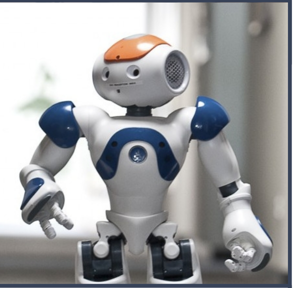
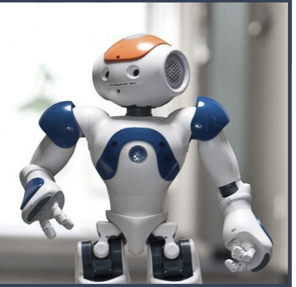
Long story short: the summary of (more than) a decade of probabilistic audio-visual learning

Xavier Alameda-Pineda and Radu Horaud
(and a long list of great people)



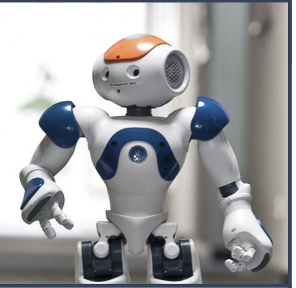
Why Probabilistic Audio-Visual Learning?



Why Probabilistic Audio-Visual Learning?

1. Scientific Challenges:

- Captured with different sensors
- Represent different phenomena
- Have different statistical patterns



Why Probabilistic Audio-Visual Learning?

1. Scientific Challenges:

- Captured with different sensors
- Represent different phenomena
- Have different statistical patterns

2. From a technological perspective:

- Sensors are mature
- Embedded in numerous devices
- Cheap perception solution

Why Probabilistic Audio-Visual Learning?

1. Scientific Challenges:

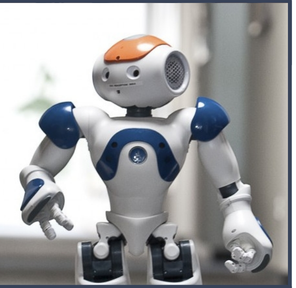
- Captured with different sensors
- Represent different phenomena
- Have different statistical patterns

2. From a technological perspective:

- Sensors are mature
- Embedded in numerous devices
- Cheap perception solution

3. Probabilistic Learning:

- Usually unsupervised
- Learn properties of noise/clutter
- Infer latent variables



Why Probabilistic Audio-Visual Learning?

1. Scientific Challenges:

- Captured with different sensors
- Represent different phenomena
- Have different statistical patterns

2. From a technological perspective:

- Sensors are mature
- Embedded in numerous devices
- Cheap perception solution

3. Probabilistic Learning:

- Usually unsupervised
- Learn properties of noise/clutter
- Infer latent variables

4. Potential impact:

- Behavior analysis
- Robotic social interaction
- Healthcare, training, security, ...

Probabilistic Audio-Visual Learning Setting

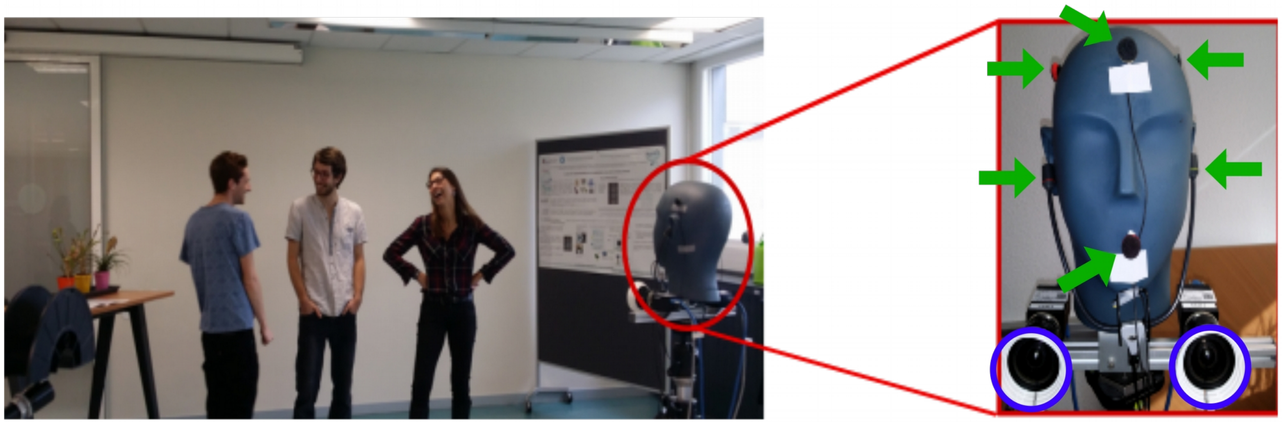
Probabilistic Audio-Visual Learning Setting

Picture a social scene with multiple people chatting and interacting.

Probabilistic Audio-Visual Learning Setting

Picture a social scene with multiple people chatting and interacting.

A **device** observes de scene with **microphones** and **cameras**.

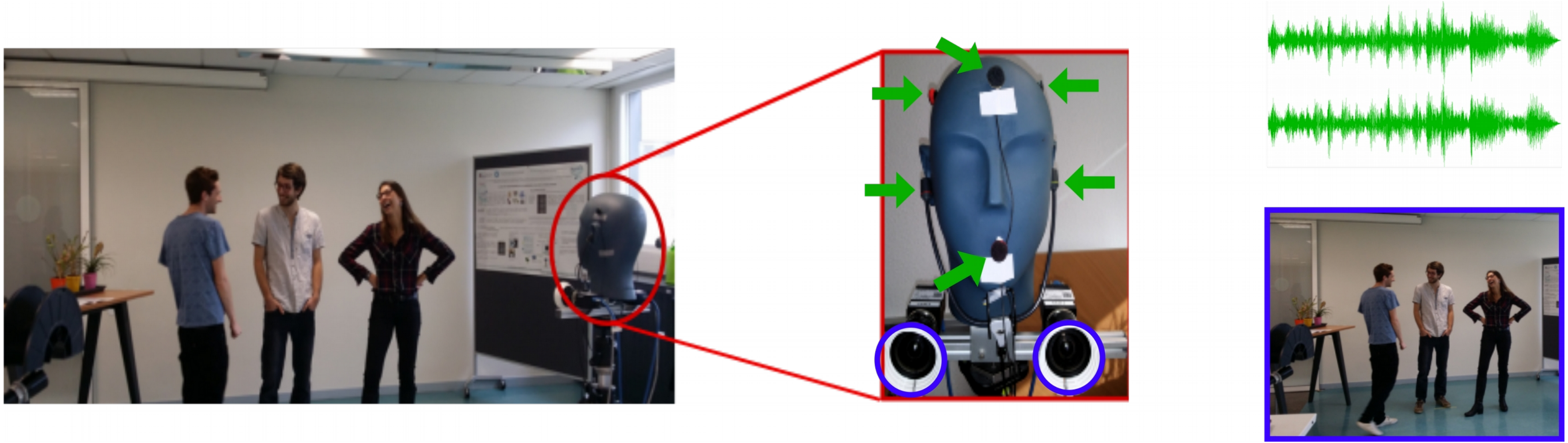


Probabilistic Audio-Visual Learning Setting

Picture a social scene with multiple people chatting and interacting.

A **device** observes de scene with **microphones** and **cameras**.

We obtain **auditory (sounds)** and **visual (images)** features.



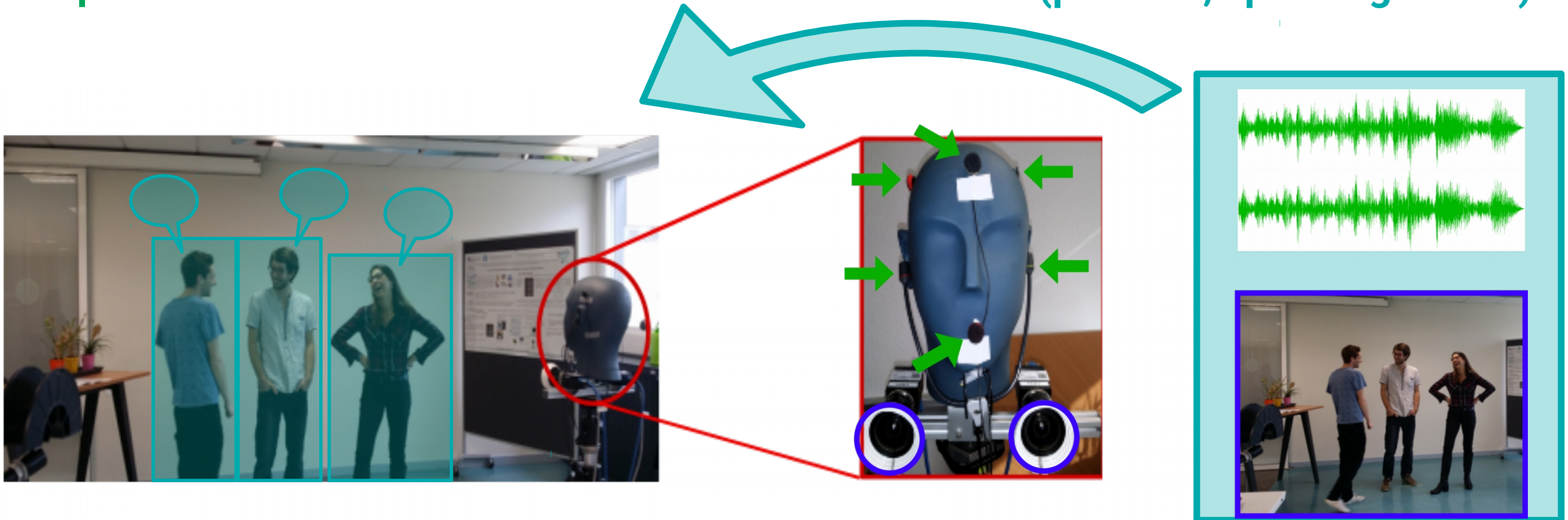
Probabilistic Audio-Visual Learning Setting

Picture a social scene with multiple people chatting and interacting.

A **device** observes de scene with **microphones** and **cameras**.

We obtain **auditory (sounds)** and **visual (images)** features.

We would like to infer **latent variables (position, speaking status)**.



Well, OK, but how?

What is the methodology?

(Apologies if the next slides are a bit dense)

Unsupervised Probabilistic Learning

Unsupervised Probabilistic Learning

Observations will be denoted by $\mathbf{x}^a$ and $\mathbf{x}^v$
Latent variables by $\mathbf{z}$

We need to set up a probabilistic
model parametrised by the set $\boldsymbol{\theta}$

$$p_{\boldsymbol{\theta}}(\mathbf{x}^a, \mathbf{x}^v|\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{z})$$

Unsupervised Probabilistic Learning

Observations will be denoted by $\mathbf{x}^a$ and $\mathbf{x}^v$
Latent variables by $\mathbf{z}$

We need to set up a probabilistic model parametrised by the set $\boldsymbol{\theta}$

$$p_{\boldsymbol{\theta}}(\mathbf{x}^a, \mathbf{x}^v | \mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{z})$$

Learning ↔ maximum likelihood:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{x}^a, \mathbf{x}^v)$$

Inference ↔ expected value (or mode)

$$\mathbf{z}^* = \mathbb{E}_{p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x}^a, \mathbf{x}^v)}\{\mathbf{z}\}$$

Unsupervised Probabilistic Learning

Observations will be denoted by $\mathbf{x}^a$ and $\mathbf{x}^v$
Latent variables by $\mathbf{z}$

We need to set up a probabilistic model parametrised by the set $\boldsymbol{\theta}$

$$p_{\boldsymbol{\theta}}(\mathbf{x}^a, \mathbf{x}^v | \mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{z})$$

Learning ↔ maximum likelihood:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{x}^a, \mathbf{x}^v)$$

Inference ↔ expected value (or mode)

$$\mathbf{z}^* = \mathbb{E}_{p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x}^a, \mathbf{x}^v)}\{\mathbf{z}\}$$

Examples: Gaussian mixture models, hidden Markov models, conditional random fields, linear dynamical systems (Kalman filter), probabilistic PCA, variational autoencoders (and dynamical ones), normalising flow, diffusion models, ...

