Exploring the Multidimensional Representation of Unidimensional Speech Acoustic Parameters Extracted by Deep Unsupervised Models

Journée commune AFIA-TLH / AFCP

Maxime Jacquelin

Maëva Garnier, Olivier Perrotin, Rémy Vincent, Laurent Girin

CNRS, Univ. Grenoble-Alpes & Grenoble-INP
GIPSA-Lab, équipe CRISSP & Vogo, équipe Innovation

Décembre 2023

1. Introduction
2. Analysis & Results
   - Multidimensional representation of acoustic features
   - Interpretation of the learnt dimensions
   - Universal vs. speaker-specific variations
   - Control of the acoustic parameters
3. Conclusion
4. References

1 Introduction

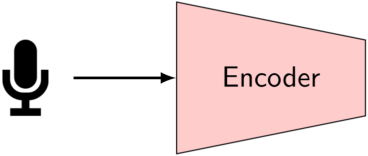
2 Analysis & Results
- Multidimensional representation of acoustic features
- Interpretation of the learnt dimensions
- Universal vs. speaker-specific variations
- Control of the acoustic parameters

3 Conclusion

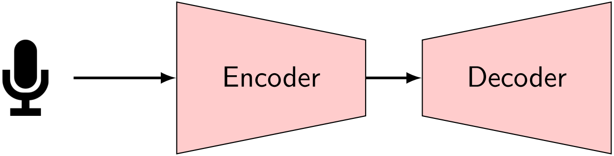
4 References

What is a Vocoder ?

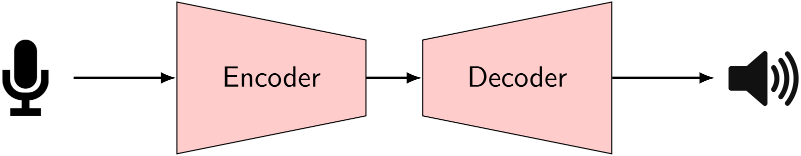
What is a Vocoder ?



What is a Vocoder ?

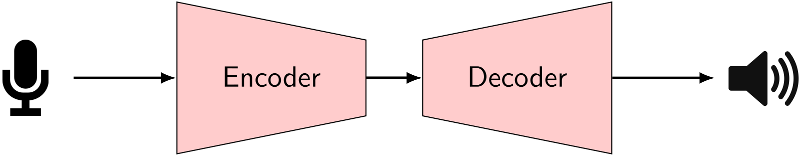


What is a Vocoder ?

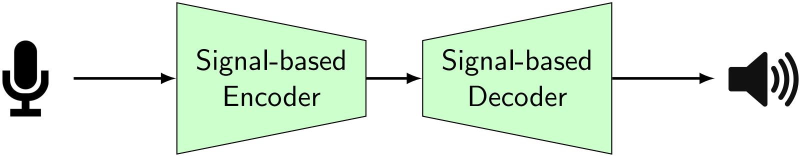


What is a Vocoder ?

