.
→ Provide explanation for why the likelihood fails fro OODD for HVAEs.
→ Proposed a likelihood-ratio score, $\mathrm{LLR}^{>k}$, that uses the conditional prior for the bottom-most latent variables in the hierarchy and showed its effectiveness.

Presentation

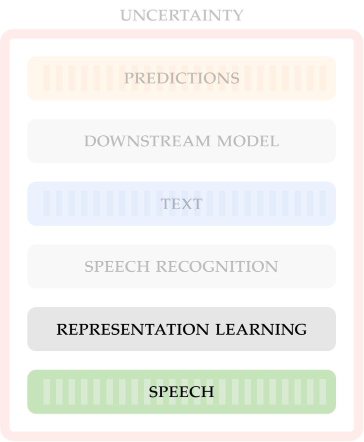


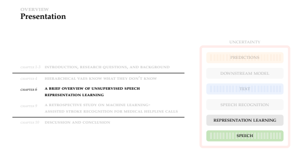


1. We now move on to our review paper of speech representation learning.

OVERVIEW

Presentation

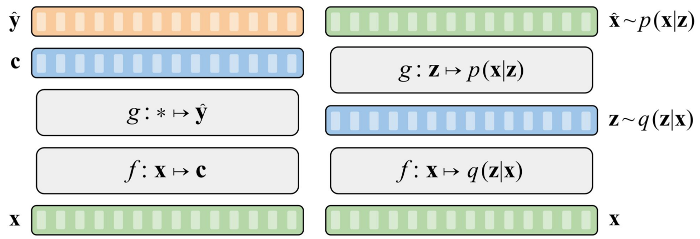




1. We now move on to our review paper of speech representation learning.

Overview: Representation Learning for Speech

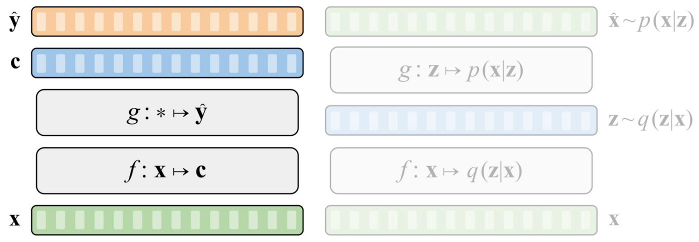
- Reviews two learning paradigms:
  - Self-supervised learning (SSL)
  - Probabilistic latent variable models (LVMs)
- Recent developments have been driven by self-supervised learning.
- A model-by-model overview for selected self-supervised models.

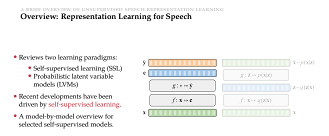


1. Bit more focus on the self-supervised learning models.

Overview: Representation Learning for Speech

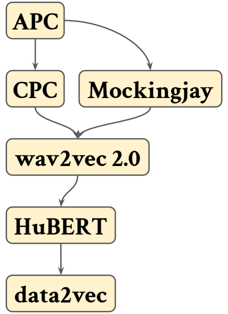
- Reviews two learning paradigms:
  - Self-supervised learning (SSL)
  - Probabilistic latent variable models (LVMs)
- Recent developments have been driven by self-supervised learning.
- A model-by-model overview for selected self-supervised models.

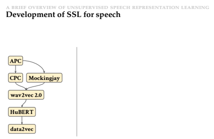




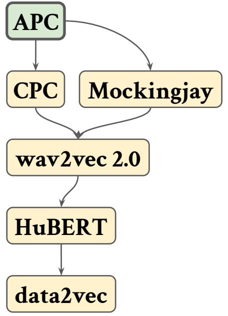
1. Bit more focus on the self-supervised learning models.

Development of SSL for speech

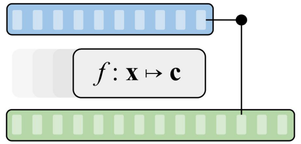


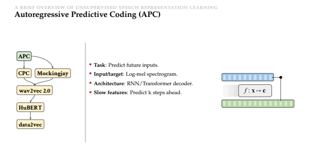


Autoregressive Predictive Coding (APC)

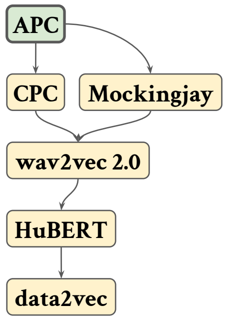


- **Task**: Predict future inputs.
- **Input/target**: Log-mel spectrogram.
- **Architecture**: RNN/Transformer decoder.
- **Slow features**: Predict k steps ahead.

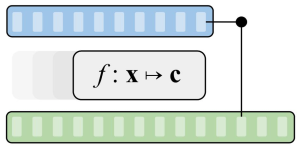


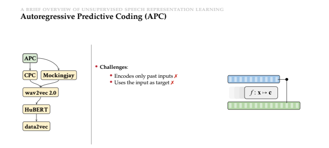


Autoregressive Predictive Coding (APC)



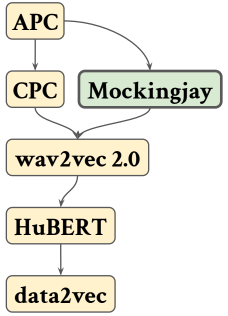
- **Challenges**:
  - Encodes only past inputs ✗
  - Uses the input as target ✗



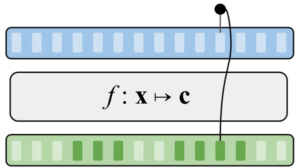


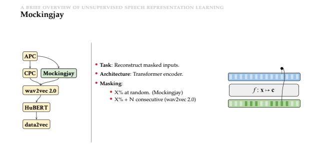
1. Encoding only past inputs limits the model's ability to learn from future context.
2. Input as target limits the model's ability to learn abstract representations.

Mockingjay



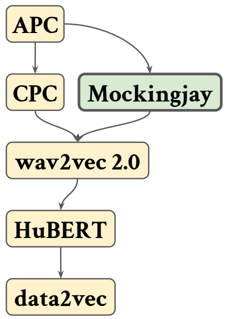
- **Task**: Reconstruct masked inputs.
- **Architecture**: Transformer encoder.
- **Masking**:
  - X% at random. (Mockingjay)
  - X% + N consecutive (wav2vec 2.0)



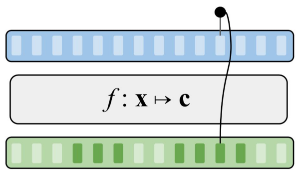


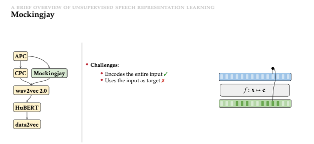
1. Replaces autoregressiveness with a masked reconstruction task.
2. Masking strategy to force learning long-term context.

Mockingjay

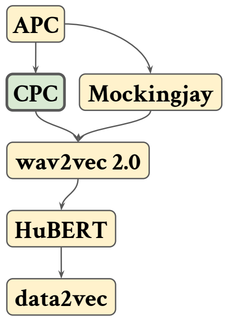


- **Challenges**:
  - Encodes the entire input ✓
  - Uses the input as target ✗

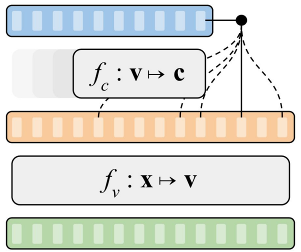


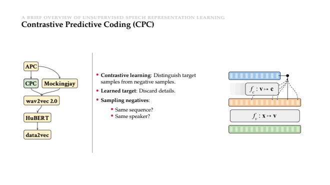


Contrastive Predictive Coding (CPC)



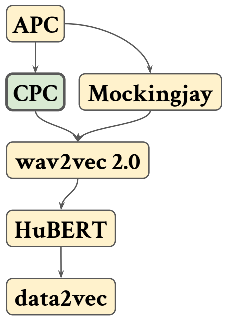
- **Contrastive learning**: Distinguish target samples from negative samples.
- **Learned target**: Discard details.
- **Sampling negatives**:
  - Same sequence?
  - Same speaker?



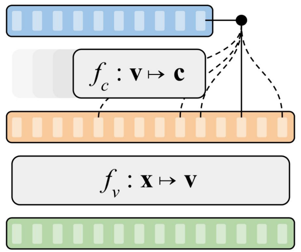


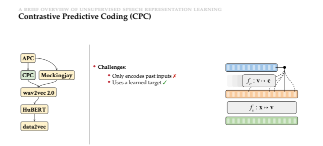
1. Contrastive models allow using a learned target. Discard unimportant details of input.
2. Negative sampling is a challenge and requires careful design.
3. Sampling different speakers as negatives learns strong speaker-identities.

Contrastive Predictive Coding (CPC)

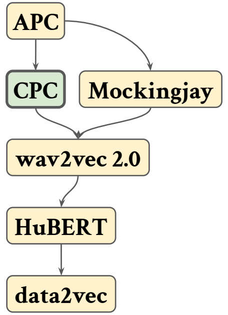


- **Challenges**:
  - Only encodes past inputs ✗
  - Uses a learned target ✓

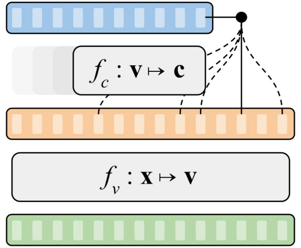


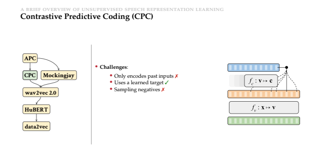


Contrastive Predictive Coding (CPC)

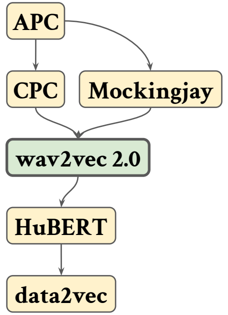


- **Challenges**:
  - Only encodes past inputs ✗
  - Uses a learned target ✓
  - Sampling negatives ✗

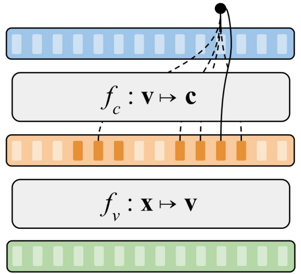


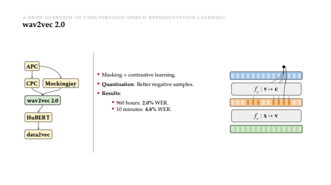


wav2vec 2.0



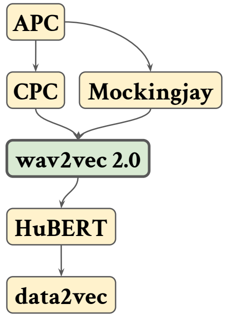
- Masking + contrastive learning.
- **Quantisation**: Better negative samples.
- **Results**:
  - 960 hours: **2.0%** WER.
  - 10 minutes: **4.8%** WER.