How to learn and infer? [v1 – Exact EM]

How to learn and infer? [v1 – Exact EM]

Direct optimisation not analytically solvable:

$$\arg\max_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{x})$$

Optimise the expected complete-data log-like:

$$\arg\max_{\boldsymbol{\theta}} \mathbb{E}_{p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})}\{\log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})\}$$

How to learn and infer? [v1 – Exact EM]

Direct optimisation not analytically solvable:

$$\arg\max_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{x})$$

Optimise the expected complete-data log-like:

$$\arg\max_{\boldsymbol{\theta}} \mathbb{E}_{p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})}\{\log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})\}$$

Expectation: given $\boldsymbol{\theta}^*$, compute:

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \mathbb{E}_{p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})}\{\log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})\}$$

Maximisation: set up the new $\boldsymbol{\theta}^*$ to:

$$\boldsymbol{\theta}^* \leftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$$

How to learn and infer? [v1 – Exact EM]

Direct optimisation not analytically solvable:

$$\arg\max_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{x})$$

Optimise the expected complete-data log-like:

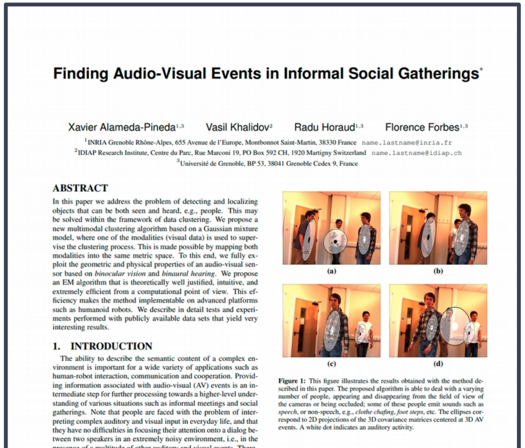
$$\arg\max_{\boldsymbol{\theta}} \mathbb{E}_{p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})}\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\}$$

Expectation: given $\boldsymbol{\theta}^*$, compute:

$$\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^*) = \mathbb{E}_{p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})}\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\}$$

Maximisation: set up the new $\boldsymbol{\theta}^*$ to:

$$\boldsymbol{\theta}^* \leftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$$



Finding Audio-Visual Events in Informal Social Gatherings*

Xavier Alameda-Pineda[1,3] Vasil Khalidov[2] Radu Horaud[1,3] Florence Forbes[1,3]

[1]INRIA Grenoble Rhône-Alpes, 655 Avenue de l'Europe, Montbonnot Saint-Martin, 38330 France name.lastname@inria.fr
[2]IDIAP Research Institute, Centre du Parc, Rue Marconi 19, PO Box 592 CH, 1920 Martigny Switzerland name.lastname@idiap.ch
[3]Université de Grenoble, BP 53, 38041 Grenoble Cedex 9, France

ABSTRACT

In this paper we address the problem of detecting and localizing objects that can be both seen and heard, e.g., people. This may be solved within the framework of data clustering. We propose a new multimodal clustering algorithm based on a Gaussian mixture model, where one of the modalities (visual data) is used to supervise the clustering process. This is made possible by mapping both modalities into the same metric space. To this end, we fully exploit the geometric and physical properties of an audio-visual sensor based on *binocular vision* and *binaural hearing*. We propose an EM algorithm that is theoretically well justified, intuitive, and extremely efficient from a computational point of view. This efficiency makes the method implementable on advanced platforms such as humanoid robots. We describe in detail tests and experiments performed with publicly available data sets that yield very interesting results.

1. INTRODUCTION

The ability to describe the semantic content of a complex environment is important for a wide variety of applications such as human-robot interaction, communication and cooperation. Providing information associated with audio-visual (AV) events is an intermediate step for further processing towards a higher-level understanding of various situations such as informal meetings and social gatherings. Note that people are faced with the problem of interpreting complex auditory and visual input in everyday life, and that they have no difficulties in focusing their attention onto a dialog between two speakers in an extremely noisy environment, i.e., in the

Figure 1: This figure illustrates the results obtained with the method described in this paper. The proposed algorithm is able to deal with a varying number of people, appearing and disappearing from the field of view of the cameras or being occluded; some of these people emit sounds such as *speech*, or non-speech, e.g., *clothe chafing*, *foot steps*, etc. The ellipses correspond to 2D projections of the 3D covariance matrices centered at 3D AV events. A white dot indicates an auditory activity.

Finding Audio-Visual Events in Informal Social Gatherings, ICMI 2011 – **Outstanding Paper Award**

How to learn and infer? [v1 – Exact EM]

Direct optimisation not analytically solvable:

$$\arg\max_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{x})$$

Optimise the expected complete-data log-like:

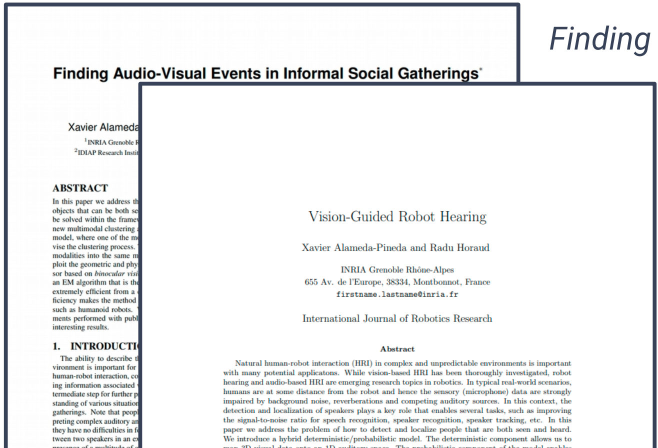
$$\arg\max_{\boldsymbol{\theta}} \mathbb{E}_{p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})}\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\}$$

Expectation: given θ^*, compute:

$$\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^*) = \mathbb{E}_{p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})}\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\}$$

Maximisation: set up the new θ^* to:

$$\boldsymbol{\theta}^* \leftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$$



Finding Audio-Visual Events in Informal Social Gatherings*

Xavier Alameda

[1]INRIA Grenoble R
[2]IDIAP Research Instit

ABSTRACT

In this paper we address th
objects that can be both se
be solved within the framew
new multimodal clustering a
model, where one of the mo
vise the clustering process.
modalities into the same m
ploit the geometric and phy
sor based on *binocular visi*
an EM algorithm that is the
extremely efficient from a
ficiency makes the method
such as humanoid robots.
ments performed with publ
interesting results.

1. INTRODUCTI

The ability to describe t
vironment is important for
human-robot interaction, co
ing information associated
termediate step for further p
standing of various situation
gatherings. Note that peopl
preting complex auditory an
they have no difficulties in f
tween two speakers in an ex

Vision-Guided Robot Hearing

Xavier Alameda-Pineda and Radu Horaud

INRIA Grenoble Rhône-Alpes
655 Av. de l'Europe, 38334, Montbonnot, France
`firstname.lastname@inria.fr`

International Journal of Robotics Research

Abstract

Natural human-robot interaction (HRI) in complex and unpredictable environments is important with many potential applicatons. While vision-based HRI has been thoroughly investigated, robot hearing and audio-based HRI are emerging research topics in robotics. In typical real-world scenarios, humans are at some distance from the robot and hence the sensory (microphone) data are strongly impaired by background noise, reverberations and competing auditory sources. In this context, the detection and localization of speakers plays a key role that enables several tasks, such as improving the signal-to-noise ratio for speech recognition, speaker recognition, speaker tracking, etc. In this paper we address the problem of how to detect and localize people that are both seen and heard. We introduce a hybrid deterministic/probabilistic model. The deterministic component allows us to

Finding Audio-Visual Events in Informal Social Gatherings, ICMI 2011 – **Outstanding Paper Award**

Visually-Guided Robot Hearing, IJRR 2012

How to learn and infer? [v1 – Exact EM]

Direct optimisation not analytically solvable:

$$\arg\max_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{x})$$

Optimise the expected complete-data log-like:

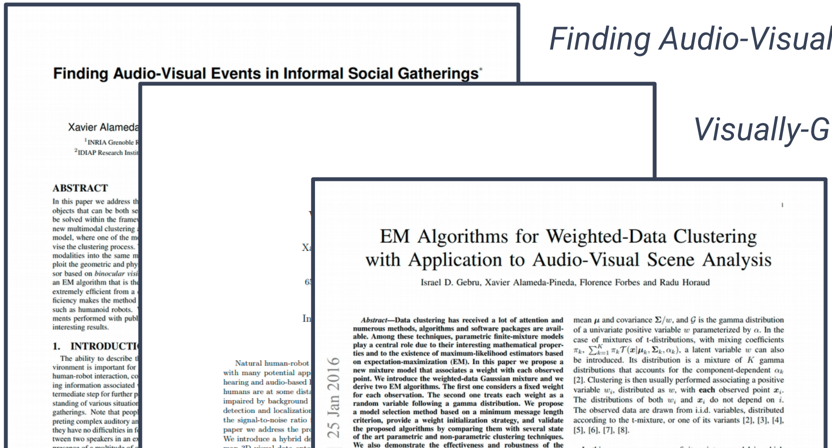
$$\arg\max_{\boldsymbol{\theta}} \mathbb{E}_{p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})}\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\}$$

Expectation: given θ^*, compute:

$$\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^*) = \mathbb{E}_{p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})}\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\}$$

Maximisation: set up the new θ^* to:

$$\boldsymbol{\theta}^* \leftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$$



Finding Audio-Visual Events in Informal Social Gatherings*

EM Algorithms for Weighted-Data Clustering with Application to Audio-Visual Scene Analysis

Israel D. Gebru, Xavier Alameda-Pineda, Florence Forbes and Radu Horaud

25 Jan 2016

Abstract—Data clustering has received a lot of attention and numerous methods, algorithms and software packages are available. Among these techniques, parametric finite-mixture models play a central role due to their interesting mathematical properties and to the existence of maximum-likelihood estimators based on expectation-maximization (EM). In this paper we propose a new mixture model that associates a weight with each observed point. We introduce the weighted-data Gaussian mixture and we derive two EM algorithms. The first one considers a fixed weight for each observation. The second one treats each weight as a random variable following a gamma distribution. We propose a model selection method based on a minimum message length criterion, provide a weight initialization strategy, and validate the proposed algorithms by comparing them with several state of the art parametric and non-parametric clustering techniques. We also demonstrate the effectiveness and robustness of the

mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}/w$, and $\mathcal{G}$ is the gamma distribution of a univariate positive variable w parameterized by α. In the case of mixtures of t-distributions, with mixing coefficients π_k, $\sum_{k=1}^K \pi_k \mathcal{T}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \alpha_k)$, a latent variable w can also be introduced. Its distribution is a mixture of K gamma distributions that accounts for the component-dependent α_k [2]. Clustering is then usually performed associating a positive variable w_i, distributed as w, with **each** observed point $\boldsymbol{x}_i$. The distributions of both w_i and $\boldsymbol{x}_i$ do not depend on i. The observed data are drawn from i.i.d. variables, distributed according to the t-mixture, or one of its variants [2], [3], [4], [5], [6], [7], [8].

Finding Audio-Visual Events in Informal Social Gatherings, ICMI 2011 – **Outstanding Paper Award**

Visually-Guided Robot Hearing, IJRR 2012

EM Algorithms for Weighted-Data Clustering for AV Scene Analysis, TPAMI 2016.

