Exploring the multidimensional representation of unidimensional speech acoustic parameters extracted by deep unsupervised models

M. Jacquelin[1,2], M. Garnier[1], L. Girin[1], R. Vincent[2], O. Perrotin[1]
[1] GIPSA-Lab, Université Grenoble Alpes [2] Vogo

Context & problematic

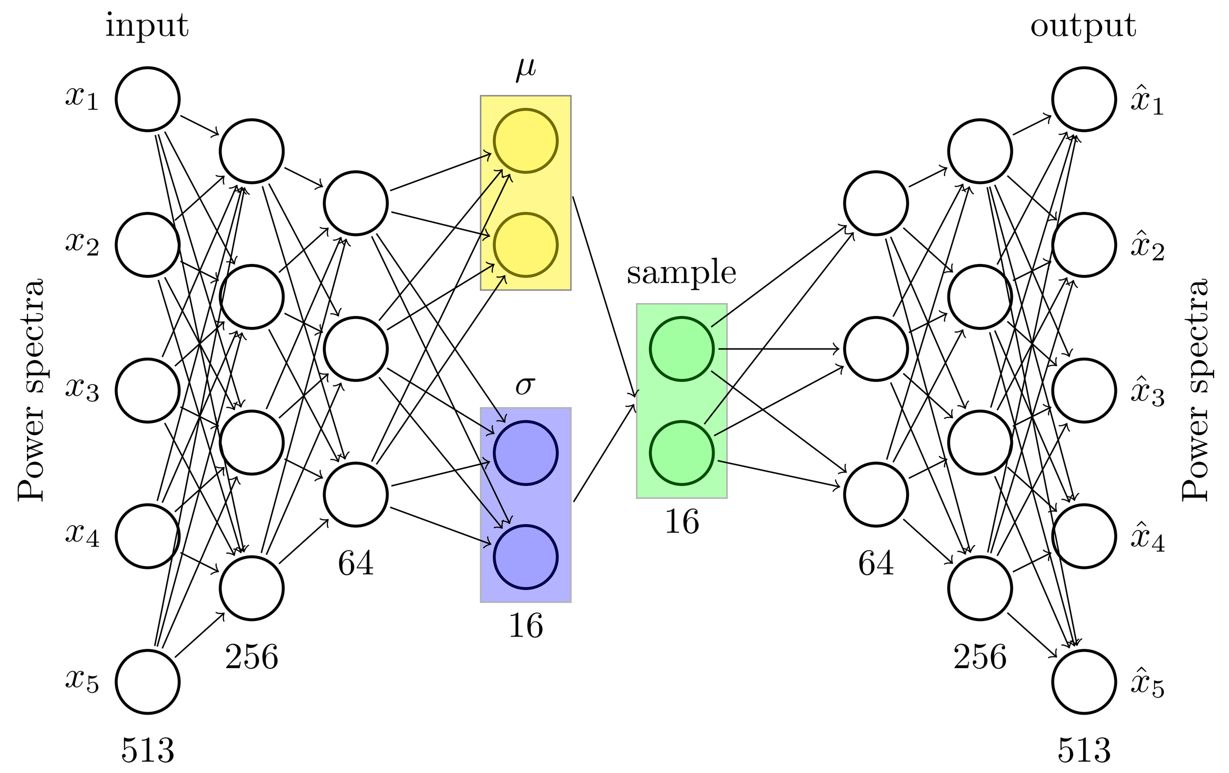
The work of [**1**] demonstrate that the fundamental frequency (f_0) and the frequency of the first three formants ($F_{1,2,3}$) are encoded in multiples dimensions in the latent space of unsupervised models. This raises the following questions:

- Why unsupervised models need multiple dimensions to encode acoustic parameters ?
- What type of information or acoustic variability is captured by each of these latent dimensions ?
- Can we control the latent space of our model to transform the variability of the acoustic parameters ?

Model used

The variational autoencoder architecture used in this work is similar to that used in [**1**]:

- The model was trained on 20 hours of VCTK on speakers not used for testing;
- The loss function is the weighted sum of the Itakura-Saito divergence;



Datas & analysis methods

Datasets:

- A natural speech test set called $D^{\text{test}}_{\text{NS},x}$ was created by extracting 3 hours from VCTK [**2**] ;
- We have built a synthetic speech test set called $D^{\text{test}}_{\text{SS},x}$ with Soundgen [**3**];
- Acoustic parameters $\mathcal{F} \in \{f_0, F_1, F_2, F_3\}$;

Methods:

- Principal Component Analysis (PCA) maximizes the variance of the projected data;
- Linear Regression (LR) estimate the relationship between variables that can be separated by classes ($\mathcal{C}$);

