Latent Representation Learning for Artificial Bandwidth Extension using a Conditional Variational Auto-encoder

Pramod Bachhav, Massimiliano Todisco and Nicholas Evans, EURECOM, France, emails: {bachhav, todisco, evans}@eurecom.fr

EURECOM Sophia Antipolis

Introduction

- traditional telephony infrastructure is typically limited to a bandwidth of 0.3-3.4 kHz, referred to as narrowband (NB)
- wider bandwidths generally correspond to higher quality speech
- artificial bandwidth extension (ABE) methods estimate missing highband (HB) components at 3.4-8kHz; a regression problem
- front-end *explicit memory* via neighboring speech frames augments complexity and latency of a standard regression model
- our own work addressed complexity issue using principal component analysis (PCA) [1] and semi-supervised stacked auto-encoders (SSAEs) [2]
- this work further investigates probabilistic graphical models (PGMs) for dimensionality reduction (DR), to learn better performing representations tailored specifically to ABE

Conditional variational auto-encoder (CVAE)

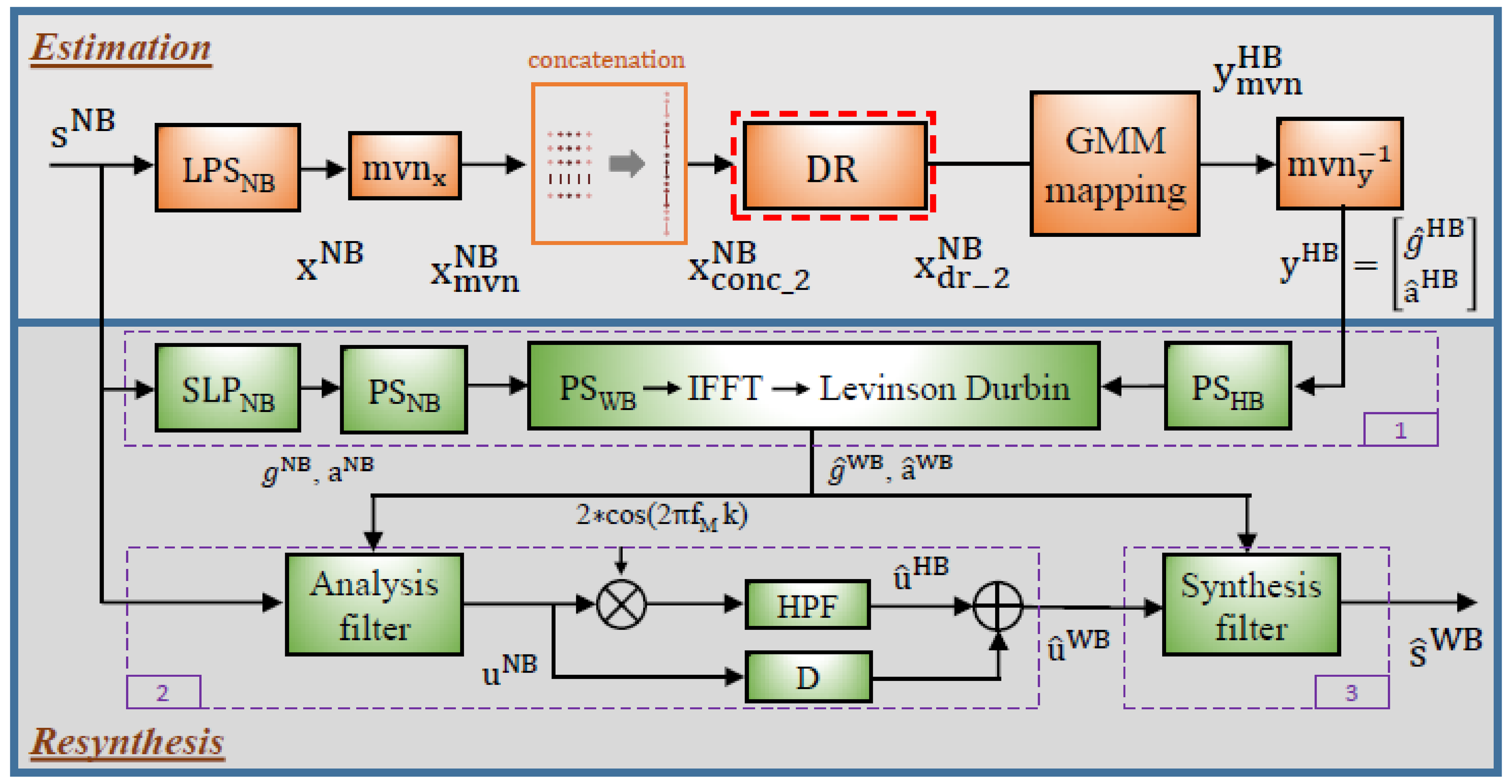
- PGMs such as VAEs and CVAEs are capable of modelling complex data distributions
- they produce probabilistic latent representations and are used to generate new data