How to learn and infer? [v1 – Exact EM]

Direct optimisation not analytically solvable:

$$\arg\max_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{x})$$

Optimise the expected complete-data log-like:

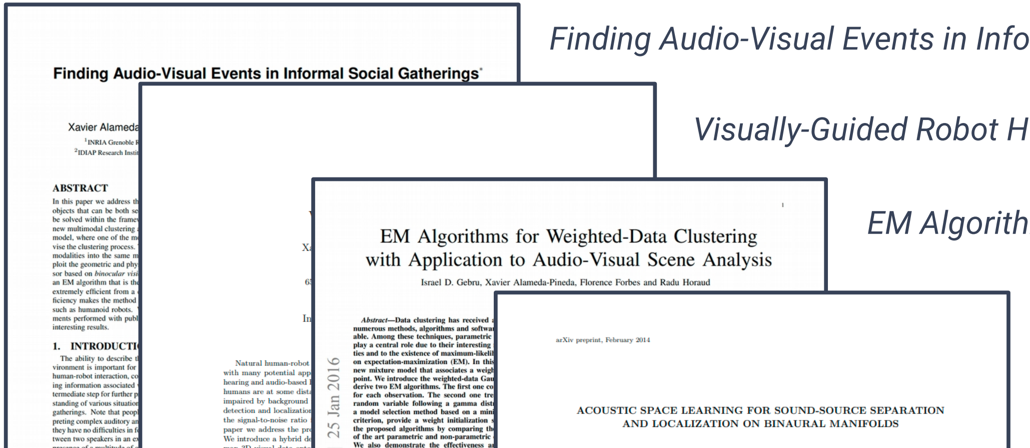
$$\arg\max_{\boldsymbol{\theta}} \mathbb{E}_{p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})}\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\}$$

Expectation: given $\boldsymbol{\theta}^*$, compute:

$$\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^*) = \mathbb{E}_{p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})}\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\}$$

Maximisation: set up the new $\boldsymbol{\theta}^*$ to:

$$\boldsymbol{\theta}^* \leftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$$



Finding Audio-Visual Events in Informal Social Gatherings*

EM Algorithms for Weighted-Data Clustering with Application to Audio-Visual Scene Analysis

Israel D. Gebru, Xavier Alameda-Pineda, Florence Forbes and Radu Horaud

25 Jan 2016

arXiv preprint, February 2014

ACOUSTIC SPACE LEARNING FOR SOUND-SOURCE SEPARATION AND LOCALIZATION ON BINAURAL MANIFOLDS

Finding Audio-Visual Events in Informal Social Gatherings, ICMI 2011 – **Outstanding Paper Award**

Visually-Guided Robot Hearing, IJRR 2012

EM Algorithms for Weighted-Data Clustering for AV Scene Analysis, TPAMI 2016.

Acoustic Space Learning for Sound-source Separation and Localization on Binaural Manifolds, Neural Systems, 2015 – **Hojjat Adeli Award for Outstanding Contributions in Neural Systems**

How to learn and infer? [v1 – Exact EM]

Direct optimisation not analytically solvable:

$$\arg\max_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{x})$$

Optimise the expected complete-data log-like:

$$\arg\max_{\boldsymbol{\theta}} \mathbb{E}_{p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})}\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\}$$

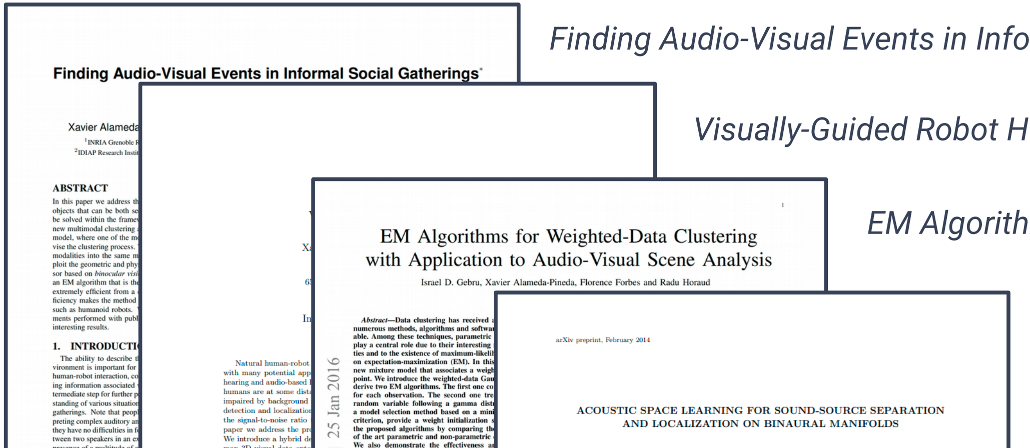
Expectation: given θ^*, compute:

$$\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^*) = \mathbb{E}_{\boxed{p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})}}\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\}$$

What if we cannot compute the posterior?

Maximisation: set up the new θ^* to:

$$\boldsymbol{\theta}^* \leftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$$



Finding Audio-Visual Events in Informal Social Gatherings*

Xavier Alameda

EM Algorithms for Weighted-Data Clustering with Application to Audio-Visual Scene Analysis

Israel D. Gebru, Xavier Alameda-Pineda, Florence Forbes and Radu Horaud

25 Jan 2016

arXiv preprint, February 2014

ACOUSTIC SPACE LEARNING FOR SOUND-SOURCE SEPARATION AND LOCALIZATION ON BINAURAL MANIFOLDS

Finding Audio-Visual Events in Informal Social Gatherings, ICMI 2011 – **Outstanding Paper Award**

Visually-Guided Robot Hearing, IJRR 2012

EM Algorithms for Weighted-Data Clustering for AV Scene Analysis, TPAMI 2016.

Acoustic Space Learning for Sound-source Separation and Localization on Binaural Manifolds, Neural Systems, 2015 – **Hojjat Adeli Award for Outstanding Contributions in Neural Systems**

How to learn and infer? [v2 – Variational EM]

Close look to the log-likelihood, for any distribution :

How to learn and infer? [v2 – Variational EM]

Close look to the log-likelihood, for any distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{q_\phi}(\mathbf{z}|\mathbf{x})\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\} + \mathcal{H}_{q_{\boldsymbol{\phi}}} + \mathcal{D}_{\mathsf{KL}}(q_{\boldsymbol{\phi}}||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))$$

How to learn and infer? [v2 – Variational EM]

Close look to the log-likelihood, for any distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{q_{\phi}}(\mathbf{z}|\mathbf{x})\{\log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})\} + \mathcal{H}_{q_{\phi}} + \mathcal{D}_{\mathsf{KL}}(q_{\boldsymbol{\phi}} \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))$$

In the exact EM, $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$ and $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is a tight lower bound. In some cases (not too complex), $p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$ is computationally intractable. Break the hidden in two (or +).

How to learn and infer? [v2 – Variational EM]

Close look to the log-likelihood, for any distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{q_\phi}(\mathbf{z}|\mathbf{x})\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\} + \mathcal{H}_{q_\phi} + \mathcal{D}_{\mathsf{KL}}(q_{\boldsymbol{\phi}} \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))$$

In the exact EM, $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$ and $\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ is a tight lower bound. In some cases (not too complex), $p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$ is computationally intractable. Break the hidden in two (or +).

$\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$ and $q_{\boldsymbol{\phi}}(\mathbf{z}) = q_{\boldsymbol{\phi}_1}(\mathbf{z}_1) q_{\boldsymbol{\phi}_2}(\mathbf{z}_2)$ $\quad q^*_{\boldsymbol{\phi}_1}, q^*_{\boldsymbol{\phi}_2} = \arg\min \mathcal{D}_{\mathsf{KL}}(q_{\boldsymbol{\phi}_1} q_{\boldsymbol{\phi}_2} \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))$

How to learn and infer? [v2 – Variational EM]

Close look to the log-likelihood, for any distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{q_{\phi}}(\mathbf{z}|\mathbf{x})\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\} + \mathcal{H}_{q_{\phi}} + \mathcal{D}_{\mathsf{KL}}(q_{\boldsymbol{\phi}}\|p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))$$

In the exact EM, $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$ and $\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ is a tight lower bound. In some cases (not too complex), $p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$ is computationally intractable. Break the hidden in two (or +).

$$\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2) \quad \text{and} \quad q_{\boldsymbol{\phi}}(\mathbf{z}) = q_{\boldsymbol{\phi}_1}(\mathbf{z}_1)q_{\boldsymbol{\phi}_2}(\mathbf{z}_2) \qquad q^*_{\boldsymbol{\phi}_1}, q^*_{\boldsymbol{\phi}_2} = \arg\min \mathcal{D}_{\mathsf{KL}}(q_{\boldsymbol{\phi}_1}q_{\boldsymbol{\phi}_2}\|p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))$$

A variational EM algorithm for the separation of moving sound sources, IEEE WASPAA 2015 – **Best Student Paper Award**

2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics October 18-21, 2015, New Paltz, NY

A VARIATIONAL EM ALGORITHM FOR THE SEPARATION OF MOVING SOUND SOURCES

Dionyssos Kounades-Bastian[1], *Laurent Girin*[1,2], *Xavier Alameda-Pineda*[3], *Sharon Gannot*[4], *Radu Horaud*[1]

[1] INRIA Grenoble Rhône-Alpes, France
[2] Univ. Grenoble Alpes, GIPSA-lab, France
[3] University of Trento, Dept. Information Ing. Comp. Sc., Italy
[4] Bar-Ilan University, Faculty of Engineering, Israel

ABSTRACT

This paper addresses the problem of separation of moving sound sources. We propose a probabilistic framework based on the complex Gaussian model combined with non-negative matrix factorization. The properties associated with moving sources are modeled using time-varying mixing filters described by a stochastic temporal process. We present a variational expectation-maximization (VEM) algorithm that employs a Kalman smoother to estimate the mixing filters. The sound sources are separated by means of Wiener filters, built from the estimators provided by the proposed VEM algorithm. Preliminary experiments with simulated data show that, while for static sources we obtain results comparable with the baseline method [1], in the case of moving source our method outperforms a piece-wise version of the baseline method.

Index Terms— Audio-source separation, time-varying mixing filters, moving sources, Kalman smoother, variational EM.

1. INTRODUCTION

Audio-source separation methods aim at recovering J unobserved source signals, $\mathbf{s} = [s_1, \ldots, s_J]^\top$ ($^\top$ denotes the transpose operator), from I observed mixed signals $\mathbf{x} = [x_1, \ldots, x_I]^\top$. A large body of literature deals with various source-separation configuration problems and their associated methods [2]. In this paper we consider the difficult case of convolutive mixtures of moving audio sources, i.e., the source-to-microphone channels are modeled with time-varying linear filters, thus taking into account possible motions of the sources and/or of the sensors (as may be the case in, e.g., human-robot interaction scenarios). Moreover the mixtures can be possibly underdetermined, i.e. we may have $I < J$.

To address this difficult problem, we focus on probabilistic methods based on complex-valued Gaussian models of source signals in the time-frequency (TF) domain, as initially proposed in [3]

associated estimation algorithm, based on the complex-Gaussian and NMF models and able to separate sound sources convolved with time-varying filters. Modeling convolutive mixtures with time-varying filters was already proposed in, e.g., [9, 12]. However, up to our knowledge, this is the first attempt to incorporate a latent-continuous model for the time-varying mixing filters in the TF-domain complex-Gaussian framework. Unlike [9], where the mixing system is parameterized with the angle of arrival, ruled by a discrete temporal model, our mixing model uses a more general propagation regime which is expected to be more suitable to reverberant environments. Moreover, [9] relies on binary masking for separating the sources, which is known to introduce speech distortion, whereas we use the more general Wiener filtering.

The paper is organized as follows. Section 2 describes the source model and introduces the proposed mixing model. In Section 3, we present a variational EM (VEM) algorithm for both the estimation of model parameters and inference of latent variables, in batch mode. A first series of experiments is reported in Section 4. Conclusions and future work are depicted in Section 5.