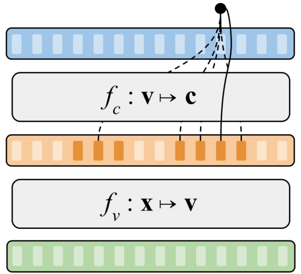


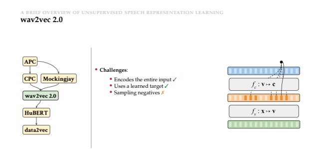
1. Training objective requires identifying the correct quantized latent audio representation in a set of distractors for each masked time step.
2. Quantisation improves negative sampling (requires approximation via Gumbel softmax).

wav2vec 2.0

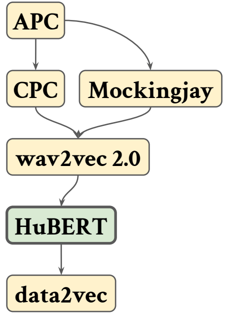


- **Challenges**:
  - Encodes the entire input ✓
  - Uses a learned target ✓
  - Sampling negatives ✗

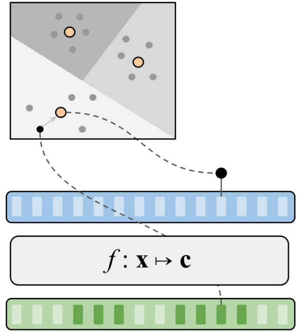


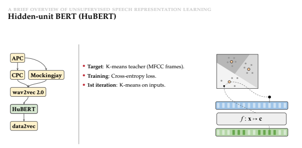


Hidden-unit BERT (HuBERT)



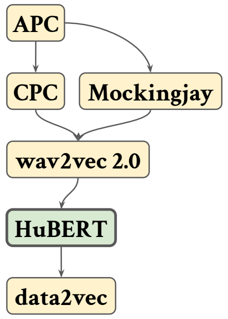
- **Target**: K-means teacher (MFCC frames).
- **Training**: Cross-entropy loss.
- **1st iteration**: K-means on inputs.



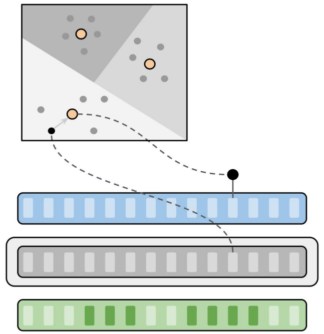


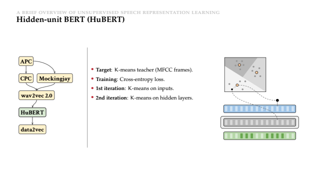
1. HuBERT approach predicts hidden cluster assignments of masked frames

Hidden-unit BERT (HuBERT)



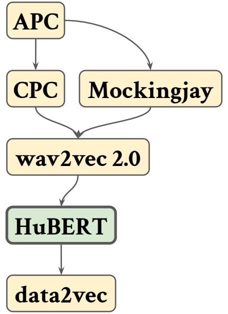
- **Target**: K-means teacher (MFCC frames).
- **Training**: Cross-entropy loss.
- **1st iteration**: K-means on inputs.
- **2nd iteration**: K-means on hidden layers.



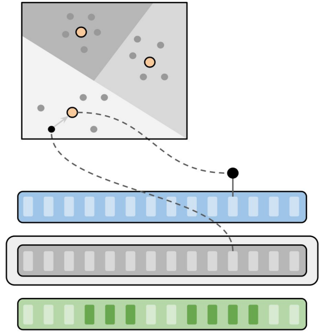


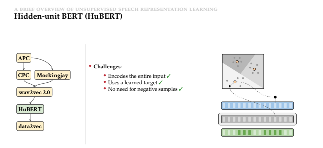
1. HuBERT approach predicts hidden cluster assignments of masked frames

Hidden-unit BERT (HuBERT)



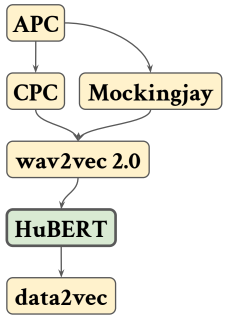
- **Challenges**:
  - Encodes the entire input ✓
  - Uses a learned target ✓
  - No need for negative samples ✓



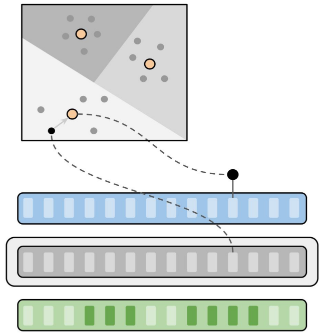


1. HuBERT approach predicts hidden cluster assignments of masked frames
2. Targets are still quantised although we no longer solve a contrastive sampling problem. Might reduce quality.

Hidden-unit BERT (HuBERT)

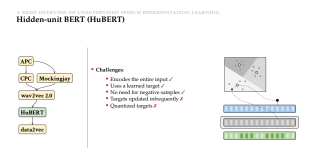


- **Challenges**:
  - Encodes the entire input ✓
  - Uses a learned target ✓
  - No need for negative samples ✓
  - Targets updated infrequently ✗
  - Quantized targets ✗



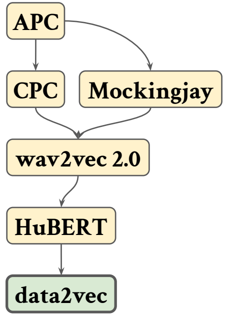
2024-03-04

UNCERTAINTY AND THE MEDICAL INTERVIEW
└a brief overview of unsupervised speech representation learning
└Hidden-unit BERT (HuBERT)

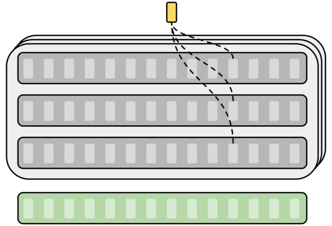


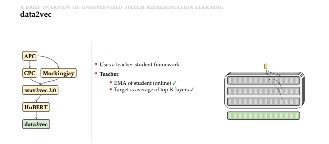
1. HuBERT approach predicts hidden cluster assignments of masked frames
2. Targets are still quantised although we no longer solve a contrastive sampling problem. Might reduce quality.

data2vec

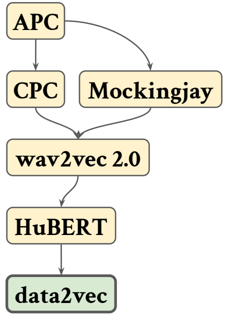


- Uses a teacher-student framework.
- **Teacher**:
  - EMA of student (online) ✓
  - Target is average of top K layers ✓

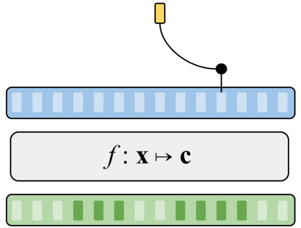




data2vec



- Uses a teacher-student framework.
- **Teacher**:
  - EMA of student (online) ✓
  - Target is average of top K layers ✓
- **Student training**: Smooth ℓ_1 loss.

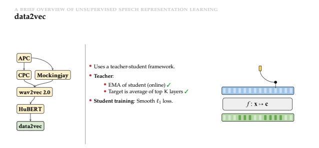


2024-03-04

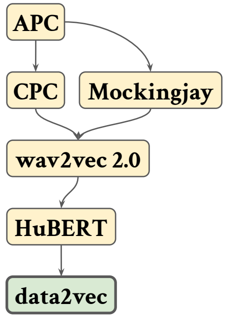
UNCERTAINTY AND THE MEDICAL INTERVIEW

└a brief overview of unsupervised speech representation learning

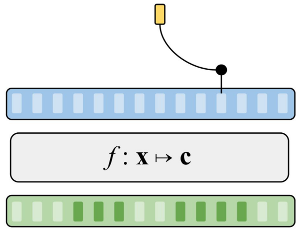
└data2vec

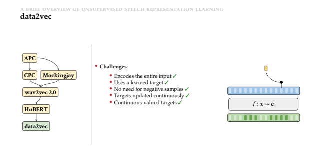


data2vec



- **Challenges**:
  - Encodes the entire input ✓
  - Uses a learned target ✓
  - No need for negative samples ✓
  - Targets updated continuously ✓
  - Continuous-valued targets ✓





Conclusions

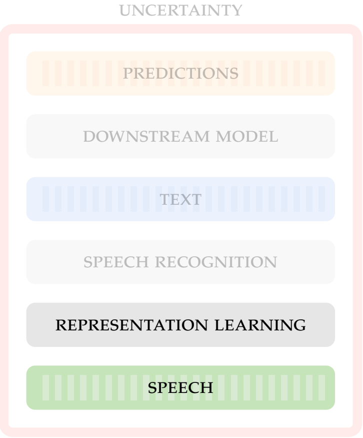
- **Main conclusions**:
  - The most popular self-supervised speech models can be compactly described by a few core design choices.
  - Many of these design choices are mirrored in earlier work on speech embedding models.
- **Open questions and limitations**:
  - Which design choices benefit which downstream tasks?
  - It is difficult to compare methods as model size and evaluation procedures differ widely between papers.

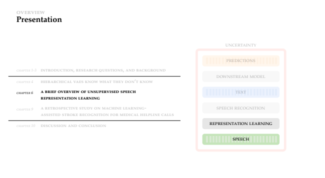
A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING
Conclusions
• Main conclusions:
• The most popular self-supervised speech models can be compactly described by a few core design choices.
• Many of these design choices are mirrored in earlier work on speech embedding models.
• Open questions and limitations:
• Which design choices benefit which downstream tasks?
• It is difficult to compare methods as model size and evaluation procedures differ widely between papers.

1. Predictive, Contrastive, Masking, Learned targets, Quantization, Teacher-student
2. Audio Word2vec, Speech2Vec, Unspeech which also used masking, and predictive and contrastive setups.

OVERVIEW

Presentation





OVERVIEW

Presentation

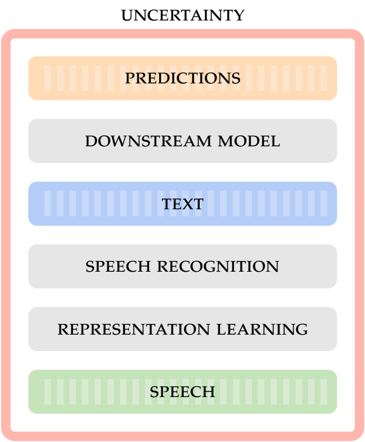
CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

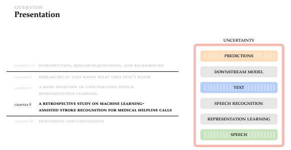
CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

***CHAPTER 9* A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS**

CHAPTER 10 DISCUSSION AND CONCLUSION





Stroke

- Stroke is the second leading cause of death (11.6%) and third leading cause of death and disability combined (5.7%) worldwide [18, 27, 31].
- Effective treatment is very time-sensitive [4, 50].
- The gateway to ambulance transport and hospital admittance is through prehospital telehealth services.
- Mobile stroke units have made it possible to deliver advanced treatment faster [20, 38].
- The effectiveness of mobile stroke units hinges on call-taker recognition of stroke [20, 38].
- Approximately half of all patients with stroke do not receive the correct triage for their condition from call-takers [7, 41, 52].

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS
Stroke
- Stroke is the second leading cause of death (11.6%) and third leading cause of death and disability combined (5.7%) worldwide [18, 27, 31].
- Effective treatment is very time-sensitive [4, 50].
- The gateway to ambulance transport and hospital admittance is through prehospital telehealth services.
- Mobile stroke units have made it possible to deliver advanced treatment faster [20, 38].
- The effectiveness of mobile stroke units hinges on call-taker recognition of stroke [20, 38].
- Approximately half of all patients with stroke do not receive the correct triage for their condition from call-takers [7, 41, 52].

1. [41]: Systematic review, recall = 41-83%, precision = 42-68%.
2. [52]: 1-1-2 calls at CEMS, recall = 66.2%, precision = 30.2%.
3. [7]: Systematic review, recall = 66.2%, precision = 30.2%.
4. Recognition rates also vary widely by call-center.