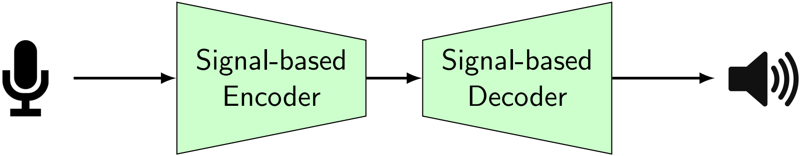
Dimensionality reduction

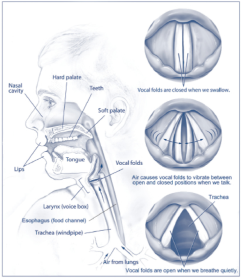


Signal-based Vocoder

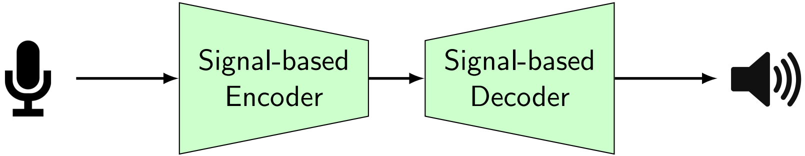


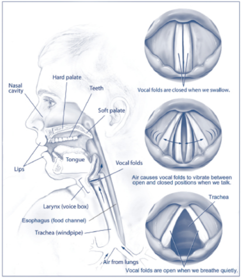
Signal-based Vocoder





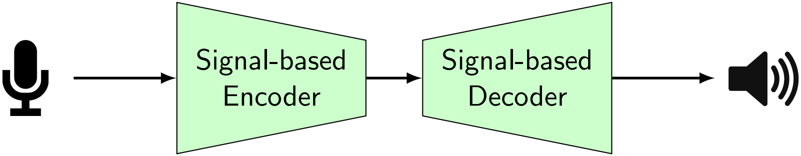
Signal-based Vocoder

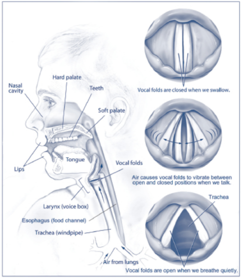




- Fundamental frequency (f_0)
- Formants frequency ($F_{1,2,3}$)

Signal-based Vocoder



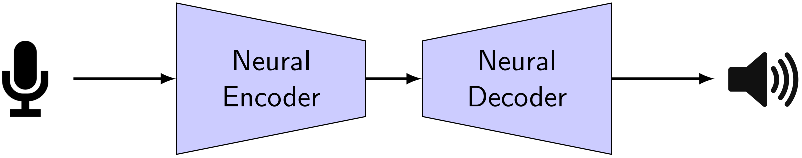


- Fundamental frequency (f_0)
- Formants frequency ($F_{1,2,3}$)

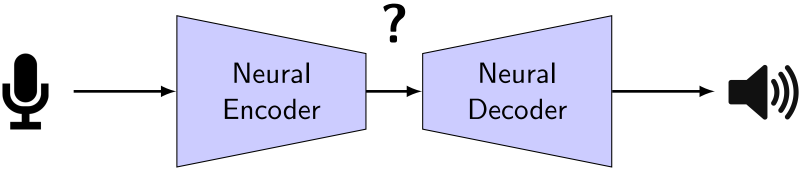
- CELP [1]
- STRAIGHT [2]
- WORLD [3]

[1] Schroeder et al., Code-excited linear prediction(CELP): High-quality speech at very low bit rates, ICASSP, 1985
[2] Kawahara et al., Restructuring speech representations using a pitch-adaptive [...] extraction, Speech communication, 1999
[3] Morise et al., WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications, IEICE, 2016

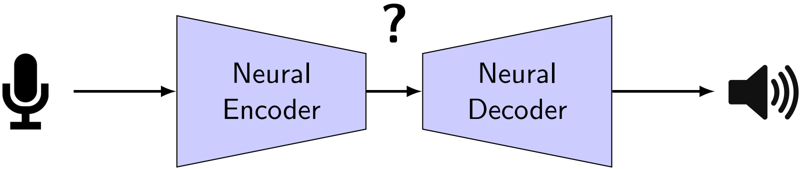
Neural Vocoder



Neural Vocoder

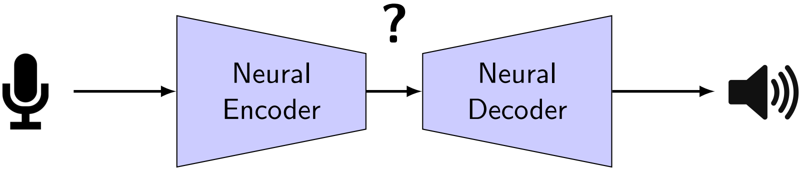


Neural Vocoder



Are acoustic parameters encoded in unsupervised models ?

Neural Vocoder



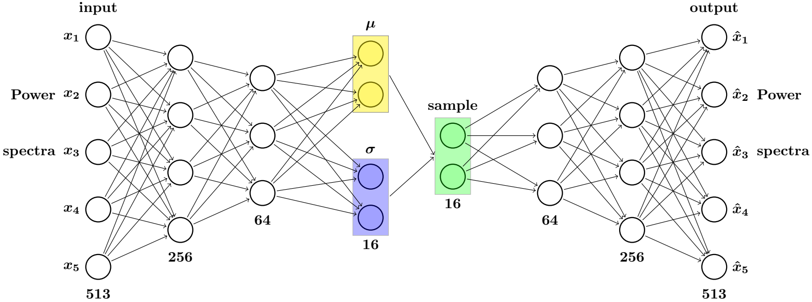
Are acoustic parameters encoded in unsupervised models ?