2. SOUND MIXTURES WITH TIME-VARYING FILTERS

2.1. The Source Model

Assuming that we work in the TF domain, as a result of applying the short-time Fourier transform (STFT) to the time-domain signals, the following notations are introduced: $f \in [1, F]$ denotes the frequency bin index, $\ell \in [1, L]$ denotes the time frame index, $\{\mathcal{K}_j\}_j$ denotes a non-trivial partition of $\{1, \ldots, K\}$, $K \geq J$ (K_j denotes the cardinal of $\mathcal{K}_j$). Following [1], each source $s_{j,f\ell}$ at TF bin (f, ℓ) is modeled as the sum of K_j latent components $c_{k,f\ell}$, $k \in \mathcal{K}_j$, namely:

$$s_{j,f\ell} = \sum_{k\in\mathcal{K}_j} c_{k,f\ell} \Leftrightarrow \mathbf{s}_{f\ell} = \mathbf{G}\mathbf{c}_{f\ell}, \tag{1}$$

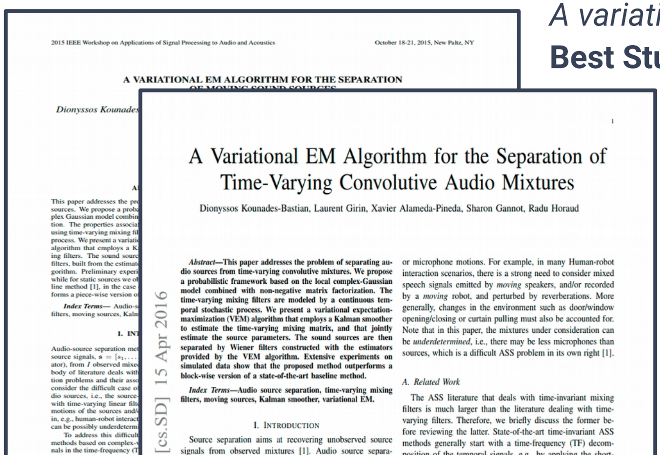
How to learn and infer? [v2 – Variational EM]

Close look to the log-likelihood, for any distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{q_\phi}(\mathbf{z}|\mathbf{x})\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\} + \mathcal{H}_{q_\phi} + \mathcal{D}_{\mathsf{KL}}(q_{\boldsymbol{\phi}} \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))$$

In the exact EM, $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$ and $\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ is a tight lower bound. In some cases (not too complex), $p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$ is computationally intractable. Break the hidden in two (or +).

$$\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2) \quad \text{and} \quad q_{\boldsymbol{\phi}}(\mathbf{z}) = q_{\boldsymbol{\phi}_1}(\mathbf{z}_1) q_{\boldsymbol{\phi}_2}(\mathbf{z}_2) \qquad q^*_{\boldsymbol{\phi}_1}, q^*_{\boldsymbol{\phi}_2} = \arg\min \mathcal{D}_{\mathsf{KL}}(q_{\boldsymbol{\phi}_1} q_{\boldsymbol{\phi}_2} \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))$$



2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics October 18-21, 2015, New Paltz, NY

A VARIATIONAL EM ALGORITHM FOR THE SEPARATION

Dionyssos Kounades

1

A Variational EM Algorithm for the Separation of Time-Varying Convolutive Audio Mixtures

Dionyssos Kounades-Bastian, Laurent Girin, Xavier Alameda-Pineda, Sharon Gannot, Radu Horaud

[cs.SD] 15 Apr 2016

***Abstract*—This paper addresses the problem of separating audio sources from time-varying convolutive mixtures. We propose a probabilistic framework based on the local complex-Gaussian model combined with non-negative matrix factorization. The time-varying mixing filters are modeled by a continuous temporal stochastic process. We present a variational expectation-maximization (VEM) algorithm that employs a Kalman smoother to estimate the time-varying mixing matrix, and that jointly estimate the source parameters. The sound sources are then separated by Wiener filters constructed with the estimators provided by the VEM algorithm. Extensive experiments on simulated data show that the proposed method outperforms a block-wise version of a state-of-the-art baseline method.**

***Index Terms*—Audio source separation, time-varying mixing filters, moving sources, Kalman smoother, variational EM.**

I. Introduction

Source separation aims at recovering unobserved source signals from observed mixtures [1]. Audio source separa-

or microphone motions. For example, in many Human-robot interaction scenarios, there is a strong need to consider mixed speech signals emitted by *moving* speakers, and/or recorded by a *moving* robot, and perturbed by reverberations. More generally, changes in the environment such as door/window opening/closing or curtain pulling must also be accounted for. Note that in this paper, the mixtures under consideration can be *underdetermined*, i.e., there may be less microphones than sources, which is a difficult ASS problem in its own right [1].

A. Related Work

The ASS literature that deals with time-invariant mixing filters is much larger than the literature dealing with time-varying filters. Therefore, we briefly discuss the former before reviewing the latter. State-of-the-art time-invariant ASS methods generally start with a time-frequency (TF) decom-

A variational EM algorithm for the separation of moving sound sources, IEEE WASPAA 2015 – **Best Student Paper Award**

A Variational EM Algorithm for the Separation of Time-Varying Convolutive Audio Mixtures, IEEE/ACM TASLP 2016

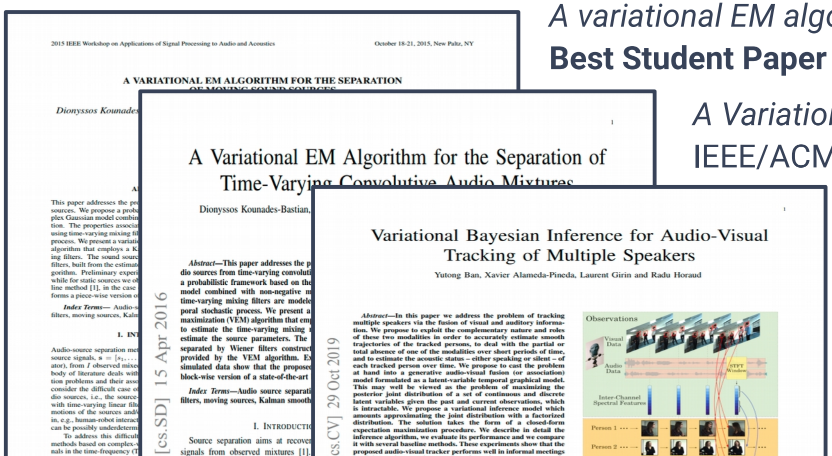
How to learn and infer? [v2 – Variational EM]

Close look to the log-likelihood, for any distribution $q_{\phi}(\mathbf{z}|\mathbf{x})$:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{q_{\phi}}(\mathbf{z}|\mathbf{x})\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\} + \mathcal{H}_{q_{\phi}} + \mathcal{D}_{\mathsf{KL}}(q_{\phi}\|p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))$$

In the exact EM, $q_{\phi}(\mathbf{z}|\mathbf{x}) = p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$ and $\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ is a tight lower bound. In some cases (not too complex), $p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$ is computationally intractable. Break the hidden in two (or +).

$$\mathbf{z} = (\mathbf{z}_1,\mathbf{z}_2) \quad \text{and} \quad q_{\phi}(\mathbf{z}) = q_{\phi_1}(\mathbf{z}_1)q_{\phi_2}(\mathbf{z}_2) \qquad q^*_{\phi_1}, q^*_{\phi_2} = \arg\min \mathcal{D}_{\mathsf{KL}}(q_{\phi_1}q_{\phi_2}\|p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))$$



2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics October 18-21, 2015, New Paltz, NY

A VARIATIONAL EM ALGORITHM FOR THE SEPARATION OF MOVING SOUND SOURCES

Dionyssos Kounades

A Variational EM Algorithm for the Separation of Time-Varying Convolutive Audio Mixtures

Dionyssos Kounades-Bastian,

[cs.SD] 15 Apr 2016

Variational Bayesian Inference for Audio-Visual Tracking of Multiple Speakers

Yutong Ban, Xavier Alameda-Pineda, Laurent Girin and Radu Horaud

[cs.CV] 29 Oct 2019

Abstract—In this paper we address the problem of tracking multiple speakers via the fusion of visual and auditory information. We propose to exploit the complementary nature and roles of these two modalities in order to accurately estimate smooth trajectories of the tracked persons, to deal with the partial or total absence of one of the modalities over short periods of time, and to estimate the acoustic status – either speaking or silent – of each tracked person over time. We propose to cast the problem at hand into a generative audio-visual fusion (or association) model formulated as a latent-variable temporal graphical model. This may well be viewed as the problem of maximizing the posterior joint distribution of a set of continuous and discrete latent variables given the past and current observations, which is intractable. We propose a variational inference model which amounts approximating the joint distribution with a factorized distribution. The solution takes the form of a closed-form expectation maximization procedure. We describe in detail the inference algorithm, we evaluate its performance and we compare it with several baseline methods. These experiments show that the proposed audio-visual tracker performs well in informal meetings

A variational EM algorithm for the separation of moving sound sources, IEEE WASPAA 2015 – **Best Student Paper Award**

A Variational EM Algorithm for the Separation of Time-Varying Convolutive Audio Mixtures, IEEE/ACM TASLP 2016

Variational Bayesian Inference for Audio-Visual Tracking of Multiple Speakers, IEEE TPAMI 2019.

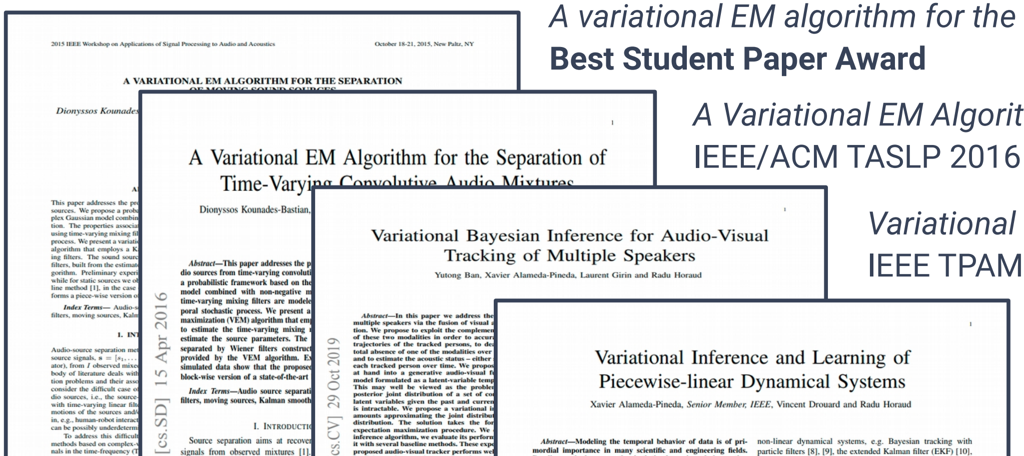
How to learn and infer? [v2 – Variational EM]

Close look to the log-likelihood, for any distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{q_\phi}(\mathbf{z}|\mathbf{x})\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\} + \mathcal{H}_{q_\phi} + \mathcal{D}_{\mathsf{KL}}(q_{\boldsymbol{\phi}}\|p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))$$

In the exact EM, $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$ and $\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ is a tight lower bound. In some cases (not too complex), $p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$ is computationally intractable. Break the hidden in two (or +).

$$\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2) \quad \text{and} \quad q_{\boldsymbol{\phi}}(\mathbf{z}) = q_{\boldsymbol{\phi}_1}(\mathbf{z}_1) q_{\boldsymbol{\phi}_2}(\mathbf{z}_2) \qquad q^*_{\boldsymbol{\phi}_1}, q^*_{\boldsymbol{\phi}_2} = \arg\min \mathcal{D}_{\mathsf{KL}}(q_{\boldsymbol{\phi}_1} q_{\boldsymbol{\phi}_2} \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))$$



2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics October 18-21, 2015, New Paltz, NY

A VARIATIONAL EM ALGORITHM FOR THE SEPARATION OF MOVING SOUND SOURCES

A Variational EM Algorithm for the Separation of Time-Varying Convolutive Audio Mixtures

Variational Bayesian Inference for Audio-Visual Tracking of Multiple Speakers

Yutong Ban, Xavier Alameda-Pineda, Laurent Girin and Radu Horaud

Variational Inference and Learning of Piecewise-linear Dynamical Systems

Xavier Alameda-Pineda, *Senior Member, IEEE*, Vincent Drouard and Radu Horaud

A variational EM algorithm for the separation of moving sound sources, IEEE WASPAA 2015 – **Best Student Paper Award**

A Variational EM Algorithm for the Separation of Time-Varying Convolutive Audio Mixtures, IEEE/ACM TASLP 2016

Variational Bayesian Inference for Audio-Visual Tracking of Multiple Speakers, IEEE TPAMI 2019.