The study

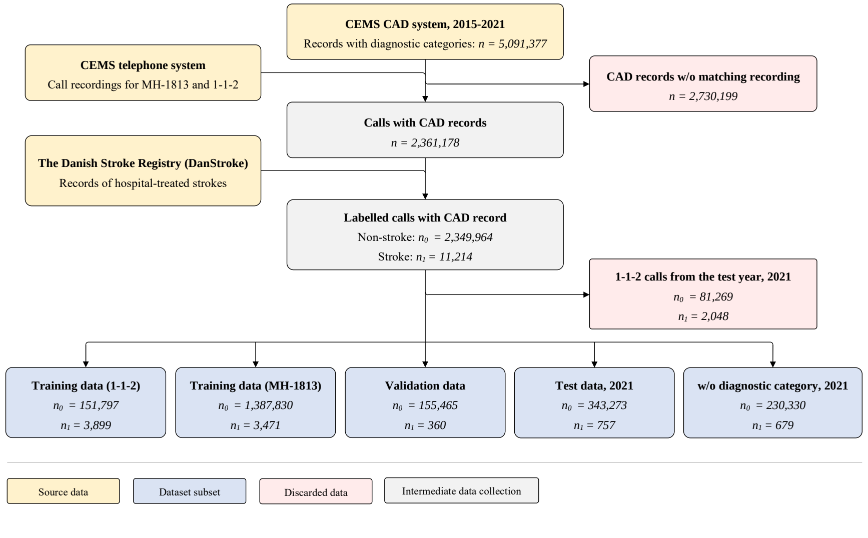
- Collaboration between Corti and the Copenhagen Emergency Medical Services (CEMS) ("Region Hovedstadens Akutberedskab").
- CEMS provides prehospital telehealth services in the Capital Region of Denmark (1.9M people).
- CEMS operates the 1-1-2 emergency line (similar to 9-1-1) and the 1813 medical helpline (non-life-threatening conditions when general practitioner is unavailable).

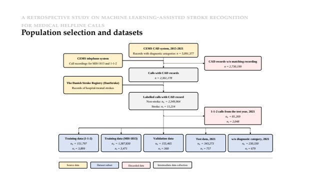
- We wanted to investigate if a machine learning model could assist call-takers of 1813 in recognizing stroke.

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS
The study
- Collaboration between Corti and the Copenhagen Emergency Medical Services (CEMS) ("Region Hovedstadens Akutberedskab").
- CEMS provides prehospital telehealth services in the Capital Region of Denmark (1.9M people).
- CEMS operates the 1-1-2 emergency line (similar to 9-1-1) and the 1813 medical helpline (non-life-threatening conditions when general practitioner is unavailable).
- We wanted to investigate if a machine learning model could assist call-takers of 1813 in recognizing stroke.

corti DTU

Population selection and datasets





1. Test data is MH-1813 2021.
2. All 1-1-2 data is used for training except 2021.
3. Validation data is sampled with stratified sampling from MH-1813 from 2015-2020.

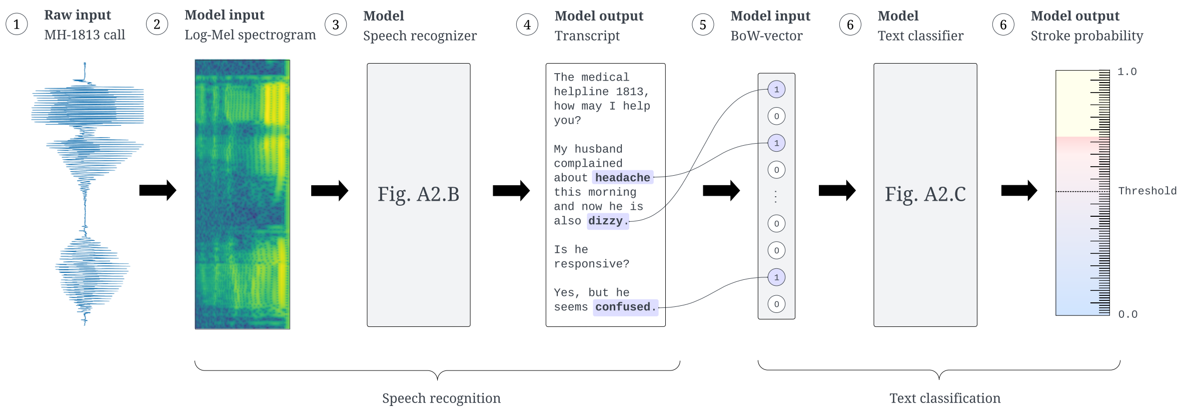
corti DTU

Population characteristics

		Training (112)	Training (MH-1813)	Validation	Test	2021 w/o category
All calls	**Num. calls**	155,696	1,391,301	155,825	344,030	231,009
	Female	74,640 (47.94%)	792,783 (56.98%)	86,959 (55.81%)	190,974 (55.51%)	134,324 (58.14%)
	Male	79,564 (51.10%)	596,760 (42.89%)	68,866 (44.19%)	153,050 (44.49%)	96,258 (41.67%)
	65+ years	72,930 (46.84%)	335,146 (24.09%)	30,313 (19.45%)	65,652 (19.08%)	81,488 (35.27%)
	Age (mean ± std.)	59.47 ± 21.24	47.12 ± 21.38	44.63 ± 20.08	44.31 ± 20.10	50.36 ± 22.77
Stroke calls	**Num. calls**	3,899	3,471	360	757	679
	Female	1,784 (45.76%)	1,654 (47.65%)	161 (44.72%)	349 (46.10%)	366 (53.90%)
	Male	2,115 (54.24%)	1,815 (52.29%)	199 (55.28%)	408 (53.90%)	313 (46.10%)
	65+ years	2,968 (76.12%)	2,421 (69.75%)	250 (69.44%)	555 (73.32%)	567 (83.51%)
	Age (mean ± std.)	72.91 ± 12.77	70.68 ± 13.85	70.93 ± 13.83	71.51 ± 13.41	73.41 ± 14.11
Non-stroke	**Num. calls**	151,797	1,387,830	155,465	343,273	230,330
	Female	72,856 (48.00%)	791,129 (57.00%)	86,798 (55.83%)	190,625 (55.53%)	133,958 (58.16%)
	Male	77,449 (51.02%)	594,945 (42.87%)	68,667 (44.17%)	152,642 (44.47%)	95,945 (41.66%)
	65+ years	69,962 (46.09%)	332,725 (23.97%)	30,063 (19.34%)	65,097 (18.96%)	80,921 (35.13%)
	Age (mean ± std.)	59.12 ± 21.30	47.06 ± 21.36	44.57 ± 20.05	44.25 ± 20.08	50.29 ± 22.76

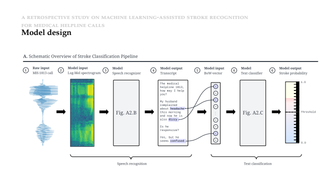
Model design

A. Schematic Overview of Stroke Classification Pipeline



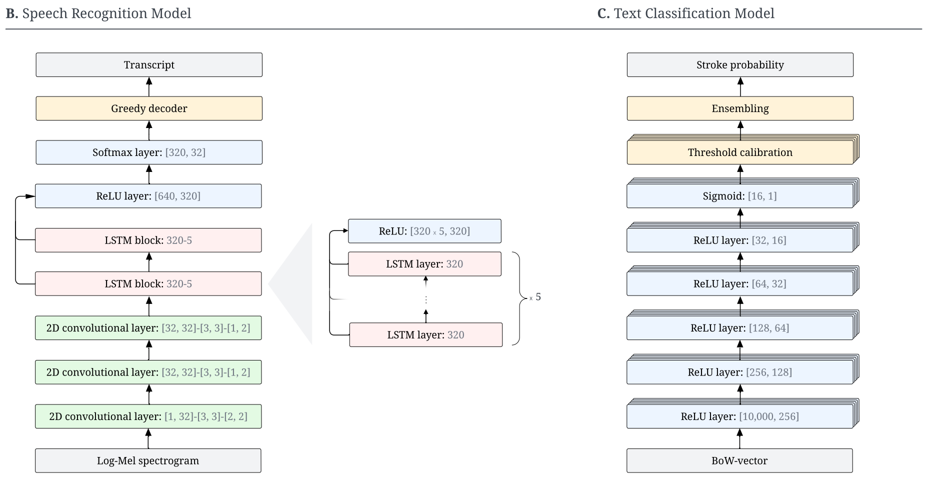
2024-03-04

UNCERTAINTY AND THE MEDICAL INTERVIEW
└a retrospective study on machine learning-assisted stroke recognition for medical helpline calls
└Model design



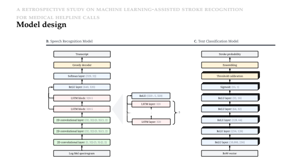
corti DTU

Model design



2024-03-04

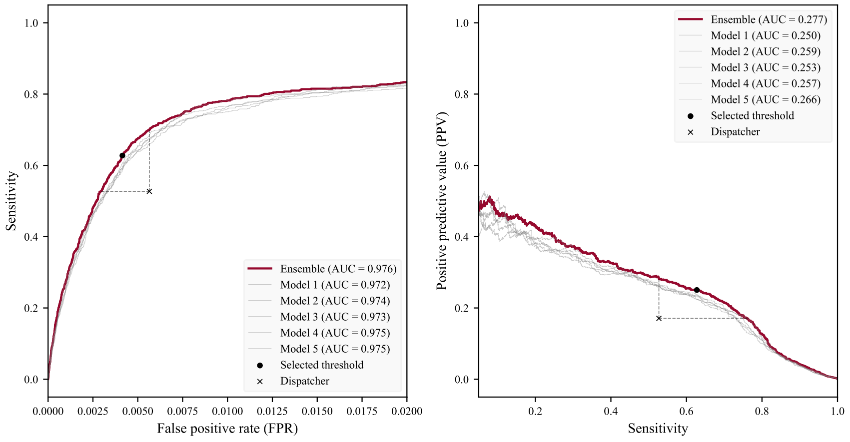
UNCERTAINTY AND THE MEDICAL INTERVIEW
└a retrospective study on machine learning-assisted stroke recognition for medical helpline calls
└Model design

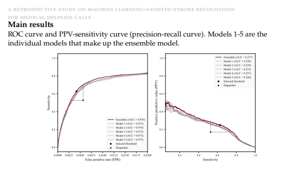


corti DTU

Main results

ROC curve and PPV-sensitivity curve (precision-recall curve). Models 1-5 are the individual models that make up the ensemble model.





Main results

Confusion matrices of predictions for call takers and the model on the test set. Numbers for the model are given as the rounded mean over eleven runs.

Call taker predictions	Ground truth labels: Positives	Ground truth labels: Negatives
Positives	True positives 399	False positives 1,938
Negatives	False negatives 358	True negatives 341,335

Model predictions	Ground truth labels: Positives	Ground truth labels: Negatives
Positives	True positives 477	False positives 1,440
Negatives	False negatives 280	True negatives 341,833

1. In absolute numbers, the model correctly identifies 78 cases of stroke missed by call-takers.
2. The model also reduces the number of false positives by about 500 from 1938 to 1440.

Main results

MH-1813 test set performance in demographic subgroups (age/sex) [mean (95% CI)].