- a CVAE is a deep generative model, $p_\theta(\mathrm{y},\mathrm{z}|\mathrm{x}) = p_\theta(\mathrm{z})p_\theta(\mathrm{y}|\mathrm{x},\mathrm{z})$ where $\mathrm{z}\sim p_\theta(\mathrm{z})$ and $\mathrm{y}\sim p_\theta(\mathrm{y}|\mathrm{x},\mathrm{z})$
- in order to maximise the conditional likelihood, $p_\theta(\mathrm{y}|\mathrm{x}) = \int p_\theta(\mathrm{z})p_\theta(\mathrm{y}|\mathrm{x},\mathrm{z})d\mathrm{z}$, CVAEs introduce a posterior distribution $q_\phi(\mathrm{z}|\mathrm{y})$ as an approximation to the intractable true posterior $p_\theta(\mathrm{z}|\mathrm{y})$
- this formulation gives the variational lower bound on likelihood, that can be optimised jointly w.r.t θ and ϕ: $\mathcal{L}(\phi,\theta;\mathrm{x},\mathrm{y}) = -D_{KL}[q_\phi(\mathrm{z}|\mathrm{y}) \,||\, p_\theta(\mathrm{z})] + E_{q_\phi(\mathrm{z}|\mathrm{y})}[\log\, p_\theta(\mathrm{y}|\mathrm{x},\mathrm{z})]$
- first term: acts as a regulariser; second term: is the negative reconstruction error
- both encoder $q_\phi(\mathrm{z}|\mathrm{y})$ and decoder $p_\theta(\mathrm{z}|\mathrm{y})$ are modelled using deep neural networks
- our initial investigations showed that vanilla VAE does not produce useful NB representations for estimation of missing HB features; signifies the importance of supervised learning