Variational Inference and Learning of Piecewise-linear Dynamical Systems, IEEE TNNLS, 2021.

How to learn and infer? [v3 – VAE]

How to learn and infer? [v3 – VAE]

What if the posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ does not have analytic form?

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{q_{\phi}}(\mathbf{z}|\mathbf{x})\{\log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})\} + \mathcal{H}_{q_{\boldsymbol{\phi}}} + \mathcal{D}_{\mathsf{KL}}(q_{\boldsymbol{\phi}}||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))$$

How to learn and infer? [v3 – VAE]

What if the posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ does not have analytic form?

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{q_{\phi}}(\mathbf{z}|\mathbf{x})\{\log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})\} + \mathcal{H}_{q_{\boldsymbol{\phi}}} + \xcancel{\mathcal{D}_{\text{KL}}(q_{\boldsymbol{\phi}} \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))}$$

Seminal case, the generative model $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ is implemented with a deep neural network.

How to learn and infer? [v3 – VAE]

What if the posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ does not have analytic form?

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{q_\phi}(\mathbf{z}|\mathbf{x})\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\} + \mathcal{H}_{q_{\boldsymbol{\phi}}} + \xcancel{\mathcal{D}_{\text{KL}}(q_{\boldsymbol{\phi}}||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))}$$

Seminal case, the generative model $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ is implemented with a deep neural network.
The posterior is approximated with a different neural network: $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \approx p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) \geq \mathbb{E}_{q_\phi}(\mathbf{z}|\mathbf{x})\{\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\} - \mathcal{D}_{\text{KL}}(q_{\boldsymbol{\phi}}||p_{\boldsymbol{\theta}}(\mathbf{z})) = \mathcal{L}_{\text{ELBO}}(\boldsymbol{\theta},\boldsymbol{\phi})$$

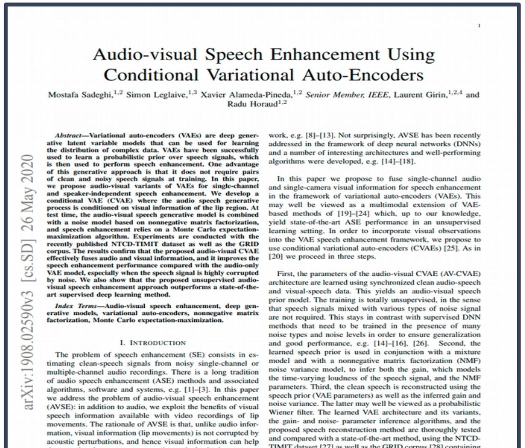
How to learn and infer? [v3 – VAE]

What if the posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ does not have analytic form?

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{q_\phi}(\mathbf{z}|\mathbf{x})\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\} + \mathcal{H}_{q_\phi} + \xcancel{\mathcal{D}_{\text{KL}}(q_\phi \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))}$$

Seminal case, the generative model $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ is implemented with a deep neural network.
The posterior is approximated with a different neural network: $q_\phi(\mathbf{z}|\mathbf{x}) \approx p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) \geq \mathbb{E}_{q_\phi}(\mathbf{z}|\mathbf{x})\{\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\} - \mathcal{D}_{\text{KL}}(q_\phi \| p_{\boldsymbol{\theta}}(\mathbf{z})) = \mathcal{L}_{\text{ELBO}}(\boldsymbol{\theta}, \phi)$$



1

Audio-visual Speech Enhancement Using Conditional Variational Auto-Encoders

Mostafa Sadeghi,[1,2] Simon Leglaive,[1,3] Xavier Alameda-Pineda,[1,2] *Senior Member, IEEE*, Laurent Girin,[1,2,4] and Radu Horaud[1,2]

***Abstract*—Variational auto-encoders (VAEs) are deep generative latent variable models that can be used for learning the distribution of complex data. VAEs have been successfully used to learn a probabilistic prior over speech signals, which is then used to perform speech enhancement. One advantage of this generative approach is that it does not require pairs of clean and noisy speech signals at training. In this paper, we propose audio-visual variants of VAEs for single-channel and speaker-independent speech enhancement. We develop a conditional VAE (CVAE) where the audio speech generative process is conditioned on visual information of the lip region. At test time, the audio-visual speech generative model is combined with a noise model based on nonnegative matrix factorization, and speech enhancement relies on a Monte Carlo expectation-maximization algorithm. Experiments are conducted with the recently published NTCD-TIMIT dataset as well as the GRID corpus. The results confirm that the proposed audio-visual CVAE effectively fuses audio and visual information, and it improves the speech enhancement performance compared with the audio-only VAE model, especially when the speech signal is highly corrupted by noise. We also show that the proposed unsupervised audio-visual speech enhancement approach outperforms a state-of-the-art supervised deep learning method.**

***Index Terms*—Audio-visual speech enhancement, deep generative models, variational auto-encoders, nonnegative matrix factorization, Monte Carlo expectation-maximization.**

I. Introduction

The problem of speech enhancement (SE) consists in estimating clean-speech signals from noisy single-channel or multiple-channel audio recordings. There is a long tradition of audio speech enhancement (ASE) methods and associated algorithms, software and systems, e.g. [1]–[3]. In this paper we address the problem of audio-visual speech enhancement (AVSE): in addition to audio, we exploit the benefits of visual speech information available with video recordings of lip movements. The rationale of AVSE is that, unlike audio information, visual information (lip movements) is not corrupted by acoustic perturbations, and hence visual information can help

work, e.g. [8]–[13]. Not surprisingly, AVSE has been recently addressed in the framework of deep neural networks (DNNs) and a number of interesting architectures and well-performing algorithms were developed, e.g. [14]–[18].

In this paper we propose to fuse single-channel audio and single-camera visual information for speech enhancement in the framework of variational auto-encoders (VAEs). This may well be viewed as a multimodal extension of VAE-based methods of [19]–[24] which, up to our knowledge, yield state-of-the-art ASE performance in an unsupervised learning setting. In order to incorporate visual observations into the VAE speech enhancement framework, we propose to use conditional variational auto-encoders (CVAEs) [25]. As in [20] we proceed in three steps.

First, the parameters of the audio-visual CVAE (AV-CVAE) architecture are learned using synchronized clean audio-speech and visual-speech data. This yields an audio-visual speech prior model. The training is totally unsupervised, in the sense that speech signals mixed with various types of noise signal are not required. This stays in contrast with supervised DNN methods that need to be trained in the presence of many noise types and noise levels in order to ensure generalization and good performance, e.g. [14]–[16], [26]. Second, the learned speech prior is used in conjunction with a mixture model and with a nonnegative matrix factorization (NMF) noise variance model, to infer both the gain, which models the time-varying loudness of the speech signal, and the NMF parameters. Third, the clean speech is reconstructed using the speech prior (VAE parameters) as well as the inferred gain and noise variance. The latter may well be viewed as a probabilistic Wiener filter. The learned VAE architecture and its variants, the gain- and noise- parameter inference algorithms, and the proposed speech reconstruction method are thoroughly tested and compared with a state-of-the-art method, using the NTCD-TIMIT dataset [27] as well as the GRID corpus [28] containing

Audio-visual speech enhancement using conditional variational auto-encoders, IEEE/ACM TASLP 2020

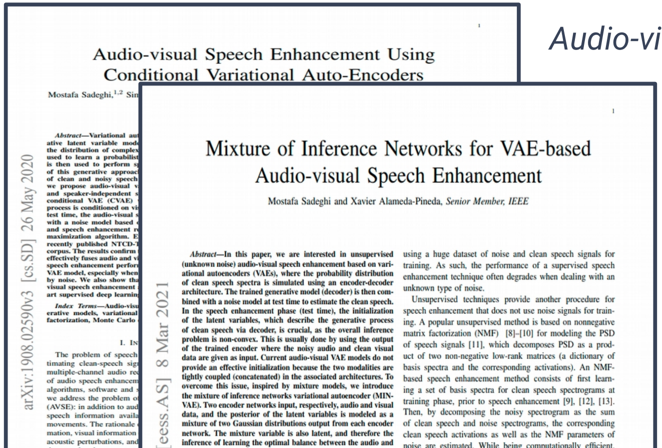
How to learn and infer? [v3 – VAE]

What if the posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ does not have analytic form?

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{q_\phi}(\mathbf{z}|\mathbf{x})\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\} + \mathcal{H}_{q_\phi} + \cancel{\mathcal{D}_{\text{KL}}(q_\phi \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))}$$

Seminal case, the generative model $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ is implemented with a deep neural network.
The posterior is approximated with a different neural network: $q_\phi(\mathbf{z}|\mathbf{x}) \approx p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) \geq \mathbb{E}_{q_\phi}(\mathbf{z}|\mathbf{x})\{\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\} - \mathcal{D}_{\text{KL}}(q_\phi \| p_{\boldsymbol{\theta}}(\mathbf{z})) = \mathcal{L}_{\text{ELBO}}(\boldsymbol{\theta},\phi)$$



Audio-visual Speech Enhancement Using
Conditional Variational Auto-Encoders

Mostafa Sadeghi,[1,2] Sin

arXiv:1908.02590v3 [cs.SD] 26 May 2020

Abstract—Variational aut
ative latent variable mode
the distribution of complex
used to learn a probabilist
is then used to perform sp
of this generative approac
of clean and noisy speech
we propose audio-visual v
and speaker-independent s
conditional VAE (CVAE)
process is conditioned on vi
test time, the audio-visual s
with a noise model based
and speech enhancement r
maximization algorithm. E
recently published NTCD-T
corpus. The results confirm
effectively fuses audio and vi
speech enhancement perfor
VAE model, especially when
by noise. We also show tha
visual speech enhancement
art supervised deep learnin

Index Terms—Audio-visu
erative models, variational
factorization, Monte Carlo

I. IN

The problem of speech
timating clean-speech sig
multiple-channel audio rec
of audio speech enhancem
algorithms, software and
we address the problem o
(AVSE): in addition to aud
speech information availa
movements. The rationale
mation, visual information
acoustic perturbations, and

1

Mixture of Inference Networks for VAE-based
Audio-visual Speech Enhancement

Mostafa Sadeghi and Xavier Alameda-Pineda, *Senior Member, IEEE*

[eess.AS] 8 Mar 2021

Abstract—In this paper, we are interested in unsupervised (unknown noise) audio-visual speech enhancement based on variational autoencoders (VAEs), where the probability distribution of clean speech spectra is simulated using an encoder-decoder architecture. The trained generative model (decoder) is then combined with a noise model at test time to estimate the clean speech. In the speech enhancement phase (test time), the initialization of the latent variables, which describe the generative process of clean speech via decoder, is crucial, as the overall inference problem is non-convex. This is usually done by using the output of the trained encoder where the noisy audio and clean visual data are given as input. Current audio-visual VAE models do not provide an effective initialization because the two modalities are tightly coupled (concatenated) in the associated architectures. To overcome this issue, inspired by mixture models, we introduce the mixture of inference networks variational autoencoder (MIN-VAE). Two encoder networks input, respectively, audio and visual data, and the posterior of the latent variables is modeled as a mixture of two Gaussian distributions output from each encoder network. The mixture variable is also latent, and therefore the inference of learning the optimal balance between the audio and

using a huge dataset of noise and clean speech signals for training. As such, the performance of a supervised speech enhancement technique often degrades when dealing with an unknown type of noise.