Subset	Predictor	F1-score [%] ↑	Sensitivity [%] ↑	PPV [%] ↑	FOR [%] ↓ (1 - specificity)	FPR [%] ↓ (1 - NPV)
Overall	**Call-takers**	25.8 (23.7-27.9)	52.7 (49.2-56.4)	17.1 (15.5-18.6)	0.105 (0.094-0.116)	0.565 (0.539-0.590)
	Model	35.7 (35.0-36.4)	63.0 (62.0-64.1)	24.9 (24.3-25.5)	0.082 (0.079-0.085)	0.419 (0.413-0.426)
18-64 years	**Call-takers**	15.9 (13.1-18.5)	50.5 (43.6-57.2)	9.40 (7.61-11.2)	0.036 (0.028-0.043)	0.353 (0.331-0.375)
	Model	22.9 (21.8-24.0)	54.1 (52.1-56.3)	14.5 (13.8-15.3)	0.033 (0.031-0.035)	0.231 (0.226-0.236)
65+ years	**Call-takers**	32.9 (30.1-35.7)	53.5 (49.4-57.6)	23.7 (21.4-26.0)	0.401 (0.352-0.449)	1.467 (1.373-1.560)
	Model	42.8 (41.9-43.7)	66.3 (65.1-67.5)	31.6 (30.8-32.4)	0.290 (0.278-0.303)	1.224 (1.198-1.249)
Male	**Call-takers**	30.2 (27.2-33.3)	53.9 (49.1-58.9)	21.0 (18.5-23.5)	0.124 (0.105-0.141)	0.542 (0.506-0.580)
	Model	39.0 (38.0-40.1)	63.7 (62.3-65.2)	28.1 (27.3-29.0)	0.097 (0.093-0.102)	0.435 (0.425-0.445)
Female	**Call-takers**	21.9 (19.1-24.6)	51.3 (46.0-56.6)	13.9 (12.0-15.8)	0.090 (0.076-0.103)	0.582 (0.547-0.616)
	Model	32.4 (31.4-33.4)	62.3 (60.7-63.8)	21.9 (21.1-22.7)	0.069 (0.066-0.073)	0.407 (0.399-0.416)

Occlusion analysis — Which features are evidence?

Features with positive ranking score ($r^{(w)} > 0$) computed on stroke positive predictions ($D = 1{,}897$)					
Rank	Word, *w (translated)*	Occurrences, $D^{(w)}$	Rank	Word, *w (translated)*	Occurrences, $D^{(w)}$
1.	Ambulance	1,680	16.	Difficulties speaking	44
2.	Blood clot	895	17.	Hemorrhagic stroke	133
3.	Left	1,108	18.	Hand	297
4.	Right	1,050	19.	The ambulance	521
5.	Double vision	84	20.	Slurred speech	58
6.	The words	344	21.	Blood clots	224
7.	Suddenly	783	22.	Fast	663
8.	Arm	709	23.	Express	44
9.	Side	1,139	24.	Blood thinner	259
10.	Stroke	117	25.	Incoherent	15
11.	Double	113	26.	Lopsided	211
12.	Control	134	27.	Reduced	528
13.	Call	39	28.	Hangs	628
14.	Numb	94	29.	Transient	48
15.	Minutes	763	30.	Not making sense	14

[Recognition, Symptom, Urgency/Time]

1. Wanted to examine the features that were important for predictions.
2. Performed an occlusion analysis where we remove one word from the input at a time.
3. Sort the words by their impact on the model's output logit.

Occlusion analysis — Which features are evidence?

Features with positive ranking score ($r^{(w)} > 0$) **computed on stroke positive predictions** ($D = 1,897$)

Rank	Word, *w (translated)*	Occurrences, $D^{(w)}$	Rank	Word, *w (translated)*	Occurrences, $D^{(w)}$
1.	Ambulance	1,680	16.	Difficulties speaking	44
2.	Blood clot	895	17.	Hemorrhagic stroke	133
3.	Left	1,108	18.	Hand	297
4.	Right	1,050	19.	The ambulance	521
5.	Double vision	84	20.	Slurred speech	58
6.	The words	344	21.	Blood clots	224
7.	Suddenly	783	22.	Fast	663
8.	Arm	709	23.	Express	44
9.	Side	1,139	24.	Blood thinner	259
10.	Stroke	117	25.	Incoherent	15
11.	Double	113	26.	Lopsided	211
12.	Control	134	27.	Reduced	528
13.	Call	39	28.	Hangs	628
14.	Numb	94	29.	Transient	48
15.	Minutes	763	30.	Not making sense	14

[Recognition, Symptom, Urgency/Time]

1. Wanted to examine the features that were important for predictions.
2. Performed an occlusion analysis where we remove one word from the input at a time.
3. Sort the words by their impact on the model's output logit.

Occlusion analysis — Which features are evidence?

Features with positive ranking score ($r^{(w)} > 0$) computed on stroke positive predictions ($D = 1,897$)

Rank	Word, *w (translated)*	Occurrences, $D^{(w)}$	Rank	Word, *w (translated)*	Occurrences, $D^{(w)}$
1.	Ambulance	1,680	16.	Difficulties speaking	44
2.	Blood clot	895	17.	Hemorrhagic stroke	133
3.	Left	1,108	18.	Hand	297
4.	Right	1,050	19.	The ambulance	521
5.	Double vision	84	20.	Slurred speech	58
6.	The words	344	21.	Blood clots	224
7.	Suddenly	783	22.	Fast	663
8.	Arm	709	23.	Express	44
9.	Side	1,139	24.	Blood thinner	259
10.	Stroke	117	25.	Incoherent	15
11.	Double	113	26.	Lopsided	211
12.	Control	134	27.	Reduced	528
13.	Call	39	28.	Hangs	628
14.	Numb	94	29.	Transient	48
15.	Minutes	763	30.	Not making sense	14

[Recognition, Symptom, Urgency/Time]

1. Wanted to examine the features that were important for predictions.
2. Performed an occlusion analysis where we remove one word from the input at a time.
3. Sort the words by their impact on the model's output logit.

Occlusion analysis — Which features are evidence?

Features with positive ranking score ($r^{(w)} > 0$) computed on stroke positive predictions ($D = 1,897$)

Rank	Word, *w* (translated)	Occurrences, $D^{(w)}$	Rank	Word, *w* (translated)	Occurrences, $D^{(w)}$
1.	Ambulance	1,680	16.	Difficulties speaking	44
2.	Blood clot	895	17.	Hemorrhagic stroke	133
3.	Left	1,108	18.	Hand	297
4.	Right	1,050	19.	The ambulance	521
5.	Double vision	84	20.	Slurred speech	58
6.	The words	344	21.	Blood clots	224
7.	Suddenly	783	22.	Fast	663
8.	Arm	709	23.	Express	44
9.	Side	1,139	24.	Blood thinner	259
10.	Stroke	117	25.	Incoherent	15
11.	Double	113	26.	Lopsided	211
12.	Control	134	27.	Reduced	528
13.	Call	39	28.	Hangs	628
14.	Numb	94	29.	Transient	48
15.	Minutes	763	30.	Not making sense	14

[Recognition, Symptom, Urgency/Time]

1. Wanted to examine the features that were important for predictions.
2. Performed an occlusion analysis where we remove one word from the input at a time.
3. Sort the words by their impact on the model's output logit.

Occlusion analysis – Which features are counter-evidence?

Rank	Word, w (translated)	Occurrences, $D^{(w)}$	Rank	Word, w (translated)	Occurrences, $D^{(w)}$
Features with negative ranking score ($r^{(w)} < 0$) computed on stroke-negative predictions ($D = 342,133$)					
1.	Tetanus	4,378	16.	The pharmacy	10,085
2.	Pregnant	8,749	17.	The stomach	42,105
3.	Cut	7,592	18.	Psychiatric	3,688
4.	Bandage	4,561	19.	Pneumonia	7,597
5.	Amager (a location)	23,776	20.	Stomach pain	10,551
6.	O'clock	94,436	21.	Stool	19,155
7.	The emergency room	42,809	22.	The ribs	3,928
8.	The police	2,903	23.	Bleed	10,501
9.	Swollen	60,559	24.	Bleeding	24,313
10.	Over the counter (OTC)	4,641	25.	Ribs	2,941
11.	The neck	30,151	26.	Broken	19,415
12.	Fever	112,586	27.	Inflammation	10,050
13.	Prescription	5,450	28.	Common cold	8,127
14.	Centimeter	12,026	29.	Morning or morrow	78,558
15.	The knee	8,875	30.	Swelling	17,762

[Recognition, Symptom, Urgency/Time]

corti DTU

Occlusion analysis – Which features are counter-evidence?

Features with negative ranking score ($r^{(w)} < 0$) computed on stroke-negative predictions ($D = 342,133$)

Rank	Word, *w (translated)*	Occurrences, $D^{(w)}$	Rank	Word, *w (translated)*	Occurrences, $D^{(w)}$
1.	Tetanus	4,378	16.	The pharmacy	10,085
2.	Pregnant	8,749	17.	The stomach	42,105
3.	Cut	7,592	18.	Psychiatric	3,688
4.	Bandage	4,561	19.	Pneumonia	7,597
5.	Amager (a location)	23,776	20.	Stomach pain	10,551
6.	O'clock	94,436	21.	Stool	19,155
7.	The emergency room	42,809	22.	The ribs	3,928
8.	The police	2,903	23.	Bleed	10,501
9.	Swollen	60,559	24.	Bleeding	24,313
10.	Over the counter (OTC)	4,641	25.	Ribs	2,941
11.	The neck	30,151	26.	Broken	19,415
12.	Fever	112,586	27.	Inflammation	10,050
13.	Prescription	5,450	28.	Common cold	8,127
14.	Centimeter	12,026	29.	Morning or morrow	78,558
15.	The knee	8,875	30.	Swelling	17,762

[Recognition, Symptom, Urgency/Time]

Occlusion analysis – Which features are counter-evidence?

Features with negative ranking score ($r^{(w)} < 0$) computed on stroke-negative predictions ($D = 342,133$)

Rank	Word, w *(translated)*	Occurrences, $D^{(w)}$	Rank	Word, w *(translated)*	Occurrences, $D^{(w)}$
1.	Tetanus	4,378	16.	The pharmacy	10,085
2.	Pregnant	8,749	17.	The stomach	42,105
3.	Cut	7,592	18.	Psychiatric	3,688
4.	Bandage	4,561	19.	Pneumonia	7,597
5.	Amager (a location)	23,776	20.	Stomach pain	10,551
6.	O'clock	94,436	21.	Stool	19,155
7.	The emergency room	42,809	22.	The ribs	3,928
8.	The police	2,903	23.	Bleed	10,501
9.	Swollen	60,559	24.	Bleeding	24,313
10.	Over the counter (OTC)	4,641	25.	Ribs	2,941
11.	The neck	30,151	26.	Broken	19,415
12.	Fever	112,586	27.	Inflammation	10,050
13.	Prescription	5,450	28.	Common cold	8,127
14.	Centimeter	12,026	29.	Morning or morrow	78,558
15.	The knee	8,875	30.	Swelling	17,762

[Recognition, Symptom, Urgency/Time]

UNCERTAINTY AND THE MEDICAL INTERVIEW
a retrospective study on machine learning-assisted stroke recognition for medical helpline calls
Occlusion analysis – Which features are counter-evidence?
2024-03-04

Occlusion analysis – Which features are counter-evidence?