Sadok et all, *Learning and controlling the source-filter representation of speech with a variational autoencoder*, In Speech Communication, 2023 [4]

Neural Network

Variational autoencoder (VAE)

- Simple but powerful deep generative neural networks [5]

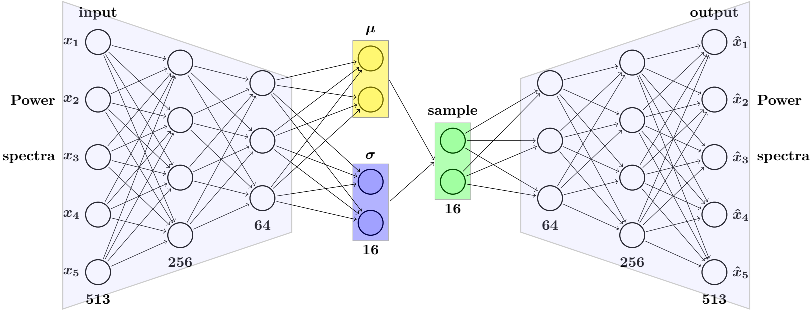


[5] Kingma et al., Auto-Encoding Variational Bayes, ICLR, 2014

Neural Network

Variational autoencoder (VAE)

- Simple but powerful deep generative neural networks [5]



[5] Kingma et al., Auto-Encoding Variational Bayes, ICLR, 2014

Neural Network

Variational autoencoder (VAE)

- Simple but powerful deep generative neural networks [5]
- [4, identify the latent subspaces encoding f_0 and the first three formant frequencies]

Neural Network

Variational autoencoder (VAE)

- Simple but powerful deep generative neural networks [5]
- [4, identify the latent subspaces encoding f_0 and the first three formant frequencies]

Why the variation of such one-dimensional feature is often explained by multiple latent dimensions ?

1. Introduction
2. **Analysis & Results**
   - Multidimensional representation of acoustic features
   - Interpretation of the learnt dimensions
   - Universal vs. speaker-specific variations
   - Control of the acoustic parameters
3. Conclusion
4. References

Multidimensional representation of acoustic features

Objective

Study the encoding of each acoustic parameter separately

Multidimensional representation of acoustic features

Methodology

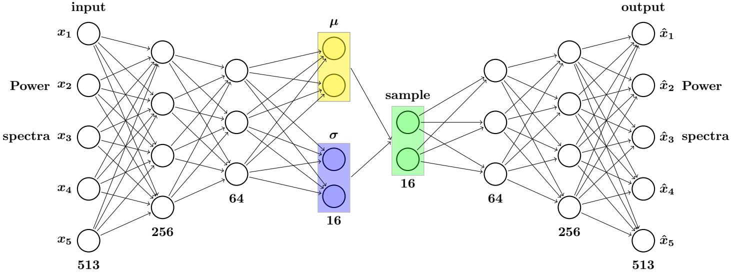
- Training dataset : VCTK [6], multi-speaker dataset, english speakers

[6] Junichi et al., VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit, 2019

Multidimensional representation of acoustic features

Methodology

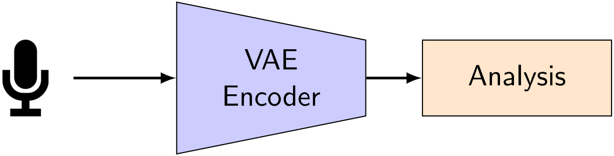
- Training dataset : VCTK [6], multi-speaker dataset, english speakers
- Model based on previous works [4]



Multidimensional representation of acoustic features

Methodology

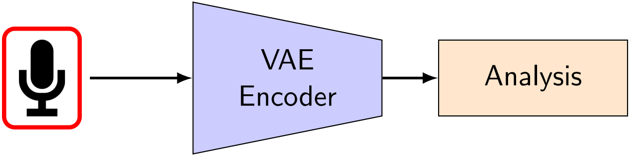
- Training dataset : VCTK [6], multi-speaker dataset, english speakers
- Model based on previous works [4]
- Linear analysis methods to identify the directions in the latent space that capture the variability for each acoustic parameter