Contributions

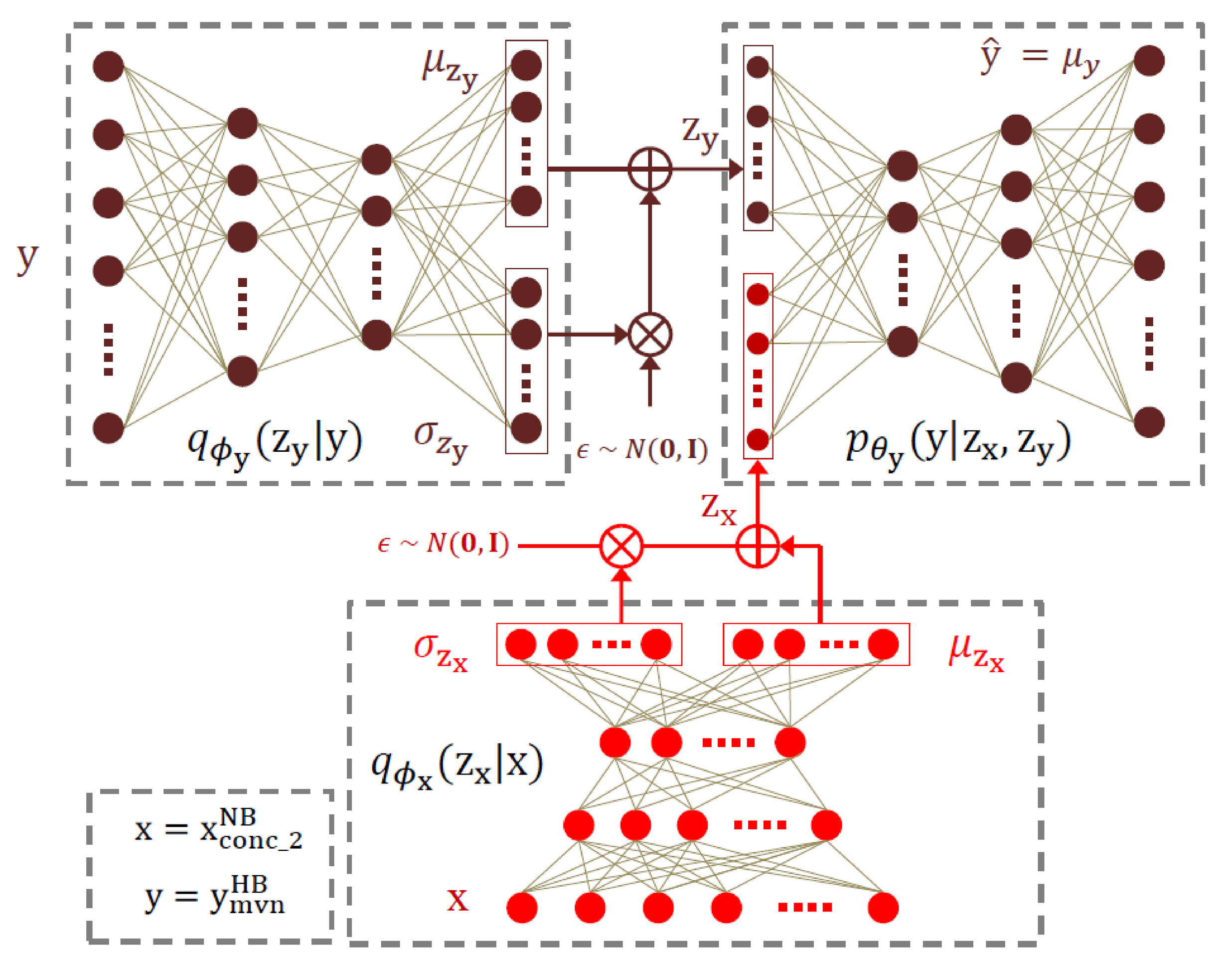
- the first application of CVAEs to DR for regression tasks such as ABE
- the combination of CVAE with a probabilistic encoder in the form of an auxiliary neural network which derives the *conditioning variable*
- the joint optimisation to extract compact probabilistic NB latent representations for estimation of missing HB components
- a thorough comparison of CVAE performance to alternative DR techniques such as PCA, SAE, SSAE

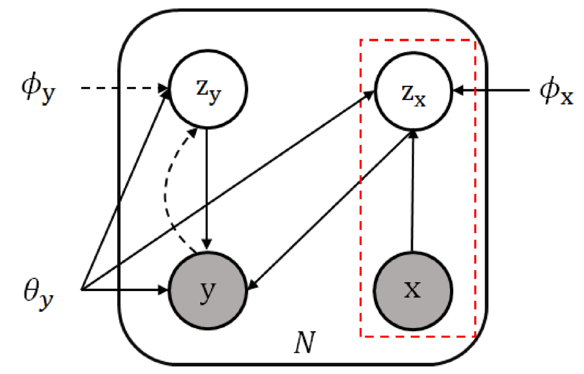
Experimental setup and results

- databases: TIMIT database divided into training (4848 utterances) and validation (192 utterances) sets. TSP speech database (1278 utterances) used as test set
- implementation details: 20ms frame duration; 10ms overlap; 1024-point FFT; square root Hann window (for analysis and synthesis)



A block diagram of the baseline ABE system.