Features with negative ranking score ($r^{(w)} < 0$) computed on stroke-negative predictions ($D = 342,133$)

Rank	Word, w *(translated)*	Occurrences, $D^{(w)}$	Rank	Word, w *(translated)*	Occurrences, $D^{(w)}$
1.	Tetanus	4,378	16.	The pharmacy	10,085
2.	Pregnant	8,749	17.	The stomach	42,105
3.	Cut	7,592	18.	Psychiatric	3,688
4.	Bandage	4,561	19.	Pneumonia	7,597
5.	Amager (a location)	23,776	20.	Stomach pain	10,551
6.	O'clock	94,436	21.	Stool	19,155
7.	The emergency room	42,809	22.	The ribs	3,928
8.	The police	2,903	23.	Bleed	10,501
9.	Swollen	60,559	24.	Bleeding	24,313
10.	Over the counter (OTC)	4,641	25.	Ribs	2,941
11.	The neck	30,151	26.	Broken	19,415
12.	Fever	112,586	27.	Inflammation	10,050
13.	Prescription	5,450	28.	Common cold	8,127
14.	Centimeter	12,026	29.	Morning or morrow	78,558
15.	The knee	8,875	30.	Swelling	17,762

[Recognition, Symptom, Urgency/Time]

Future work

- Machine learning
  - Learning to predict directly from audio data (SSL).
  - Learning to defer to predict methods [51].
- Clinical applications
  - Mental health: Screening for suicide risk in emergency and medical helpline calls.
  - Maternity ward: Screening for serious pregnancy complications.

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS
Future work
• Machine learning
• Learning to predict directly from audio data (SSL).
• Learning to defer to predict methods [51].
• Clinical applications
• Mental health: Screening for suicide risk in emergency and medical helpline calls.
• Maternity ward: Screening for serious pregnancy complications.

1. Learning to defer to predict is related to uncertainty estimation.

OVERVIEW
Presentation

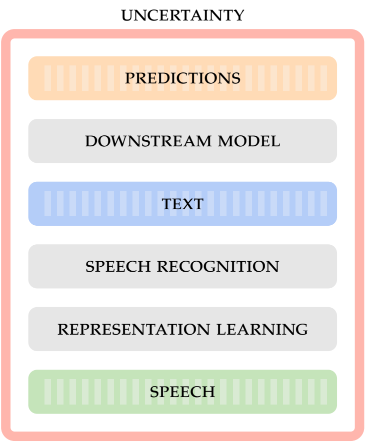
CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

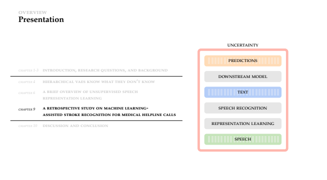
CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

CHAPTER 10 DISCUSSION AND CONCLUSION

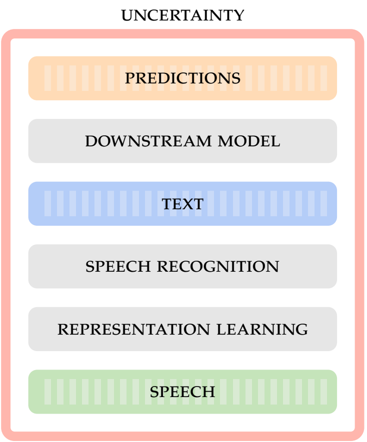


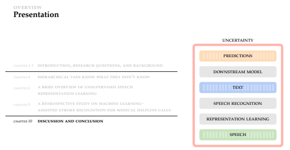
2024-03-04

UNCERTAINTY AND THE MEDICAL INTERVIEW
└overview
└Presentation



OVERVIEW
Presentation





The broad picture: The thesis topic since 2020

2020 **Project start**

- Out-of-distribution detection with generative models: Mysterious new topic.
- Speech representation/recognition: Inflection point between supervised methods and new self-supervised approaches.

DISCUSSION
The broad picture: The thesis topic since 2020
2020 Project start
- Out-of-distribution detection with generative models: Mysterious new topic.
- Speech representation/recognition: Inflection point between supervised methods and new self-supervised approaches.

The broad picture: The thesis topic since 2020

2020 Project start

- Out-of-distribution detection with generative models: Mysterious new topic.
- Speech representation/recognition: Inflection point between supervised methods and new self-supervised approaches.

2024 Project end

- Out-of-distribution detection is a mature field with a wide range of methods.
- Self-supervised learning is the dominant paradigm in speech recognition - challenged by weak labelling.

DISCUSSION
The broad picture: The thesis topic since 2020
2020 Project start
- Out-of-distribution detection with generative models: Mysterious new topic.
- Speech representation/recognition: Inflection point between supervised methods and new self-supervised approaches.
2024 Project end
- Out-of-distribution detection is a mature field with a wide range of methods.
- Self-supervised learning is the dominant paradigm in speech recognition - challenged by weak labelling.

The broad picture: The thesis topic since 2020

corti DTU

2020 Project start

- Out-of-distribution detection with generative models: Mysterious new topic.
- Speech representation/recognition: Inflection point between supervised methods and new self-supervised approaches.

2024 Project end

- Out-of-distribution detection is a mature field with a wide range of methods.
- Self-supervised learning is the dominant paradigm in speech recognition - challenged by weak labelling.
- Clinical research is increasingly becoming interested in the use of machine learning.

The broad picture: The thesis topic since 2020
2020 Project start
- Out-of-distribution detection with generative models: Mysterious new topic.
- Speech representation/recognition: Inflection point between supervised methods and new self-supervised approaches.
2024 Project end
- Out-of-distribution detection is a mature field with a wide range of methods.
- Self-supervised learning is the dominant paradigm in speech recognition - challenged by weak labelling.
- Clinical research is increasingly becoming interested in the use of machine learning.

What lies ahead

corti DTU

- Selective out-of-distribution detection

  Two pairs of distributions may have identical divergence, but in different dimensions. How do we control features in black-box models?

- Self-supervised learning in the wild

  Does the recent progress on academic datasets translate to this real-world setting?

  Speech recognition has been the cornerstone benchmarking task. How do we target spoken language understanding directly?

- Large language models in medical dialogue

  LLMs will likely play a central role in the future of medical documentation and communication. How do we get a grip of their uncertainty?

What lies ahead
Selective out-of-distribution detection
Two pairs of distributions may have identical divergence, but in different dimensions. How do we control features in black-box models?
Self-supervised learning in the wild
Does the recent progress on academic datasets translate to this real-world setting?
Speech recognition has been the cornerstone benchmarking task. How do we target spoken language understanding directly?
Large language models in medical dialogue
LLMs will likely play a central role in the future of medical documentation and communication. How do we get a grip of their uncertainty?

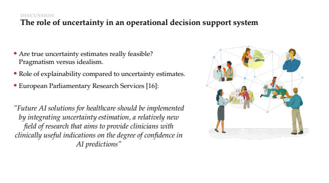
1. We saw how certain features can dominate an uncertainty score. How do we properly control which features we want to use to detect out-of-distribution data?
2. Self-supervised learning has made great strides in academic benchmarks. How does this translate to real-world settings?
3. LLMs will be important for medical documentation and communication going forward. What does uncertainty look like in this context?

DISCUSSION

The role of uncertainty in an operational decision support system

- Are true uncertainty estimates really feasible? Pragmatism versus idealism.
- Role of explainability compared to uncertainty estimates.
- European Parliamentary Research Services [16]:

"Future AI solutions for healthcare should be implemented by integrating uncertainty estimation, a relatively new field of research that aims to provide clinicians with clinically useful indications on the degree of confidence in AI predictions"



1. LVMs were difficult to scale to the problems we care about.
2. LVMs did not convincingly outperform more pragmatic approaches (e.g. deferring).
3. Bayesian methods, deferring to predict, discriminative uncertainty.
4. What will happen if uncertainty estimates become a regulatory requirement?

Thank you for your attention.

Thank you for your attention.

Bibliography I

[1] Aksan, E., Hilliges, O., "STCN: Stochastic Temporal Convolutional Networks". In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. New Orleans, LA, USA, 2019 (cited on pages 111, 112).

[2] Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., Auli, M., *Data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language*. Facebook AI Research blog, 2022 (cited on page 112).

[3] Baevski, A., Zhou, H., Mohamed, A., Auli, M., "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*. Virtual, 2020. arXiv: `2006.11477` (cited on page 112).

[4] Berge, E., Whiteley, W., Audebert, H., De Marchis, G. M., Fonseca, A. C., Padiglioni, C., Pérez de la Ossa, N., Strbian, D., Tsivgoulis, G., Turc, G., "European Stroke Organisation (ESO) Guidelines on Intravenous Thrombolysis for Acute Ischaemic Stroke". In: *European Stroke Journal* 6.1 (2021) (cited on page 65).

[5] Bishop, C. M. "Novelty Detection and Neural-Network Validation". In: *IEE Proceedings - Vision, Image and Signal Processing* 141.4 (1994). ISSN: 1350245x, 13597108 (cited on pages 31, 32).

[6] Blomberg, S. N., Christensen, H. C., Lippert, F., Ersbøll, A. K., Torp-Petersen, C., Sayre, M. R., Kudenchuk, P. J., Folke, F., "Effect of Machine Learning on Dispatcher Recognition of Out-of-Hospital Cardiac Arrest during Calls to Emergency Medical Services: A Randomized Clinical Trial". In: *JAMA Network Open* 4.1 (2021) (cited on page 113).

[7] Bohm, K., Kurland, L., "The Accuracy of Medical Dispatch - A Systematic Review". In: *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 26 (2018) (cited on page 65).

Bibliography I
[1] Aksan, E., Hilliges, O., "STCN: Stochastic Temporal Convolutional Networks". In: Proceedings of the 7th International Conference on Learning Representations (ICLR). New Orleans, LA, USA, 2019 (cited on pages 111, 112).
[2] Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., Auli, M., Data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. Facebook AI Research blog, 2022 (cited on page 112).
[3] Baevski, A., Zhou, H., Mohamed, A., Auli, M., "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS). Virtual, 2020. arXiv: 2006.11477 (cited on page 112).
[4] Berge, E., Whiteley, W., Audebert, H., De Marchis, G. M., Fonseca, A. C., Padiglioni, C., Pérez de la Ossa, N., Strbian, D., Tsivgoulis, G., Turc, G., "European Stroke Organisation (ESO) Guidelines on Intravenous Thrombolysis for Acute Ischaemic Stroke". In: European Stroke Journal 6.1 (2021) (cited on page 65).
[5] Bishop, C. M. "Novelty Detection and Neural-Network Validation". In: IEE Proceedings - Vision, Image and Signal Processing 141.4 (1994). ISSN: 1350245x, 13597108 (cited on pages 31, 32).
[6] Blomberg, S. N., Christensen, H. C., Lippert, F., Ersbøll, A. K., Torp-Petersen, C., Sayre, M. R., Kudenchuk, P. J., Folke, F., "Effect of Machine Learning on Dispatcher Recognition of Out-of-Hospital Cardiac Arrest during Calls to Emergency Medical Services: A Randomized Clinical Trial". In: JAMA Network Open 4.1 (2021) (cited on page 113).
[7] Bohm, K., Kurland, L., "The Accuracy of Medical Dispatch - A Systematic Review". In: Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine 26 (2018) (cited on page 65).

Bibliography II

[8] Burda, Y., Grosse, R., Salakhutdinov, R. R., "Importance Weighted Autoencoders". In: *Proceedings of the 4th International Conference on Learning Representations (ICLR)*. San Juan, Puerto Rico, 2016 (cited on page 37).

[9] Buse, A. "The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note". In: *The American Statistician* 36 (3a 1982) (cited on pages 36, 102).

[10] Carver, N., Gupta, V., Hipskind, J. E., "Medical Errors". In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2024. pmid: 28613514 (cited on pages 7, 8).

[11] Choi, H., Jang, E., Alemi, A. A., *WAIC, but Why? Generative Ensembles for Robust Anomaly Detection*. 2019. arXiv: 1810.01392 (cited on page 39).

[12] Chung, Y.-A., Hsu, W.-N., Tang, H., Glass, J., *An Unsupervised Autoregressive Model for Speech Representation Learning*. 2019. arXiv: 1904.03240 (cited on page 112).