Multidimensional representation of acoustic features

Methodology

- Training dataset : VCTK [6], multi-speaker dataset, english speakers
- Model based on previous works [4]
- Linear analysis methods to identify the directions in the latent space that capture the variability for each acoustic parameter

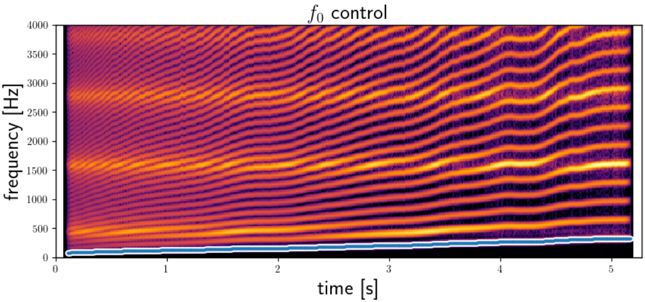


Study the encoding of each acoustic parameter separately

Multidimensional representation of acoustic features

Proposed method

- Soundgen [7] : generate signals with variation of either f_0, F_1, F_2, F_3 while the three other parameters remained constant

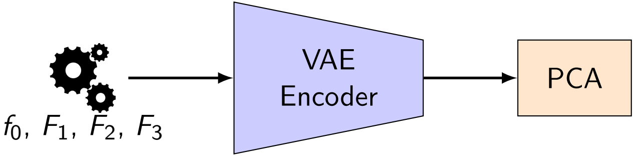


[7] Anikin et al., Soundgen: an open-source tool for synthesizing nonverbal vocalizations, Behavior research methods, 2019

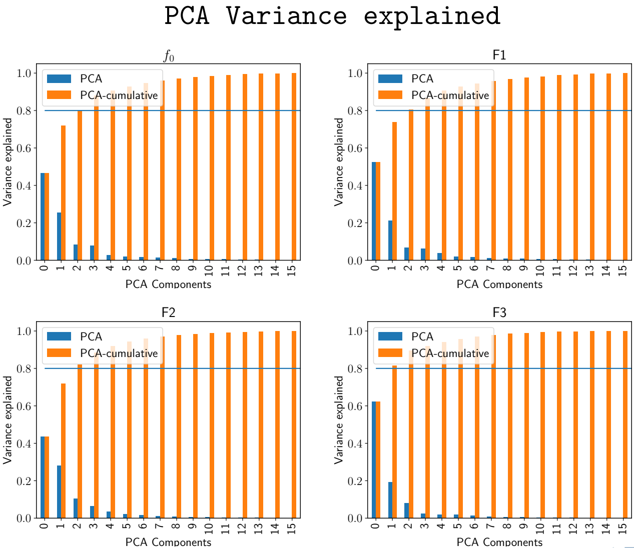
Multidimensional representation of acoustic features

Proposed method

- Soundgen [7] : generate signals with variation of either f_0, F_1, F_2, F_3 while the three other parameters remained constant
- Principal Components Analysis (PCA) : identify the directions that explain the variation of the acoustic parameter