Unsupervised techniques provide another procedure for speech enhancement that does not use noise signals for training. A popular unsupervised method is based on nonnegative matrix factorization (NMF) [8]–[10] for modeling the PSD of speech signals [11], which decomposes PSD as a product of two non-negative low-rank matrices (a dictionary of basis spectra and the corresponding activations). An NMF-based speech enhancement method consists of first learning a set of basis spectra for clean speech spectrograms at training phase, prior to speech enhancement [9], [12], [13]. Then, by decomposing the noisy spectrogram as the sum of clean speech and noise spectrograms, the corresponding clean speech activations as well as the NMF parameters of noise are estimated. While being computationally efficient,

Audio-visual speech enhancement using conditional variational auto-encoders, IEEE/ACM TASLP 2020

Mixture of inference networks for VAE-based audio-visual speech enhancement, IEEE TSP 2021

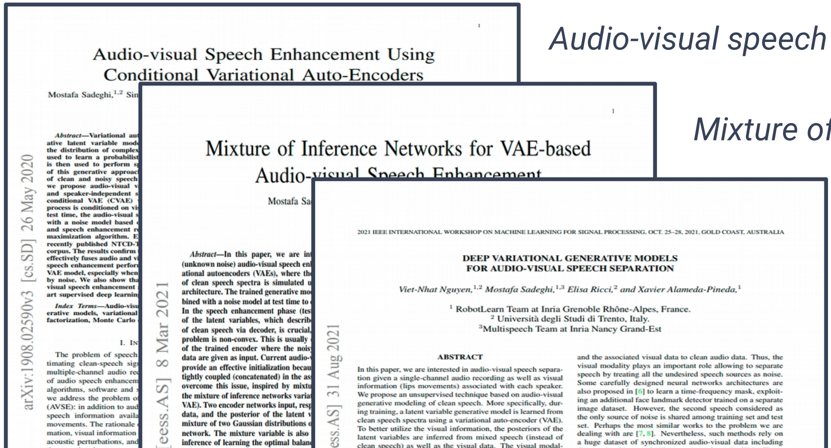
How to learn and infer? [v3 – VAE]

What if the posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ does not have analytic form?

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{q_\phi}(\mathbf{z}|\mathbf{x})\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\} + \mathcal{H}_{q_\phi} + \xcancel{\mathcal{D}_{\text{KL}}(q_\phi \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))}$$

Seminal case, the generative model $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ is implemented with a deep neural network.
The posterior is approximated with a different neural network: $q_\phi(\mathbf{z}|\mathbf{x}) \approx p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) \geq \mathbb{E}_{q_\phi}(\mathbf{z}|\mathbf{x})\{\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\} - \mathcal{D}_{\text{KL}}(q_\phi \| p_{\boldsymbol{\theta}}(\mathbf{z})) = \mathcal{L}_{\text{ELBO}}(\boldsymbol{\theta}, \phi)$$



Audio-visual Speech Enhancement Using Conditional Variational Auto-Encoders

Mostafa Sadeghi,[1,2] Sin

arXiv:1908.02590v3 [cs.SD] 26 May 2020

Mixture of Inference Networks for VAE-based Audio-visual Speech Enhancement

Mostafa Sa

[eess.AS] 8 Mar 2021

2021 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, OCT. 25–28, 2021, GOLD COAST, AUSTRALIA

DEEP VARIATIONAL GENERATIVE MODELS FOR AUDIO-VISUAL SPEECH SEPARATION

Viet-Nhat Nguyen,[1,2] Mostafa Sadeghi,[1,3] Elisa Ricci,[2] and Xavier Alameda-Pineda,[1]

[1] RobotLearn Team at Inria Grenoble Rhône-Alpes, France.
[2] Università degli Studi di Trento, Italy.
[3]Multispeech Team at Inria Nancy Grand-Est

ABSTRACT

In this paper, we are interested in audio-visual speech separation given a single-channel audio recording as well as visual information (lips movements) associated with each speaker. We propose an unsupervised technique based on audio-visual generative modeling of clean speech. More specifically, during training, a latent variable generative model is learned from clean speech spectra using a variational auto-encoder (VAE). To better utilize the visual information, the posteriors of the latent variables are inferred from mixed speech (instead of clean speech) as well as the visual data. The visual modal-

and the associated visual data to clean audio data. Thus, the visual modality plays an important role allowing to separate speech by treating all the undesired speech sources as noise. Some carefully designed neural networks architectures are also proposed in [6] to learn a time-frequency mask, exploiting an additional face landmark detector trained on a separate image dataset. However, the second speech considered as the only source of noise is shared among training set and test set. Perhaps the most similar works to the problem we are dealing with are [7, 8]. Nevertheless, such methods rely on a huge dataset of synchronized audio-visual data including

[eess.AS] 31 Aug 2021

Audio-visual speech enhancement using conditional variational auto-encoders, IEEE/ACM TASLP 2020

Mixture of inference networks for VAE-based audio-visual speech enhancement, IEEE TSP 2021

Deep variational generative models for audio-visual speech separation, IEEE MLSP 2021

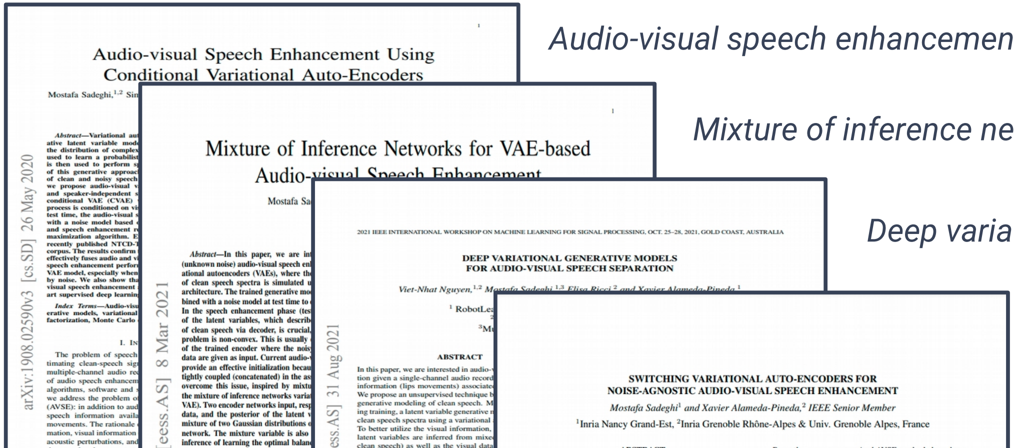
How to learn and infer? [v3 – VAE]

What if the posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ does not have analytic form?

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{q_\phi}(\mathbf{z}|\mathbf{x})\{\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\} + \mathcal{H}_{q_{\boldsymbol{\phi}}} + \xcancel{\mathcal{D}_{\mathsf{KL}}(q_{\boldsymbol{\phi}}\|p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))}$$

Seminal case, the generative model $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ is implemented with a deep neural network.
The posterior is approximated with a different neural network: $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \approx p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) \geq \mathbb{E}_{q_\phi}(\mathbf{z}|\mathbf{x})\{\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\} - \mathcal{D}_{\mathsf{KL}}(q_{\boldsymbol{\phi}}\|p_{\boldsymbol{\theta}}(\mathbf{z})) = \mathcal{L}_{\mathsf{ELBO}}(\boldsymbol{\theta},\boldsymbol{\phi})$$



Audio-visual Speech Enhancement Using Conditional Variational Auto-Encoders

Mixture of Inference Networks for VAE-based Audio-visual Speech Enhancement

2021 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, OCT. 25–28, 2021, GOLD COAST, AUSTRALIA

DEEP VARIATIONAL GENERATIVE MODELS FOR AUDIO-VISUAL SPEECH SEPARATION

SWITCHING VARIATIONAL AUTO-ENCODERS FOR NOISE-AGNOSTIC AUDIO-VISUAL SPEECH ENHANCEMENT

Mostafa Sadeghi[1] *and Xavier Alameda-Pineda,*[2] *IEEE Senior Member*

[1]Inria Nancy Grand-Est, [2]Inria Grenoble Rhône-Alpes & Univ. Grenoble Alpes, France

Audio-visual speech enhancement using conditional variational auto-encoders, IEEE/ACM TASLP 2020

Mixture of inference networks for VAE-based audio-visual speech enhancement, IEEE TSP 2021

Deep variational generative models for audio-visual speech separation, IEEE MLSP 2021

Switching variational auto-encoders for noise-agnostic audio-visual speech enhancement, IEEE ICASSP, 2021.

How to learn and infer? [v4 – Dynamical VAE]

How to learn and infer? [v4 – Dynamical VAE]

How to model temporal dependencies with deep generative models?

How to learn and infer? [v4 – Dynamical VAE]

How to model temporal dependencies with deep generative models?

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}) \geq \mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})}\{\log p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|\mathbf{z}_{1:T})\} - \mathcal{D}_{\mathsf{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})\|p_{\boldsymbol{\theta}}(\mathbf{z}_{1:T}))$$

Both the generative model and approximate posterior need to model temporal dependencies implemented via, e.g., recurrent or transformer networks.

How to learn and infer? [v4 – Dynamical VAE]

How to model temporal dependencies with deep generative models?

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}) \geq \mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})}\{\log p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|\mathbf{z}_{1:T})\} - \mathcal{D}_{\mathsf{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})\|p_{\boldsymbol{\theta}}(\mathbf{z}_{1:T}))$$

Both the generative model and approximate posterior need to model temporal dependencies implemented via, e.g., recurrent or transformer networks.

The dependencies of $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ can be implemented in $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ (still an approximation).

How to learn and infer? [v4 – Dynamical VAE]

How to model temporal dependencies with deep generative models?

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}) \geq \mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})}\{\log p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|\mathbf{z}_{1:T})\} - \mathcal{D}_{\mathsf{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})\|p_{\boldsymbol{\theta}}(\mathbf{z}_{1:T}))$$

Both the generative model and approximate posterior need to model temporal dependencies implemented via, e.g., recurrent or transformer networks.

The dependencies of $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ can be implemented in $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ (still an approximation).

A RECURRENT VARIATIONAL AUTOENCODER FOR SPEECH ENHANCEMENT

Simon Leglaive[1,2] Xavier Alameda-Pineda[2] Laurent Girin[2,3] Radu Horaud[2]

[1]CentraleSupélec, IETR, France [2]Inria Grenoble Rhône-Alpes, France
[3]Univ. Grenoble Alpes, Grenoble INP, GIPSA-lab, France

ABSTRACT

This paper presents a generative approach to speech enhancement based on a recurrent variational autoencoder (RVAE). The deep generative speech model is trained using clean speech signals only, and it is combined with a nonnegative matrix factorization noise model for speech enhancement. We propose a variational expectation-maximization algorithm where the encoder of the RVAE is fine-tuned at test time, to approximate the distribution of the latent variables given the noisy speech observations. Compared with previous approaches based on feed-forward fully-connected architectures, the proposed recurrent deep generative speech model induces a posterior temporal dynamic over the latent variables, which is shown to improve the speech enhancement results.

Index Terms— Speech enhancement, recurrent variational autoencoders, nonnegative matrix factorization, variational inference.

1. INTRODUCTION

Speech enhancement is an important problem in audio signal processing [1]. The objective is to recover a clean speech signal from a noisy mixture signal. In this work, we focus on single-channel (i.e. single-microphone) speech enhancement.

Discriminative approaches based on deep neural networks have been extensively used for speech enhancement. They try to estimate a clean speech spectrogram or a time-frequency mask from a noisy speech spectrogram, see e.g. [2, 3, 4, 5, 6]. Recently, deep generative speech models based on variational autoencoders (VAEs) [7] have been investigated for single-channel [8, 9, 10, 11] and multichannel speech enhancement [12, 13, 14]. A pre-trained deep generative speech model is combined with a nonnegative matrix factorization (NMF) [15] noise model whose parameters are estimated at test time, from the observation of the noisy mixture signal only. Compared with discriminative approaches, these generative methods do not require pairs of clean and noisy speech signal for training. This setting was referred to as "semi-supervised source separation" in previous works [16, 17, 18], which should not be confused with the supervised/unsupervised terminology of machine learning.

To the best of our knowledge, the aforementioned works on VAE-based deep generative models for speech enhancement have

of the RVAE is fine-tuned to approximate the posterior distribution of the latent variables, given the noisy speech observations. This model induces a posterior temporal dynamic over the latent variables, which is further propagated to the speech estimate. Experimental results show that this approach outperforms its feed-forward and fully-connected counterpart.