[13] Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C., Bengio, Y., "A Recurrent Latent Variable Model for Sequential Data". In: *Proceedings of the 29th Conference on Neural Information Processing Systems (NeurIPS)*. Montréal, Quebec, Canada, 2015 (cited on pages 111, 112).

[14] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. arXiv: 1810.04805. (Visited on 11 February 2019) (cited on page 115).

[15] Ebbers, J., Heymann, J., Drude, L., Glarner, T., Haeb-Umbach, R., Raj, B., "Hidden Markov Model Variational Autoencoder for Acoustic Unit Discovery". In: *Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2017 (cited on page 111).

Bibliography III

[16] European Parliament, Directorate-General for Parliamentary Research Services, Lekadir, K., Quaglio, G., Tselioudis Garmendia, A., Gallin, C., *Artificial Intelligence in Healthcare – Applications, Risks, and Ethical and Societal Impacts*. European Parliament, 2022 (cited on page 89).

[17] Fraccaro, M., Sønderby, S. K., Paquet, U., Winther, O., "Sequential Neural Models with Stochastic Layers". In: *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS)*. Barcelona, Spain, 2016 (cited on pages 111, 112).

[18] GBD 2019 Stroke Collaborators, "Global, Regional, and National Burden of Stroke and Its Risk Factors, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019". In: *The Lancet Neurology* 20.10 (2021). ISSN: 1474-4422 (cited on page 65).

[19] Glarner, T., Hanebrink, P., Ebbers, J., Haeb-Umbach, R., "Full Bayesian Hidden Markov Model Variational Autoencoder for Acoustic Unit Discovery". In: *Proceedings of the 19th Annual Conference of the International Speech Communication Association (Interspeech)*. Hyderabad, India: ISCA, 2018 (cited on page 111).

[20] Hariharan, P., Tariq, M. B., Grotta, J. C., Czap, A. L., "Mobile Stroke Units: Current Evidence and Impact". In: *Current Neurology and Neuroscience Reports* 22.1 (2022) (cited on page 65).

[21] Hendrycks, D., Mazeika, M., Dietterich, T. G., "Deep Anomaly Detection with Outlier Exposure". In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. New Orleans, LA, USA, 2019 (cited on page 39).

Bibliography IV

[22] Horwitz, L. I., Jenq, G. Y., Brewster, U. C., Chen, C., Kanade, S., Van Ness, P. H., Araujo, K. L. B., Ziaeian, B., Moriarty, J. P., Fogerty, R., Krumholz, H. M., "Comprehensive Quality of Discharge Summaries at an Academic Medical Center". In: *Journal of hospital medicine : an official publication of the Society of Hospital Medicine* 8.8 (2013). ISSN: 1553-5592. pmid: 23526813 (cited on pages 9, 10).

[23] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., Mohamed, A., "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units". In: (2021) (cited on page 112).

[24] Hsu, W.-N., Zhang, Y., Glass, J., *Learning Latent Representations for Speech Generation and Transformation*. 2017. arXiv: 1704.04222 (cited on pages 111, 112).

[25] Hsu, W.-N., Zhang, Y., Glass, J., "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data". In: *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*. Long Beach, CA, USA, 2017 (cited on pages 111, 112).

[26] Joukes, E., Abu-Hanna, A., Cornet, R., De Keizer, N., "Time Spent on Dedicated Patient Care and Documentation Tasks Before and After the Introduction of a Structured and Standardized Electronic Health Record". In: *Applied Clinical Informatics* 09.01 (2018). ISSN: 1869-0327 (cited on pages 9, 10).

[27] Katan, M., Luft, A., "Global Burden of Stroke". In: *Seminars in Neurology*. Volume 38. 02. Thieme Medical Publishers, 2018 (cited on page 65).

Bibliography V

[28] Khurana, S., Joty, S. R., Ali, A., Glass, J., "A Factorial Deep Markov Model for Unsupervised Disentangled Representation Learning from Speech". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, United Kingdom: IEEE, 2019. ISBN: 978-1-4799-8131-1 (cited on pages 111, 112).

[29] Khurana, S., Laurent, A., Hsu, W.-N., Chorowski, J., Lancucki, A., Marxer, R., Glass, J., *A Convolutional Deep Markov Model for Unsupervised Speech Representation Learning*. 2020. arXiv: 2006.02547 (cited on pages 111, 112).

[30] Kingma, D. P., Welling, M., "Auto-Encoding Variational Bayes". In: *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. Banff, AB, Canada, 2014. arXiv: 1312.6114 (cited on page 33).

[31] Kyu, H. H., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., "Global, Regional, and National Disability-Adjusted Life-Years (DALYs) for 359 Diseases and Injuries and Healthy Life Expectancy (HALE) for 195 Countries and Territories, 1990–2017: A Systematic Analysis for the Global Burden of Disease Study 2017". In: *The Lancet* 392.10159 (2018) (cited on page 65).

[32] Lee, K., Lee, K., Lee, H., Shin, J., "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks". In: *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*. Montréal, Quebec, Canada, 2018 (cited on page 39).

[33] Ling, S., Liu, Y., "DeCoAR 2.0: Deep Contextualized Acoustic Representations with Vector Quantization". 2020. arXiv: 2012.06659 (cited on page 112).

[34] Liu, A. H., Chung, Y.-A., Glass, J., "Non-Autoregressive Predictive Coding for Learning Speech Representations from Local Dependencies". 2020. arXiv: 2011.00406 (cited on page 112).

Bibliography VI

corti DTU

[35] Liu, A. T., Yang, S.-w., Chi, P.-H., Hsu, P.-c., Lee, H.-y., "Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020 (cited on page 112).

[36] Maaløe, L., Fraccaro, M., Liévin, V., Winther, O., "BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling". In: *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*. Vancouver, Canada, 2019 (cited on pages 33, 101).

[37] Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., Lakshminarayanan, B., "Do Deep Generative Models Know What They Don't Know?" In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. New Orleans, LA, USA, 2019. arXiv: `1810.09136` (cited on page 40).

[38] Navi, B. B., Audebert, H. J., Alexandrov, A. W., Cadilhac, D. A., Grotta, J. C., PRESTO (Prehospital Stroke Treatment Organization) Writing Group, "Mobile Stroke Units: Evidence, Gaps, and next Steps". In: *Stroke* 53.6 (2022) (cited on page 65).

[39] Oord, A., Li, Y., Vinyals, O., *Representation Learning with Contrastive Predictive Coding*. 2018. arXiv: `1807.03748` (cited on page 112).

[40] Oord, A., Vinyals, O., Kavukcuoglu, K., "Neural Discrete Representation Learning". In: *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS)*. Long Beach, CA, USA, 2018 (cited on pages 111, 112).

[41] Oostema, J. A., Carle, T., Talia, N., Reeves, M., "Dispatcher Stroke Recognition Using a Stroke Screening Tool: A Systematic Review". In: *Cerebrovascular Diseases* 42.5-6 (2016) (cited on page 65).

Bibliography VII

[42] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., “Improving Language Understanding by Generative Pre-Training”. In: (2018) (cited on page 115).

[43] Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., Lakshminarayanan, B., “Likelihood Ratios for Out-of-Distribution Detection”. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*. Vancouver, Canada, 2019 (cited on page 39).

[44] Rezende, D. J., Mohamed, S., Wierstra, D., “Stochastic Backpropagation and Approximate Inference in Deep Generative Models”. In: *Proceedings of the 31st International Conference on Machine Learning (ICML)*. Volume 32. Beijing, China: PMLR, 2014 (cited on page 33).

[45] Schneider, S., Baevski, A., Collobert, R., Auli, M., “Wav2vec: Unsupervised Pre-training for Speech Recognition”. In: *Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech)*. Graz, Austria: ISCA, 2019. arXiv: 1904.05862 (cited on page 112).

[46] Serrà, J., Álvarez, D., Gómez, V., Slizovskaia, O., Núñez, J. F., Luque, J., “Input Complexity and Out-of-Distribution Detection with Likelihood-Based Generative Models”. In: *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. Addis Ababa, Ethiopia, 2020 (cited on page 39).

[47] Sinsky, C., Colligan, L., Li, L., Prgomet, M., Reynolds, S., Goeders, L., Westbrook, J., Tutty, M., Blike, G., “Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties”. In: *Annals of Internal Medicine* 165.11 (2016). ISSN: 1539-3704. pmid: 27595430 (cited on pages 9, 10).

Bibliography VIII

[48] Starmer, A. J., Spector, N. D., Srivastava, R., West, D. C., Rosenbluth, G., Allen, A. D., Noble, E. L., Tse, L. L., Dalal, A. K., Keohane, C. A., "Changes in Medical Errors after Implementation of a Handoff Program". In: *New England Journal of Medicine* 371.19 (2014) (cited on pages 7, 8).

[49] Tipping, M. D., Forth, V. E., O'Leary, K. J., Malkenson, D. M., Magill, D. B., Englert, K., Williams, M. V., "Where Did the Day Go?–A Time-Motion Study of Hospitalists". In: *Journal of Hospital Medicine* 5.6 (2010). ISSN: 1553-5606. pmid: 20803669 (cited on pages 9, 10).

[50] Turc, G., Bhogal, P., Fischer, U., Khatri, P., Lobotesis, K., Mazighi, M., Schellinger, P. D., Toni, D., De Vries, J., White, P., "European Stroke Organisation (ESO)-European Society for Minimally Invasive Neurological Therapy (ESMINT) Guidelines on Mechanical Thrombectomy in Acute Ischemic Stroke". In: *Journal of Neurointerventional Surgery* 11.8 (2019) (cited on page 65).

[51] Verma, R., Nalisnick, E., "Calibrated Learning to Defer with One-vs-All Classifiers". In: *International Conference on Machine Learning*. PMLR, 2022 (cited on page 82).

[52] Viereck, S., Møller, T. P., Iversen, H. K., Christensen, H., Lippert, F., "Medical Dispatchers Recognise Substantial Amount of Acute Stroke during Emergency Calls". In: *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 24 (2016) (cited on page 65).

[53] Xiao, Z., Yan, Q., Amit, Y., "Likelihood Regret: An Out-of-Distribution Detection Score for Variational Auto-Encoder". In: *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*. Virtual, 2020 (cited on page 39).

Reliability of machine learning systems

- Data: Quality, quantity, diversity, bias, privacy, ethics.
- Task: Context, domain, language, culture, purpose.

Reliability of machine learning systems
• Data: Quality, quantity, diversity, bias, privacy, ethics.
• Task: Context, domain, language, culture, purpose.

1. So what are the challenges holding us back in implementing such systems?
2.
   - Strong requirements of data.
   - Tasks that span different contexts, domains, languages, and cultures.
   - Regulatory requirements for transparency and accountability.
   - Trust and understanding of predictions.

Reliability of machine learning systems

corti DTU

- Data: Quality, quantity, diversity, bias, privacy, ethics.
- Task: Context, domain, language, culture, purpose.

- Interpretability of how a model works (transparency, accountability, regulation).
- Explainability of model predictions (trust, understanding, feedback).
- Fairness in treatment of different groups of people.
- Robustness to noise, outliers, distribution shift, and adversarial attacks.

Reliability of machine learning systems
- Data: Quality, quantity, diversity, bias, privacy, ethics.
- Task: Context, domain, language, culture, purpose.
- Interpretability of how a model works (transparency, accountability, regulation).
- Explainability of model predictions (trust, understanding, feedback).
- Fairness in treatment of different groups of people.
- Robustness to noise, outliers, distribution shift, and adversarial attacks.

1. So what are the challenges holding us back in implementing such systems?
2. – Strong requirements of data.
   – Tasks that span different contexts, domains, languages, and cultures.
   – Regulatory requirements for transparency and accountability.
   – Trust and understanding of predictions.