Signal	*Encoding*	*Latent*	*Analysis*	*Directions*
$D^{\text{test}}_{\text{SS}(\mathcal{F}),x}$	VAE	$D^{\text{test}}_{\text{SS}(\mathcal{F}),z}$	PCA	$\text{pca}_{\mathcal{F}}$
$D^{\text{test}}_{\text{NS},x}$	VAE	$D^{\text{test}}_{\text{NS},z}$	LR($\mathcal{F}$)	$\text{m}_{\mathcal{F}}$
			LR($\mathcal{F}\|\mathcal{C}$)	$\text{m}_{\mathcal{F}\|\mathcal{C}}$

Results

- **The variation of each acoustic parameter is encoded by multiple dimensions in the latent space.**

		pca_{f_0} 1	pca_{f_0} 2	pca_{f_0} 3	pca_{F_1} 1	pca_{F_1} 2	pca_{F_1} 3	pca_{F_2} 1	pca_{F_2} 2	pca_{F_2} 3	pca_{F_3} 1	pca_{F_3} 2
pca_{f_0}	1	1.00	0.00	0.00	0.33	0.49	0.74	0.07	0.26	0.58	0.12	0.54
	2	0.00	1.00	0.00	0.26	0.60	0.37	0.09	0.07	0.26	0.05	0.10
	3	0.00	0.00	1.00	0.60	0.16	0.36	0.15	0.02	0.09	0.23	0.01
pca_{F_1}	1	0.33	0.26	0.60	1.00	0.00	0.00	0.11	0.29	0.35	0.19	0.27
	2	0.49	0.60	0.16	0.00	1.00	0.00	0.24	0.14	0.66	0.35	0.33
	3	0.74	0.37	0.36	0.00	0.00	1.00	0.14	0.27	0.21	0.06	0.43
pca_{F_2}	1	0.07	0.09	0.15	0.11	0.24	0.14	1.00	0.00	0.00	0.07	0.29
	2	0.26	0.07	0.02	0.29	0.14	0.27	0.00	1.00	0.00	0.19	0.13
	3	0.58	0.26	0.09	0.35	0.66	0.21	0.00	0.00	1.00	0.11	0.51
pca_{F_3}	1	0.12	0.05	0.23	0.19	0.35	0.06	0.07	0.19	0.11	1.00	0.00
	2	0.54	0.10	0.01	0.27	0.33	0.43	0.29	0.13	0.51	0.00	1.00

Can we relate these dimensions to variations of acoustic parameters?

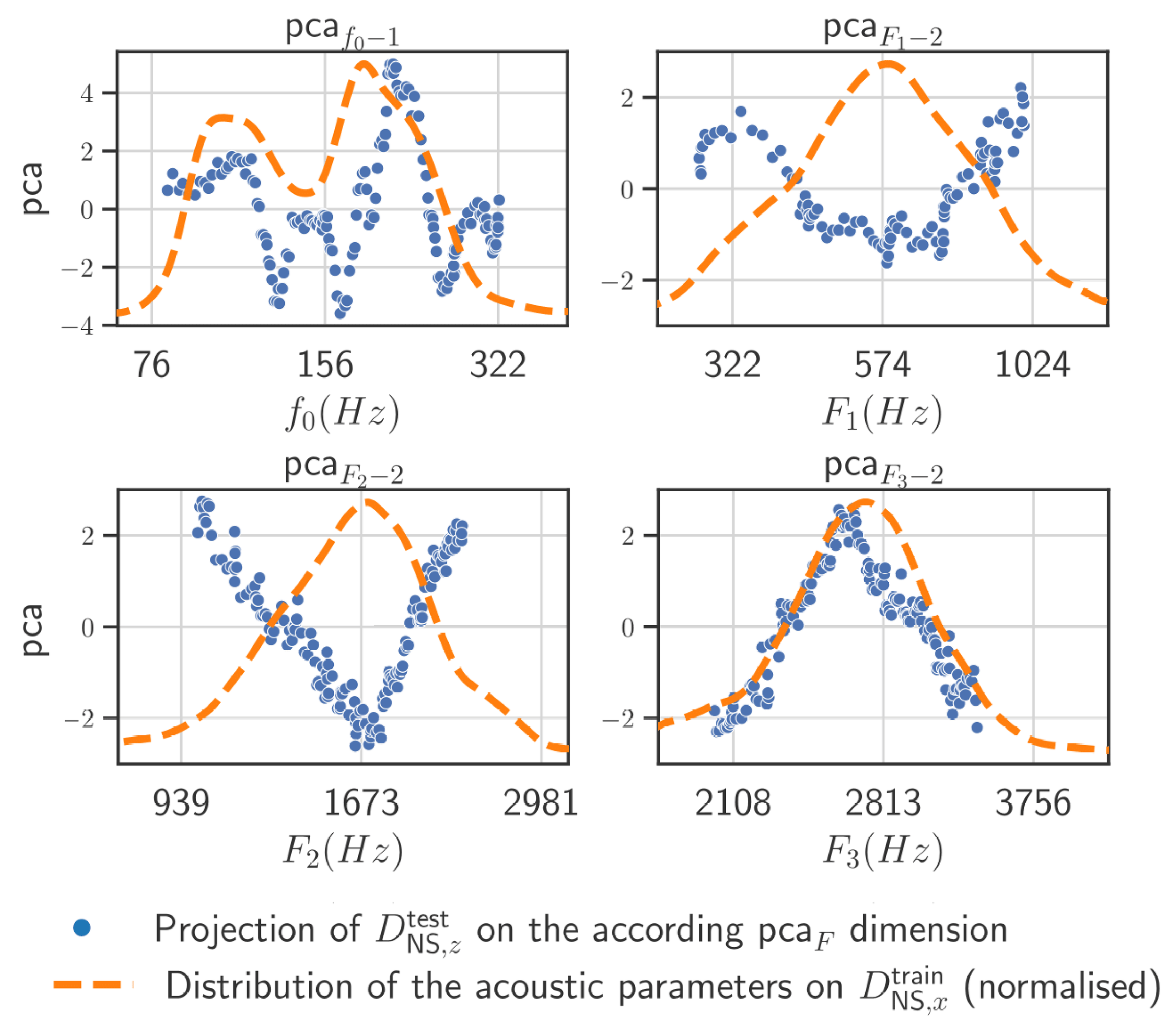
- **The multidimensional representation of acoustic parameters is closely related to the multimodality of the parameter distribution.**