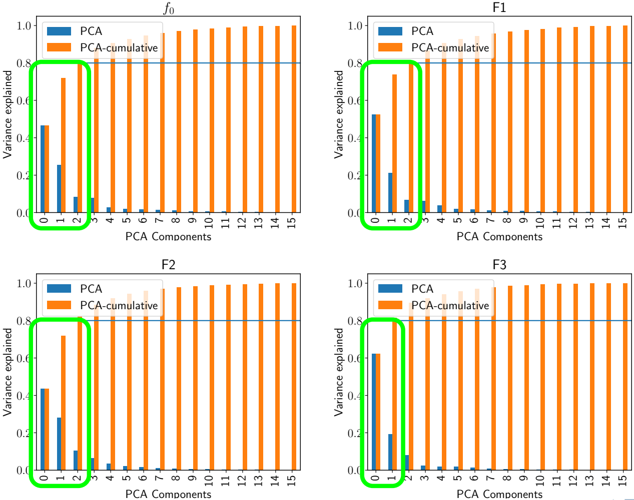


Multidimensional representation of acoustic features

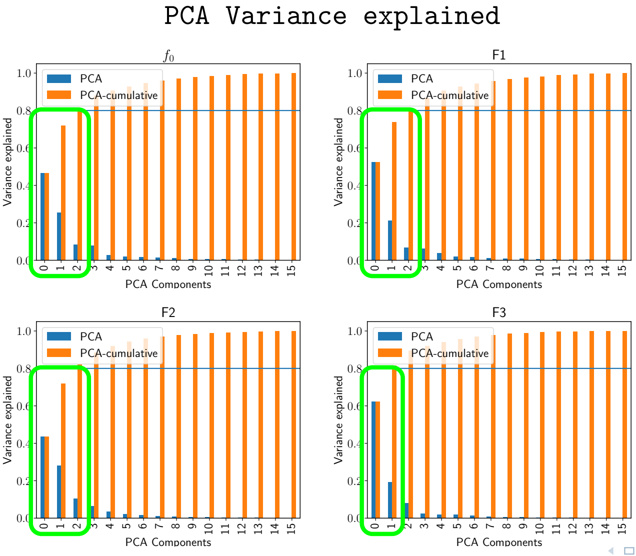


Multidimensional representation of acoustic features

PCA Variance explained

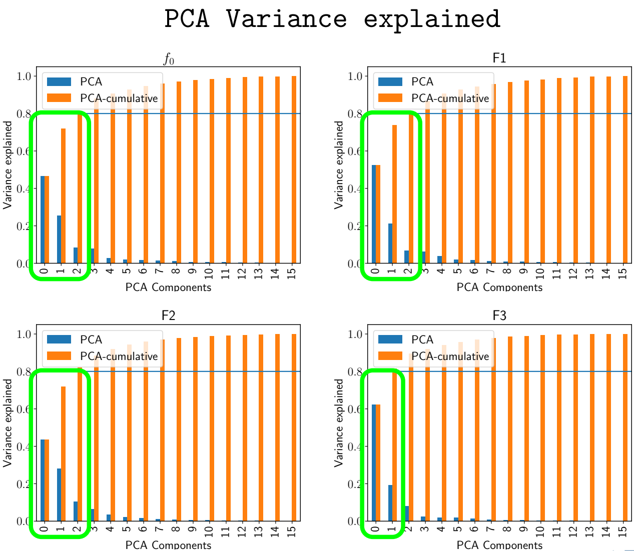


Multidimensional representation of acoustic features



- Each acoustic parameter is encoded by multiple dimensions.

Multidimensional representation of acoustic features



- Each acoustic parameter is encoded by multiple dimensions.
- What kind of information is encoded in each component ?

Interpretation of the learnt dimensions

Objective

Identify the role of these multiple dimensions, through the analysis of natural speech

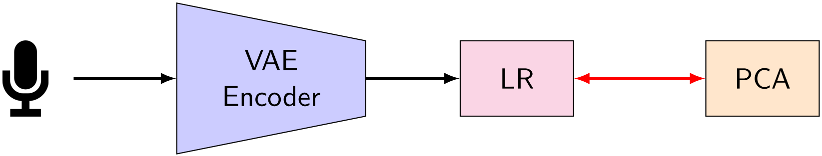
Hypothesis

The different latent dimensions reflect sources of inter- and intra-individual variability of each acoustic parameter

Interpretation of the learnt dimensions

Proposed method

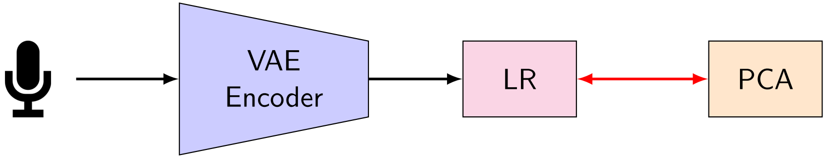
- VCTK [6] : multi-speaker dataset, english speakers not seen during training
- Linear Regression (LR) : analyze the variation of specific acoustic parameters in the natural test set



Interpretation of the learnt dimensions

Proposed method

- VCTK [6] : multi-speaker dataset, english speakers not seen during training
- Linear Regression (LR) : analyze the variation of specific acoustic parameters in the natural test set
- Analyse the possible representation of gender-related acoustic parameters