An alternative likelihood bound, $\mathcal{L}^{>k}$

corti DTU

An alternative version of the ELBO that only partially uses the approximate posterior can be written as [36]

$$\mathcal{L}^{>k}(\mathbf{x};\theta,\phi) = \mathbb{E}_{p_\theta(\mathbf{z}_{\leqslant k}|\mathbf{z}_{>k})q_\phi(\mathbf{z}_{>k}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z}_{>k})}{q_\phi(\mathbf{z}_{>k}|\mathbf{x})}\right] \quad (8)$$

Here, we have replaced the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ with a different proposal distribution that combines part of the approximate posterior with the conditional prior, namely

$$p_\theta(\mathbf{z}_{\leqslant k}|\mathbf{z}_{>k})q_\phi(\mathbf{z}_{>k}|\mathbf{x})$$

This bound uses the conditional prior for the lowest latent variables in the hierarchy.

EXTRA SLIDES
An alternative likelihood bound, $\mathcal{L}^{>k}$
An alternative version of the ELBO that only partially uses the approximate posterior can be written as [36]
$$\mathcal{L}^{>k}(\mathbf{x};\theta,\phi) = \mathbb{E}_{p_\theta(\mathbf{z}_{\leqslant k}|\mathbf{z}_{>k})q_\phi(\mathbf{z}_{>k}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z}_{>k})}{q_\phi(\mathbf{z}_{>k}|\mathbf{x})}\right] \quad (8)$$
Here, we have replaced the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ with a different proposal distribution that combines part of the approximate posterior with the conditional prior, namely
$$p_\theta(\mathbf{z}_{\leqslant k}|\mathbf{z}_{>k})q_\phi(\mathbf{z}_{>k}|\mathbf{x})$$
This bound uses the conditional prior for the lowest latent variables in the hierarchy.

1. So can we come up with a new bound that does not use the lowest latent variables in the same way?
2. So we could use this for OOD detection (as done in BIVA).

Likelihood ratios

We can use our new bound to compute the score used in a standard likelihood ratio test [9].

$$\mathrm{LLR}^{>k}(\mathbf{x}) \equiv \mathcal{L}(\mathbf{x}) - \mathcal{L}^{>k}(\mathbf{x}) \,. \tag{9}$$

We can inspect what this likelihood-ratio measures by considering the exact form of our bounds.

$$\begin{aligned} \mathcal{L} &= \log p_\theta(\mathbf{x}) - D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p_\theta(\mathbf{z}|\mathbf{x})\right), &(10)\\ \mathcal{L}^{>k} &= \log p_\theta(\mathbf{x}) - D_{\mathrm{KL}}\left(p_\theta(\mathbf{z}_{\leqslant}|\mathbf{z}_{>k}) q_\phi(\mathbf{z}_{>k}|\mathbf{x}) \,\|\, p_\theta(\mathbf{z}|\mathbf{x})\right) \,. \end{aligned}$$

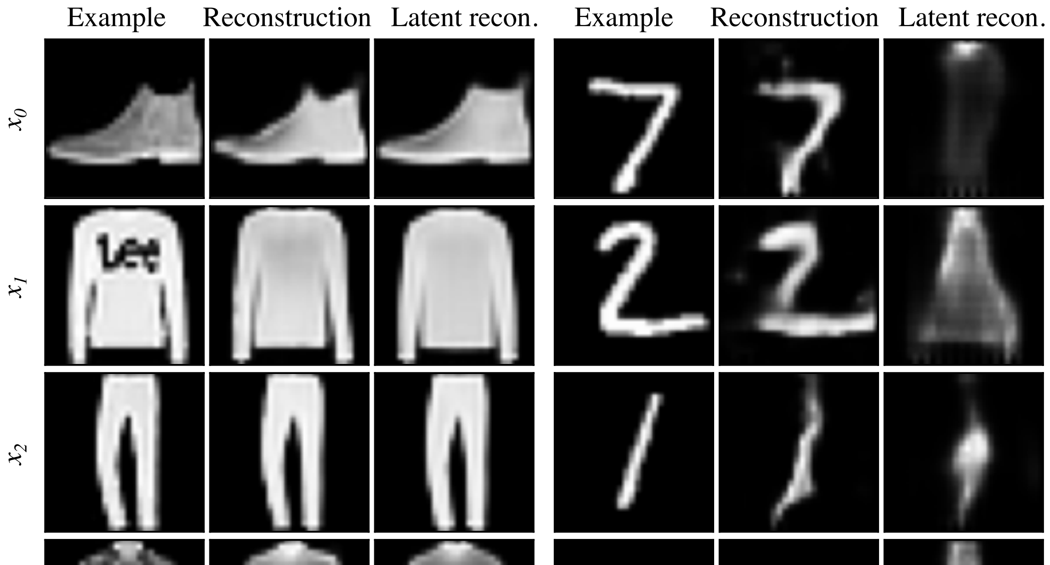
In the likelihood ratio the reconstruction terms cancel out and only the KL-divergences from the approximate to the true posterior remain.

$$\begin{aligned} \mathrm{LLR}^{>k}(\mathbf{x}) = &-D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p_\theta(\mathbf{z}|\mathbf{x})\right) &(11)\\ &+ D_{\mathrm{KL}}\left(p_\theta(\mathbf{z}_{\leqslant}|\mathbf{z}_{>k}) q_\phi(\mathbf{z}_{>k}|\mathbf{x}) \,\|\, p_\theta(\mathbf{z}|\mathbf{x})\right) \,. \end{aligned}$$

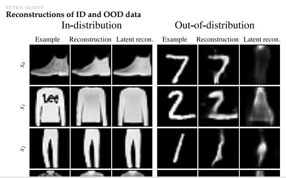
1. Write likelihood-ratio using the exact form of the bounds including intractable KL-divergence.

Reconstructions of ID and OOD data

In-distribution | Out-of-distribution

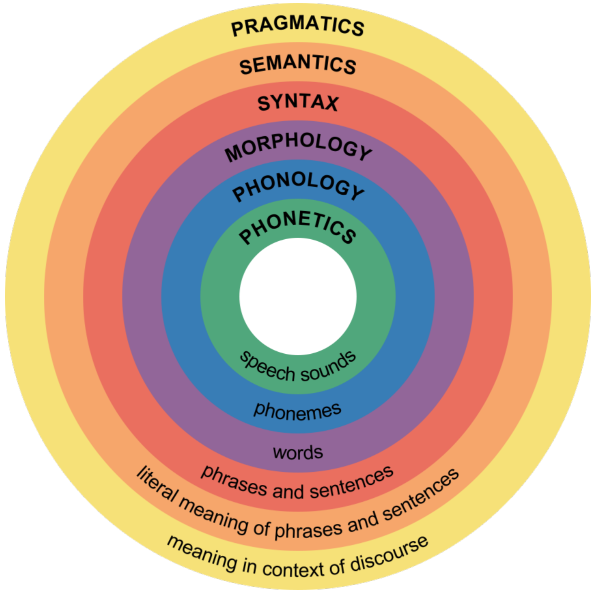


2024-03-04



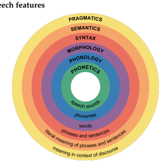
Hierarchy of speech features

corti DTU



2024-03-04

UNCERTAINTY AND THE MEDICAL INTERVIEW
└extra slides
 └hierarchical vaes know what they don't know
  └Hierarchy of speech features



Results on diverse datasets

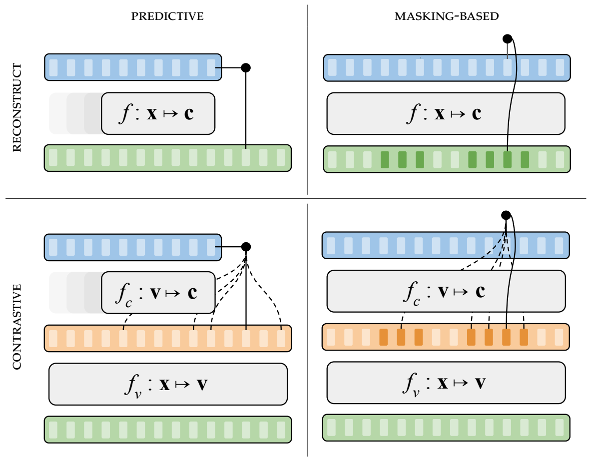
corti DTU

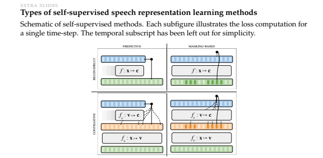
OOD dataset	Metric	AUROC↑	AUPRC↑	FPR80↓
Trained on CIFAR10				
SVHN	$\mathrm{LLR}^{>2}$	0.811	0.837	0.394
CIFAR10	$\mathrm{LLR}^{>1}$	0.469	0.479	0.835
Trained on SVHN				
CIFAR10	$\mathrm{LLR}^{>1}$	0.939	0.950	0.052
SVHN	$\mathrm{LLR}^{>1}$	0.489	0.484	0.799

OOD dataset	Metric	AUROC↑	AUPRC↑	FPR80↓
Trained on FashionMNIST				
MNIST	$\mathrm{LLR}^{>1}$	0.986	0.987	0.011
notMNIST	$\mathrm{LLR}^{>1}$	0.998	0.998	0.000
KMNIST	$\mathrm{LLR}^{>1}$	0.974	0.977	0.017
Omniglot28x28	$\mathrm{LLR}^{>2}$	1.000	1.000	0.000
Omniglot28x28Inverted	$\mathrm{LLR}^{>1}$	0.954	0.954	0.050
SmallNORB28x28	$\mathrm{LLR}^{>2}$	0.999	0.999	0.002
SmallNORB28x28Inverted	$\mathrm{LLR}^{>2}$	0.941	0.946	0.069
FashionMNIST	$\mathrm{LLR}^{>1}$	0.488	0.496	0.811
Trained on MNIST				
FashionMNIST	$\mathrm{LLR}^{>1}$	0.999	0.999	0.000
notMNIST	$\mathrm{LLR}^{>1}$	1.000	0.999	0.000
KMNIST	$\mathrm{LLR}^{>1}$	0.999	0.999	0.000
Omniglot28x28	$\mathrm{LLR}^{>1}$	1.000	1.000	0.000
Omniglot28x28Inverted	$\mathrm{LLR}^{>1}$	0.944	0.953	0.057
SmallNORB28x28	$\mathrm{LLR}^{>1}$	1.000	1.000	0.000
SmallNORB28x28Inverted	$\mathrm{LLR}^{>1}$	0.985	0.987	0.000
MNIST	$\mathrm{LLR}^{>2}$	0.515	0.507	0.792

Types of self-supervised speech representation learning methods

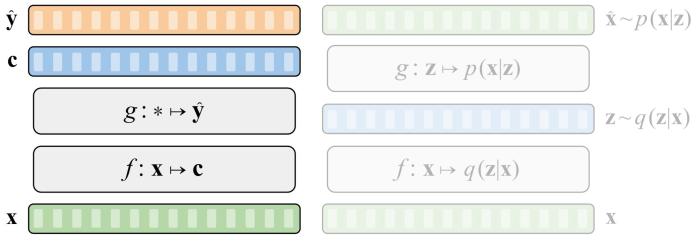
Schematic of self-supervised methods. Each subfigure illustrates the loss computation for a single time-step. The temporal subscript has been left out for simplicity.