Proposed CVAE architecture as a graphical model, dashed lines: variational inference; solid lines: generative model

Proposed dimensionality reduction/feature extraction scheme using CVAE.

VLB: $\log p_{\theta_y}(\mathrm{y}|\mathrm{z_x}) \geq \mathcal{L}(\theta_\mathrm{y},\phi_\mathrm{y},\phi_\mathrm{x};\mathrm{z_x},\mathrm{y}) = -D_{KL}[q_{\phi_\mathrm{y}}(\mathrm{z_y}|\mathrm{y})\| \,p_{\theta_\mathrm{y}}(\mathrm{z_y})] + \|\mathrm{y} - f(\mathrm{z_x},\mathrm{z_y};\theta_\mathrm{y})\|^2/\alpha$

- NB features (x^{NB}): 200-d log-power spectrum (LPS) coefficients; HB features (y^{HB}): 10 linear prediction (LP) coefficients including LP gain
- $\mathrm{x}^{NB}_{conc\_2}$ are 1000-d features obtained by concatenation of static features from 2 neighbouring frames; $\mathrm{x}^{NB}_{dr\_2}$ are 10-d features obtained after application of DR
- Mapping: GMM regression (using 128 components)

α	2	5	10	20	30
D_{KL} (training phase)	0.96	0.21	3.3e-4	1.5e-4	9.7e-5
RE (training phase)	4.73	7.40	8.93	8.97	8.97
RE (testing phase)	11.40	9.85	8.93	8.97	8.97

Effect of weighing factor α on D_{KL} and RE ($\|\mathrm{y} - f(\mathrm{z_x},\mathrm{z_y};\theta_\mathrm{y})\|^2$) during both *training* (or *reconstruction*) and *testing* (or *prediction*) phases. Results shown for the validation dataset.

DR method	$d_{RMS-LSD}$ (dB)	d_{COSH} (dB)	MOS-LQO
PCA	6.95	1.43	3.21
PCA + MVN	7.35	1.45	3.14
SAE	12.45	2.96	1.95
SAE + MVN	7.54	1.50	3.03
VAE	8.64	1.67	2.75
VAE+ MVN	8.60	1.67	2.75
SSAE	10.50	2.11	2.26
SSAE + MVN	6.80	1.34	3.28
CVAE	**6.59**	**1.31**	**3.34**
CVAE + MVN	**6.69**	**1.30**	**3.31**

Objective assessment results. $d_{RMS-LSD}$ and d_{COSH} are distance measures (lower values indicate better performance) in dB, whereas MOS-LQO values reflect quality (higher values indicate better performance), MVN – mean-variance normalisation

Comparison A → B	CMOS
CVAE → NB	0.90
CVAE→ PCA	0.13
CVAE→ SSAE + MVN	0.10
CVAE → WB	-0.96

Subjective assessment results for the ABE systems with CVAE, SSAE + MVN and PCA DR techniques in terms of CMOS

Conclusions and future work

- the first application of DR using CVAEs for ABE
- when used with standard regression ABE model, the latent, probabilistic NB features do not need any post-processing such as mean-variance normalisation
- improvements in subjective and objective results are attributed purely to the probabilistic modelling of higher dimensional spectral coefficients using CVAE
- future work should compare or combine CVAEs with other generative models such as adversarial networks

Selected References

[1] **P. Bachhav**, M. Todisco, and N. Evans, "Exploiting explicit memory inclusion for artificial bandwidth extension," *in Proc. of ICASSP*, 2018

[2] **P. Bachhav**, M. Todisco, and N. Evans, "Artificial bandwidth extension with memory inclusion using semi-supervised stacked auto-encoders," *in Proc. of INTERSPEECH*, 2018

[3] K. Sohn, et. al., "Learning structured output representation using deep conditional generative models," in Advances in *NIPS* 2015

[4] D. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv, 2013

Our implementation and speech samples are available at :

https://github.com/bachhavpramod/bandwidth_extension

http://audio.eurecom.fr/content/media

Presented by Pramod Bachhav at 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2019, Brighton, UK.