2. DEEP GENERATIVE SPEECH MODEL

2.1. Definition

Let $\mathbf{s} = \{\mathbf{s}_n \in \mathbb{C}^F\}_{n=0}^{N-1}$ denote a sequence of short-time Fourier transform (STFT) speech time frames, and $\mathbf{z} = \{\mathbf{z}_n \in \mathbb{R}^L\}_{n=0}^{N-1}$ a corresponding sequence of latent random vectors. We define the following hierarchical generative speech model independently for all time frames $n \in \{0, ..., N-1\}$:

$$\mathbf{s}_n \mid \mathbf{z} \sim \mathcal{N}_c\left(\mathbf{0}, \operatorname{diag}\{\mathbf{v}_{\mathbf{s},n}(\mathbf{z})\}\right), \quad \text{with} \quad \mathbf{z}_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad (1)$$

and where $\mathbf{v}_{\mathbf{s},n}(\mathbf{z}) \in \mathbb{R}_+^F$ will be defined by means of a *decoder* neural network. $\mathcal{N}$ denotes the multivariate Gaussian distribution for a real-valued random vector and $\mathcal{N}_c$ denotes the multivariate complex proper Gaussian distribution [21]. Multiple choices are possible to define the neural network corresponding to $\mathbf{v}_{\mathbf{s},n}(\mathbf{z})$, which will lead to different probabilistic graphical models represented in Fig. 1.

FFNN generative speech model $\mathbf{v}_{\mathbf{s},n}(\mathbf{z}) = \boldsymbol{\varphi}_{\text{dec}}^{\text{FFNN}}(\mathbf{z}_n; \boldsymbol{\theta}_{\text{dec}})$ where $\boldsymbol{\varphi}_{\text{dec}}^{\text{FFNN}}(\cdot; \boldsymbol{\theta}_{\text{dec}}) : \mathbb{R}^L \mapsto \mathbb{R}_+^F$ denotes a feed-forward fully-connected neural network (FFNN) of parameters $\boldsymbol{\theta}_{\text{dec}}$. Such an architecture was used in [8, 9, 10, 11, 12, 13, 14]. As represented in Fig. 1a, this model results in the following factorization of the complete-data likelihood:

$$p(\mathbf{s}, \mathbf{z}; \boldsymbol{\theta}_{\text{dec}}) = \textstyle\prod_{n=0}^{N-1} p(\mathbf{s}_n|\mathbf{z}_n; \boldsymbol{\theta}_{\text{dec}})p(\mathbf{z}_n). \qquad (2)$$

Note that in this case, the speech STFT time frames are not only conditionally independent, but also marginally independent, i.e. $p(\mathbf{s}; \boldsymbol{\theta}_{\text{dec}}) = \prod_{n=0}^{N-1} p(\mathbf{s}_n; \boldsymbol{\theta}_{\text{dec}})$.

RNN generative speech model $\mathbf{v}_{\mathbf{s},n}(\mathbf{z}) = \boldsymbol{\varphi}_{\text{dec},n}^{\text{RNN}}(\mathbf{z}_{0:n}; \boldsymbol{\theta}_{\text{dec}})$ where $\boldsymbol{\varphi}_{\text{dec},n}^{\text{RNN}}(\cdot; \boldsymbol{\theta}_{\text{dec}}) : \mathbb{R}^{L\times(n+1)} \mapsto \mathbb{R}_+^F$ denotes the output at time frame n of a recurrent neural network (RNN), taking as input the

A recurrent variational autoencoder for speech enhancement, IEEE ICASSP 2020

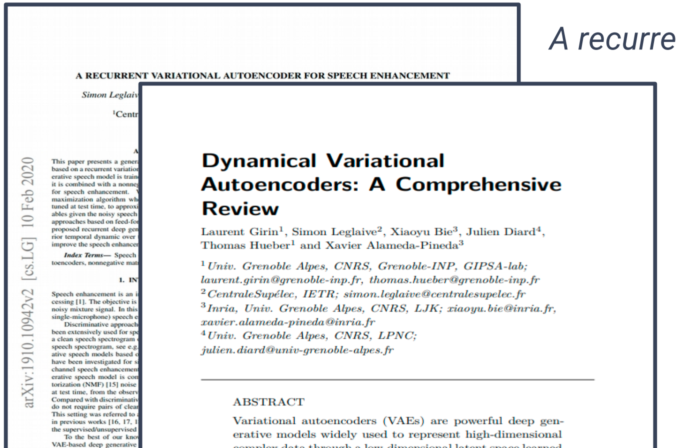
How to learn and infer? [v4 – Dynamical VAE]

How to model temporal dependencies with deep generative models?

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}) \geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})}\{\log p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|\mathbf{z}_{1:T})\} - \mathcal{D}_{\mathsf{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})\|p_{\boldsymbol{\theta}}(\mathbf{z}_{1:T}))$$

Both the generative model and approximate posterior need to model temporal dependencies implemented via, e.g., recurrent or transformer networks.

The dependencies of $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ can be implemented in $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ (still an approximation).



A RECURRENT VARIATIONAL AUTOENCODER FOR SPEECH ENHANCEMENT

Simon Leglaiv

arXiv:1910.10942v2 [cs.LG] 10 Feb 2020

Dynamical Variational Autoencoders: A Comprehensive Review

Laurent Girin[1], Simon Leglaive[2], Xiaoyu Bie[3], Julien Diard[4], Thomas Hueber[1] and Xavier Alameda-Pineda[3]

[1] *Univ. Grenoble Alpes, CNRS, Grenoble-INP, GIPSA-lab; laurent.girin@grenoble-inp.fr, thomas.hueber@grenoble-inp.fr*
[2] *CentraleSupélec, IETR; simon.leglaive@centralesupelec.fr*
[3] *Inria, Univ. Grenoble Alpes, CNRS, LJK; xiaoyu.bie@inria.fr, xavier.alameda-pineda@inria.fr*
[4] *Univ. Grenoble Alpes, CNRS, LPNC; julien.diard@univ-grenoble-alpes.fr*

ABSTRACT

Variational autoencoders (VAEs) are powerful deep generative models widely used to represent high-dimensional

A recurrent variational autoencoder for speech enhancement, IEEE ICASSP 2020

Dynamical variational autoencoders: A comprehensive review, FnT Machine Learning 2021

How to learn and infer? [v4 – Dynamical VAE]

How to model temporal dependencies with deep generative models?

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}) \geq \mathbb{E}_{q_{\phi}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})}\{\log p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|\mathbf{z}_{1:T})\} - \mathcal{D}_{\mathsf{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})\|p_{\boldsymbol{\theta}}(\mathbf{z}_{1:T}))$$

Both the generative model and approximate posterior need to model temporal dependencies implemented via, e.g., recurrent or transformer networks.

The dependencies of $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ can be implemented in $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ (still an approximation).

A RECURRENT VARIATIONAL AUTOENCODER FOR SPEECH ENHANCEMENT

Simon Leglaiv

arXiv:1910.10942v2 [cs.LG] 10 Feb 2020

Dynamical Variational Autoencoc Review

Laurent Girin[1], Sim Thomas Hueber[1] ar

ABSTRACT

[cs.CV] 4 Apr 2022

HiT-DVAE: Human Motion Generation via Hierarchical Transformer Dynamical VAE

Xiaoyu Bie[1*], Wen Guo[1,4*], Simon Leglaive[2], Lauren Girin[3], Francesc Moreno-Noguer[4], and Xavier Alameda-Pineda[1]

[1]Inria, Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, France
[2]CentraleSupélec, IETR, France
[3]Univ. Grenoble Alpes, Grenoble-INP, GIPSA-lab, France
[4]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Spain

Abstract

Studies on the automatic processing of 3D human pose data have flourished in the recent past. In this paper, we are interested in the generation of plausible and diverse future human poses following an observed 3D pose sequence. Current methods address this problem by injecting random variables from a single latent space into a deterministic motion prediction framework, which precludes the inherent multi-modality in human motion generation. In addition, previous works rarely explore the use of attention to select which frames are to be used to inform the generation process up to our knowledge. To overcome these limitations, we propose Hierarchical

A recurrent variational autoencoder for speech enhancement, IEEE ICASSP 2020

Dynamical variational autoencoders: A comprehensive review, FnT Machine Learning 2021

HiT-DVAE: Human Motion Generation via Hierarchical Transformer Dynamical VAE

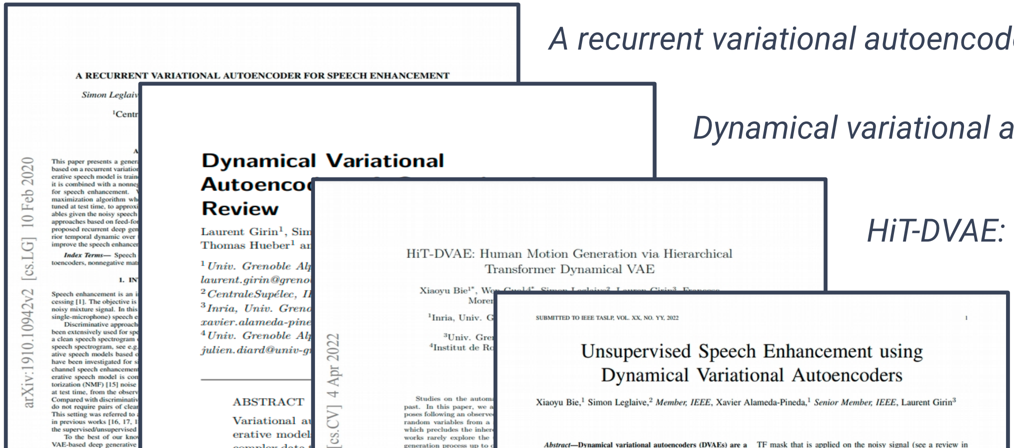
How to learn and infer? [v4 – Dynamical VAE]

How to model temporal dependencies with deep generative models?

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}) \geq \mathbb{E}_{q_{\phi}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})}\{\log p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|\mathbf{z}_{1:T})\} - \mathcal{D}_{\mathsf{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})\|p_{\boldsymbol{\theta}}(\mathbf{z}_{1:T}))$$

Both the generative model and approximate posterior need to model temporal dependencies implemented via, e.g., recurrent or transformer networks.

The dependencies of $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ can be implemented in $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ (still an approximation).



A RECURRENT VARIATIONAL AUTOENCODER FOR SPEECH ENHANCEMENT

arXiv:1910.10942v2 [cs.LG] 10 Feb 2020

Dynamical Variational Autoencoders: A Comprehensive Review

HiT-DVAE: Human Motion Generation via Hierarchical Transformer Dynamical VAE

[cs.CV] 4 Apr 2022

SUBMITTED TO IEEE TASLP, VOL. XX, NO. YY, 2022

Unsupervised Speech Enhancement using Dynamical Variational Autoencoders

Xiaoyu Bie,[1] Simon Leglaive,[2] *Member, IEEE*, Xavier Alameda-Pineda,[1] *Senior Member, IEEE*, Laurent Girin[3]

A recurrent variational autoencoder for speech enhancement, IEEE ICASSP 2020

Dynamical variational autoencoders: A comprehensive review, FnT Machine Learning 2021

HiT-DVAE: Human Motion Generation via Hierarchical Transformer Dynamical VAE

Unsupervised speech enhancement using dynamical variational autoencoders, IEEE TASLP, 2022

How to learn and infer? [v4 – Dynamical VAE]

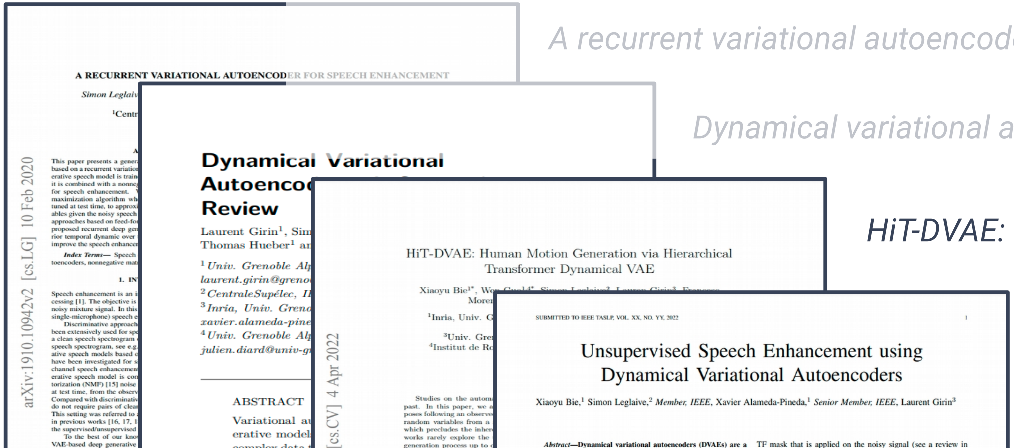
How to model temporal dependencies with deep generative models?

$$\log p_\theta(\mathbf{x}_{1:T}) \geq \mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})}\{\log p_\theta(\mathbf{x}_{1:T}|\mathbf{z}_{1:T})\} - \mathcal{D}_{\text{KL}}(q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})\|p_\theta(\mathbf{z}_{1:T}))$$

Both the generative model and approximate posterior need to model temporal dependencies implemented via, e.g., recurrent or transformer networks.