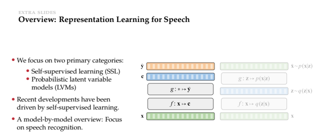




Overview: Representation Learning for Speech

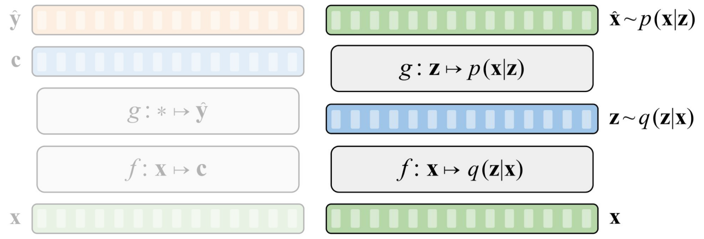
- We focus on two primary categories:
  - Self-supervised learning (SSL)
  - Probabilistic latent variable models (LVMs)
- Recent developments have been driven by self-supervised learning.
- A model-by-model overview: Focus on speech recognition.

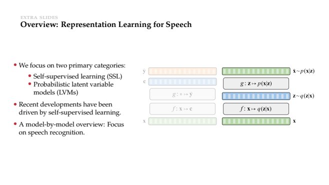




Overview: Representation Learning for Speech

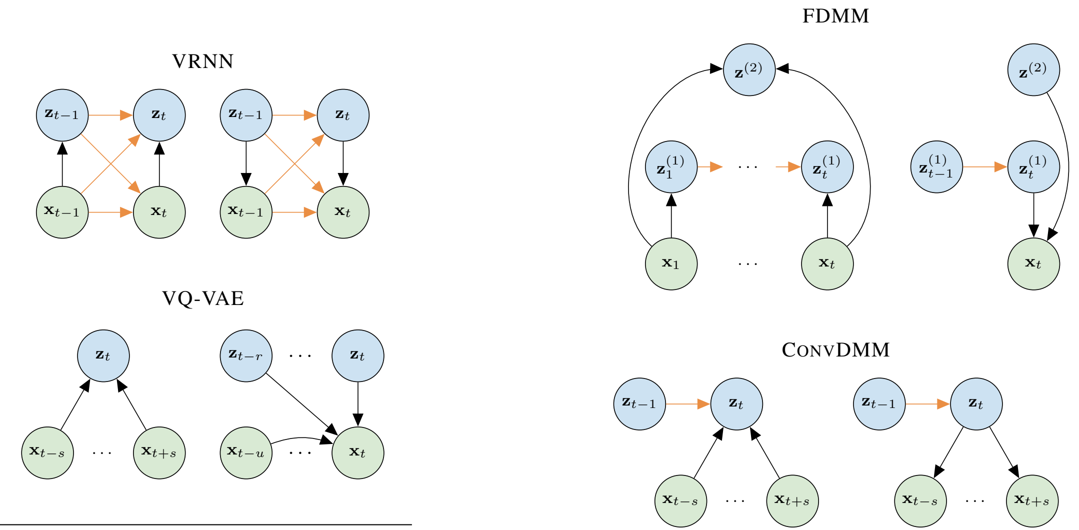
- We focus on two primary categories:
  - Self-supervised learning (SSL)
  - Probabilistic latent variable models (LVMs)
- Recent developments have been driven by self-supervised learning.
- A model-by-model overview: Focus on speech recognition.



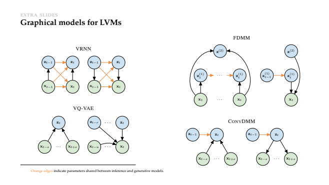


EXTRA SLIDES

Graphical models for LVMs



Orange edges indicate parameters shared between inference and generative models.



Overview of LVM probabilistic components

TYPE		FORM
	OBSERVATION MODEL	
ARX	Autoregressive on $\mathbf{x}_t$	$p(\mathbf{x}_t \mid \mathbf{x}_{1:t-1})$
LOC	Local latent variable	$p(\mathbf{x}_t \mid \mathbf{z}_{1:t})$
GLB	Global latent variable	$p(\mathbf{x}_t \mid \mathbf{z})$
	PRIOR	
ARX	Autoregressive on $\mathbf{x}_t$	$p(\mathbf{z}_t \mid \mathbf{x}_{1:t-1})$
ARZ	Autoregressive on $\mathbf{z}_t$	$p(\mathbf{z}_t \mid \mathbf{z}_{1:t-1})$
IND	Locally independent $\mathbf{z}_t$	$p(\mathbf{z}_t)$
GLB	Global latent variable	$p(\mathbf{z})$
	INFERENCE MODEL	
ARZ	Autoregressive on $\mathbf{z}_t$	$q(\mathbf{z}_t \mid \mathbf{z}_{1:t-1})$
FLT	Filtering	$q(\mathbf{z}_t \mid \mathbf{x}_{1:t})$
LSM	Local smoothing	$q(\mathbf{z}_t \mid \mathbf{x}_{t-r:t+r})$
GSM	Global smoothing	$q(\mathbf{z}_t \mid \mathbf{x}_{1:T})$
GLB	Global latent variable	$q(\mathbf{z} \mid \mathbf{x}_{1:T})$

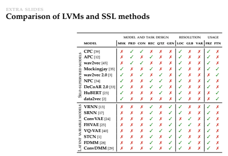
Classification of selected LVMs for speech

corti DTU

	Observation			Prior				Inference					
MODEL	ARX	LOC	GLB	ARX	ARZ	IND	GLB	ARZ	FLT	LSM	GSM	GLB	HIE
VRNN [13]	✓	✓	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗
SRNN [17]	✓	✓	✗	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗
HMM-VAE [15]	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✓
ConvVAE [24]	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓	✗
FHVAE [25]	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✓	✓	✓
VQ-VAE [40]	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗
BHMM-VAE [19]	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗
STCN [1]	✗	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
FDMM [28]	✗	✓	✓	✗	✓	✗	✓	✓	✓	✗	✗	✓	✓
ConvDMM [29]	✗	✓	✗	✗	✓	✗	✗	✓	✗	✓	✗	✗	✗

Comparison of LVMs and SSL methods

	MODEL	MODEL AND TASK DESIGN						RESOLUTION			USAGE	
		MSK	PRD	CON	REC	QTZ	GEN	LOC	GLB	VAR	FRZ	FTN
SELF-SUPERVISED MODELS	**CPC** [39]	✗	✓	✓	✗	✗	✗	✓	✗	✗	✓	✗
	APC [12]	✗	✓	✗	✓	✗	✗	✓	✗	✗	✓	✗
	wav2vec [45]	✗	✓	✓	✗	✗	✗	✓	✗	✗	✓	✗
	Mockingjay [35]	✓	✗	✗	✓	✗	✗	✓	✗	✗	✓	✓
	wav2vec 2.0 [3]	✓	✗	✓	✗	✓	✗	✓	✗	✗	✗	✓
	NPC [34]	✓	✗	✗	✓	✓	✗	✓	✗	✗	✓	✗
	DeCoAR 2.0 [33]	✓	✗	✗	✓	✓	✗	✓	✗	✗	✓	✗
	HuBERT [23]	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✓
	data2vec [2]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓
LATENT VARIABLE MODELS	**VRNN** [13]	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗
	SRNN [17]	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗
	ConvVAE [24]	✗	✗	✗	✓	✗	✓	✗	✓	✗	✓	✗
	FHVAE [25]	✗	✗	✗	✓	✗	✓	✓	✓	✗	✓	✗
	VQ-VAE [40]	✗	✗	✗	✓	✓	✓	✓	✗	✗	✓	✗
	STCN [1]	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗
	FDMM [28]	✗	✗	✗	✓	✗	✓	✓	✓	✗	✓	✗
	ConvDMM [29]	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗



Simulated prospective study

corti DTU

I. When is the model prediction presented to the call-taker?

  1. Notify the call-taker after the call ends.
  2. Notify the call-taker during the call.

II. How does prediction influence the diagnostic code the call-taker assigns to the call?

  A. Call-takers mirror model positives.
  B. Call-takers mirror model negatives.
  C. Call-takers mirror model predictions (corresponds to main results of the model itself).

To simulate the online scenario (2.), we stream the transcript to the model and make predictions every 50 words. A stroke positive is triggered only when three consecutive positive predictions are made. This is similar to the strategy implemented for a previous RCT on cardiac arrest [6].

EXTRA SLIDES
Simulated prospective study
I. When is the model prediction presented to the call-taker?
1. Notify the call-taker after the call ends.
2. Notify the call-taker during the call.
II. How does prediction influence the diagnostic code the call-taker assigns to the call?
A. Call-takers mirror model positives.
B. Call-takers mirror model negatives.
C. Call-takers mirror model predictions (corresponds to main results of the model itself).
To simulate the online scenario (2.), we stream the transcript to the model and make predictions every 50 words. A stroke positive is triggered only when three consecutive positive predictions are made. This is similar to the strategy implemented for a previous RCT on cardiac arrest [6].

Simulated prospective study

Predictor	Call-taker	Model		Call-taker supported by the model (simulated)			
When	During call	After call	During call	After call	During call	After call	During call
Method	-	-	-	neg → pos	neg → pos	pos → neg	pos → neg
F1-score [%] ↑	25.8 (23.7-27.9)	35.7 (35.0-36.4)	33.1 (32.4-33.7)	28.9 (28.3-29.5)	27.6 (27.0-28.1)	33.3 (32.5-34.1)	32.7 (31.8-33.5)
Sensitivity [%] ↑	52.7 (49.2-56.4)	63.0 (62.0-64.1)	58.7 (57.7-59.8)	72.4 (71.5-73.3)	72.3 (71.4-73.3)	43.4 (42.3-44.5)	39.1 (38.1-40.1)
PPV [%] ↑	17.1 (15.5-18.6)	24.9 (24.3-25.5)	23.0 (22.5-23.6)	18.0 (17.6-18.4)	17.0 (16.7-17.4)	27.0 (26.3-27.8)	28.1 (27.3-28.9)
FOR [%] ↓ (1 - NPV)	0.105 (0.094-0.116)	0.082 (0.079-0.085)	0.091 (0.088-0.094)	0.061 (0.059-0.064)	0.061 (0.059-0.064)	0.125 (0.121-0.129)	0.134 (0.131-0.138)
FPR [%] ↓ (1 - specificity)	0.565 (0.539-0.590)	0.419 (0.413-0.426)	0.432 (0.426-0.439)	0.726 (0.717-0.735)	0.776 (0.767-0.786)	0.258 (0.253-0.263)	0.221 (0.216-0.226)

Fine-tuning a large language model

corti DTU

- Large language models are effective in a wide range of NLP tasks [14, 42].
- Might BERT be useful for recognizing stroke?

Subset	Predictor	F1-score [%] ↑	Sensitivity [%] ↑	PPV [%] ↑	FOR [%] ↓ (1 - NPV)	FPR [%] ↓ (1 - specificity)
Overall	**Call-takers**	25.8 (23.7-27.9)	52.7 (49.2-56.4)	17.1 (15.5-18.6)	0.105 (0.094-0.116)	0.565 (0.539-0.590)
	MLP	35.7 (35.0-36.4)	63.0 (62.0-64.1)	24.9 (24.3-25.5)	0.082 (0.079-0.085)	0.419 (0.413-0.426)
	BERT (fine-tuned)	33.8 (31.5-36.2)	57.5 (53.9-60.9)	23.9 (21.9-25.9)	0.094 (0.084-0.104)	0.403 (0.381-0.424)

Table of contents I

EXTRA SLIDES
Table of contents I
• project
• introduction
• hierarchical vaes know what they don’t know
• overview
• a brief overview of unsupervised speech representation learning
• a retrospective study on machine learning-assisted stroke recognition for medical helpline calls
• discussion
• bibliography
• extra slides
• introduction
• hierarchical vaes know what they don’t know
• a brief overview of unsupervised speech representation learning
• a retrospective study on machine learning-assisted stroke recognition for medical helpline calls