	pca_{f_0} 1	pca_{f_0} 2	pca_{f_0} 3
$\text{m}_{f_0\|F}$	0.26	0.14	0.64
$\text{m}_{f_0\|M}$	0.09	0.68	0.16

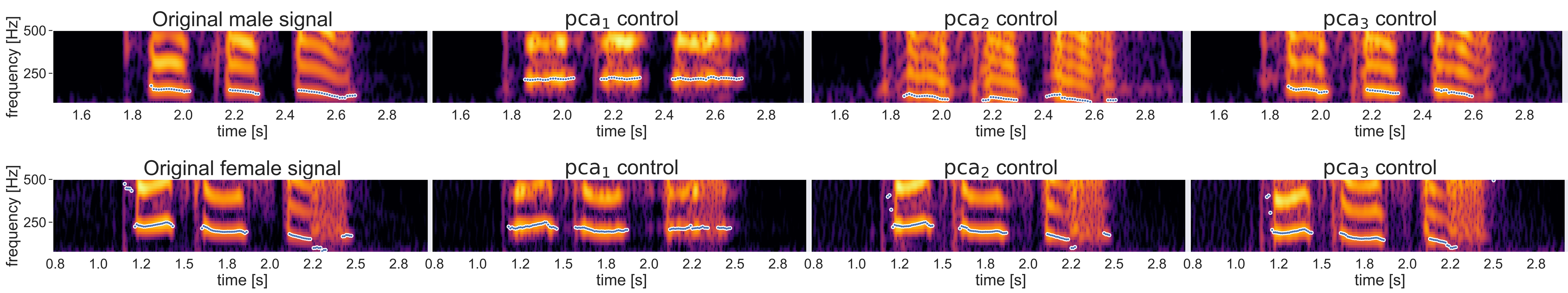
	pca_{F_1} 1	pca_{F_1} 2	pca_{F_1} 3
$\text{m}_{F_1\|F}$	0.75	0.13	0.34
$\text{m}_{F_1\|M}$	0.75	0.14	0.31

	pca_{F_2} 1	pca_{F_2} 2	pca_{F_2} 3
$\text{m}_{F_2\|F}$	0.65	0.06	0.18
$\text{m}_{F_2\|M}$	0.68	0.23	0.12

	pca_{F_3} 1	pca_{F_3} 2
$\text{m}_{F_3\|F}$	0.63	0.16
$\text{m}_{F_3\|M}$	0.61	0.17



- **The variation of acoustic parameter can be controlled directly from the encoded latent space.**



Conclusion

- ✔ **Identification of the directions that best explain the variation of selected acoustic features**
- ✔ **Highlighting the link between multimodality of parameter distribution and multidimensional representation**
- ✔ **Demonstrate the ability to control the variation of acoustic parameter with the encoded latent space**

References

[1] S. Sadok, S. Leglaive, L. Girin, X. Alameda-Pineda, R. Séguier, *Learning and controlling the source-filter representation of speech with a variational autoencoder*, in Speech Communication, 2023, vol. 148, pp. 53-65.

[2] J. Yamagishi, C. Veaux, K. MacDonald, CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit, 2019.

[3] A. Anikin, *Soundgen: an open-source tool for synthesizing non-verbal vocalizations*, in Behavior research methods, 2019, vol. 51, pp. 778–792.

Grenoble Images Parole Signal Automatique
UMR CNRS 5216 - Grenoble Campus
38400 Saint Martin d'Hères - FRANCE