How is all this possible?

The dependencies of $p_\theta(\mathbf{z}|\mathbf{x})$ can be implemented in $q_\phi(\mathbf{z}|\mathbf{x})$ (still an approximation).



A RECURRENT VARIATIONAL AUTOENCODER FOR SPEECH ENHANCEMENT

Dynamical Variational Autoencoders: A Comprehensive Review

HiT-DVAE: Human Motion Generation via Hierarchical Transformer Dynamical VAE

Unsupervised Speech Enhancement using Dynamical Variational Autoencoders

A recurrent variational autoencoder for speech enhancement, IEEE ICASSP 2020

Dynamical variational autoencoders: A comprehensive review, FnT Machine Learning 2021

HiT-DVAE: Human Motion Generation via Hierarchical Transformer Dynamical VAE

Unsupervised speech enhancement using dynamical variational autoencoders, IEEE TASLP, 2022

Scientific life lessons

(Leçons de vie scientifique)

Back to 2009, Barcelona...

Back to 2009, Barcelona...

La paciència és la mare de la ciència. (catalan proverb)

Masters Thesis

Safe environment!

Masters Thesis

Safe environment!

Take pride in what you do, but don’t let your pride guide you.

PhD Thesis – 1st Submission

Want to do things on your own, but perhaps not ready ;)

PhD Thesis – 1st Submission

Want to do things on your own, but perhaps not ready ;)

Perseverance is key.

PhD Thesis – Paper Writing

“Writing” journal paper by concatenating two short papers (obtaining a long complex paper)

PhD Thesis – Paper Writing

“Writing” journal paper by concatenating two short papers (obtaining a long complex paper)

Writing is 1/3 of your research time.

PhD Thesis – Focus

We could do that, and that, and that, ...

PhD Thesis – Focus

We could do that, and that, and that, ...

...yeah, sure, but focus.

PhD Thesis – Defence

Tell a story...

PhD Thesis – Defence

with highlights,

Tell a story...

PhD Thesis – Defence

Tell a story...

with highlights,

unexpected features,

PhD Thesis – Defence

Tell a story...

with highlights,

unexpected features,

limitations and how to overcome them,

PhD Thesis – Defence

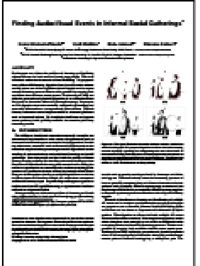
Tell a story...

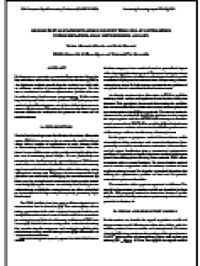
with highlights,

unexpected features,

limitations and how to overcome them,

...and be proud of your work!


ICMI 2011


EUSIPCO 2012

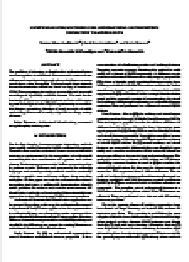
Humanoids 2012

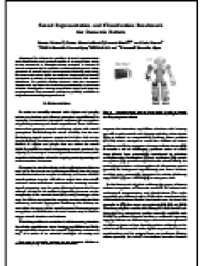
Humanoids 2012

JMUI 2013

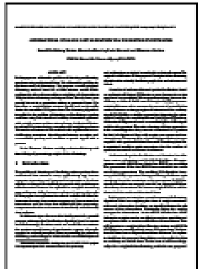
WASPAA 2013

Humanoids 2013

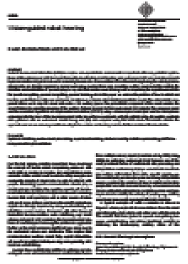

ICASSP 2013

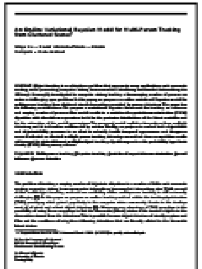

ICRA 2014

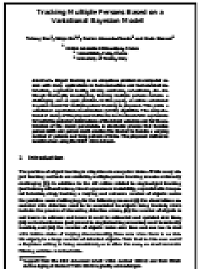
TASLP 2014

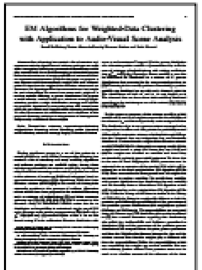

MLSP 2014

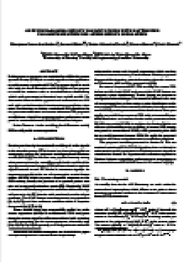
WASPAA 2015

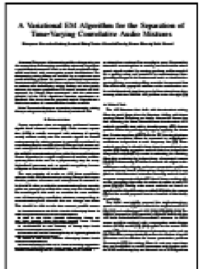

IJRR 2015

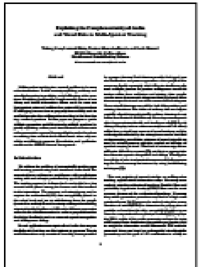

CVIU 2016

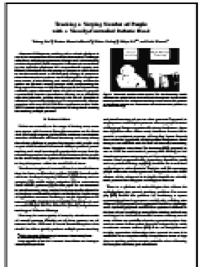

ECCV-W 2016


TPAMI 2016

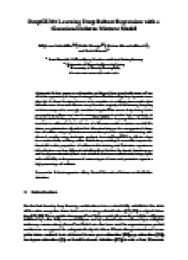

ICASSP 2016

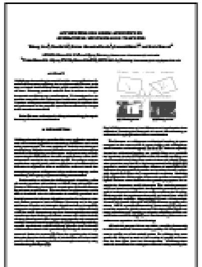

TASLP 2016


ICCV-W 2017


IROS 2017

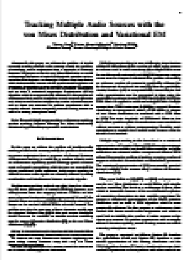
ICASSP 2017

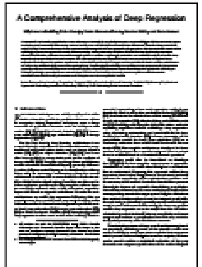

ECCV 2018

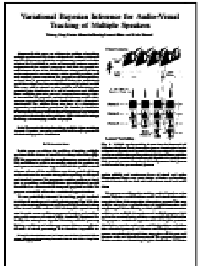

ICASSP 2018

IWAENC 2018

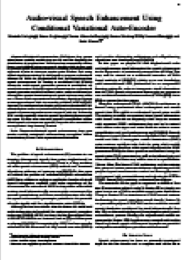
JSTSP 2019


SPL 2019

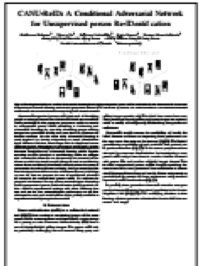

TPAMI 2019

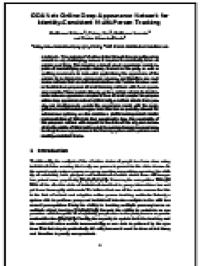

TPAMI 2020

CVPR 2020

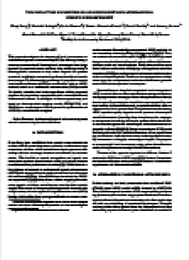

TASLP 2020

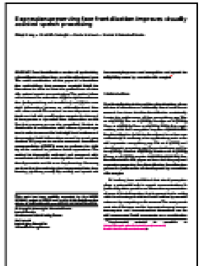
ICASSP 2020

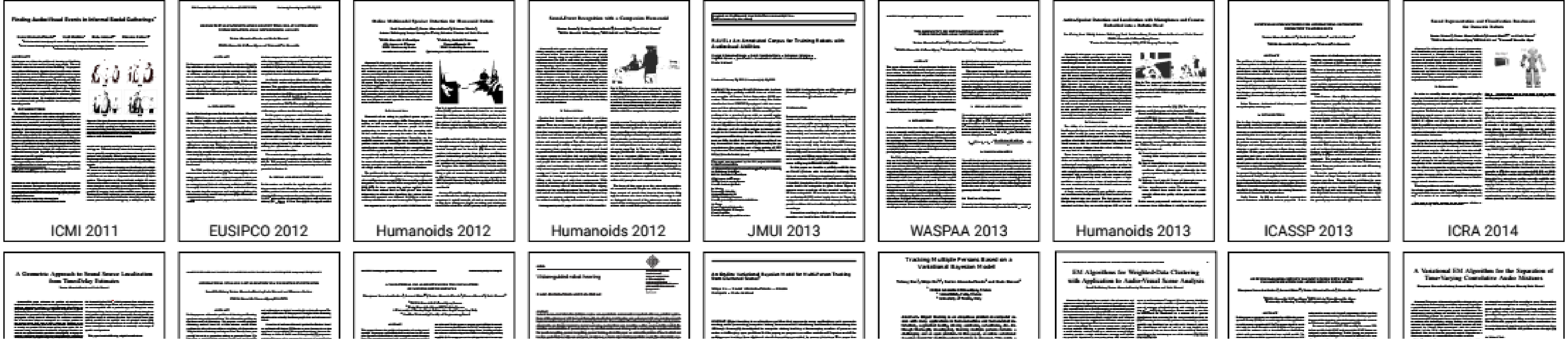

ICPR 2021


ICPR-W 2021

TNNLS 2021


ICASSP 2022

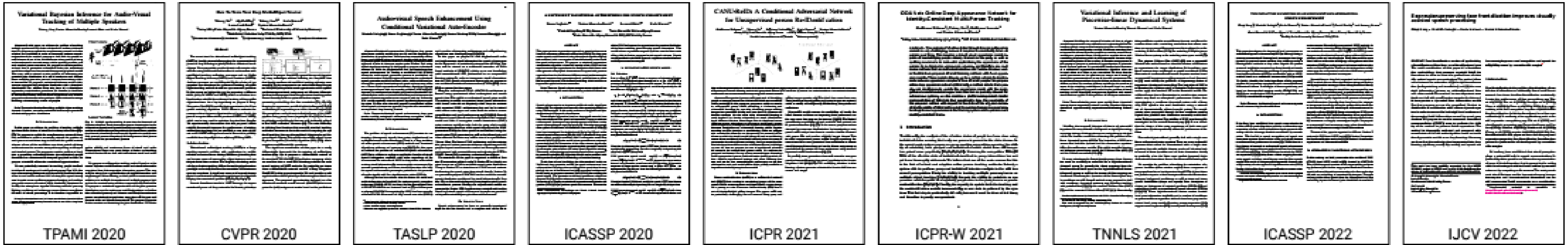

IJCV 2022



MIAI Grenoble Alpes

SPRING

ROBOT LEARN



Its the not the Destination,
it's the journey.

Merci beaucoup, Radu !