Interpretation of the learnt dimensions

Cosine similarity between LR and PCA

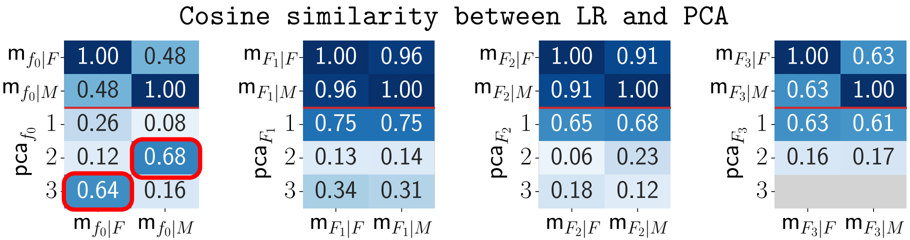
		$m_{f_0\|F}$	$m_{f_0\|M}$
$m_{f_0\|F}$		1.00	0.48
$m_{f_0\|M}$		0.48	1.00
pca_{f_0}	1	0.26	0.08
	2	0.12	0.68
	3	0.64	0.16

		$m_{F_1\|F}$	$m_{F_1\|M}$
$m_{F_1\|F}$		1.00	0.96
$m_{F_1\|M}$		0.96	1.00
pca_{F_1}	1	0.75	0.75
	2	0.13	0.14
	3	0.34	0.31

		$m_{F_2\|F}$	$m_{F_2\|M}$
$m_{F_2\|F}$		1.00	0.91
$m_{F_2\|M}$		0.91	1.00
pca_{F_2}	1	0.65	0.68
	2	0.06	0.23
	3	0.18	0.12

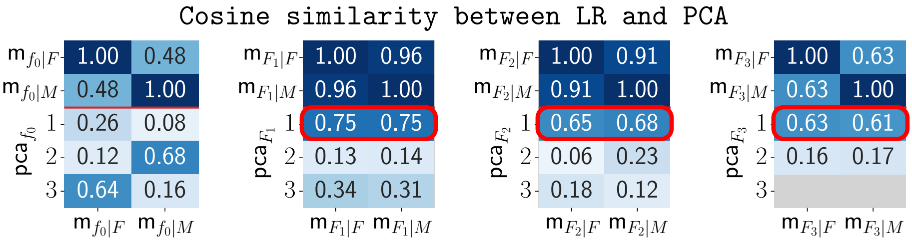
		$m_{F_3\|F}$	$m_{F_3\|M}$
$m_{F_3\|F}$		1.00	0.63
$m_{F_3\|M}$		0.63	1.00
pca_{F_3}	1	0.63	0.61
	2	0.16	0.17
	3		

Interpretation of the learnt dimensions



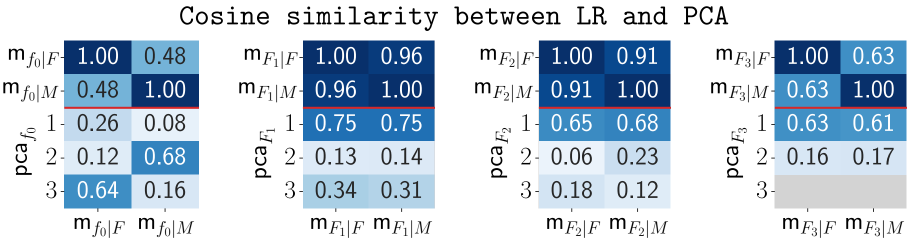
- For f_0 : each gender is encoded in a distinct component.

Interpretation of the learnt dimensions



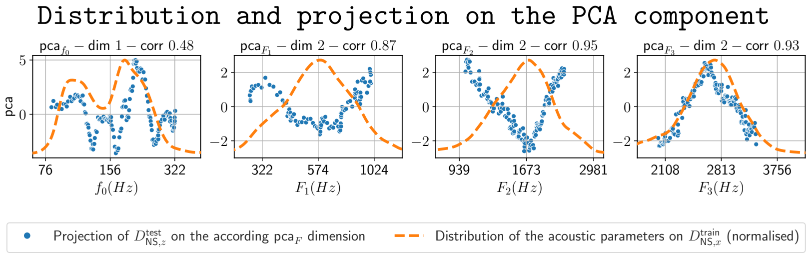
- For f_0 : each gender is encoded in a distinct component.
- For $F_{1,2,3}$: both genders are encoded in the same component.

Interpretation of the learnt dimensions

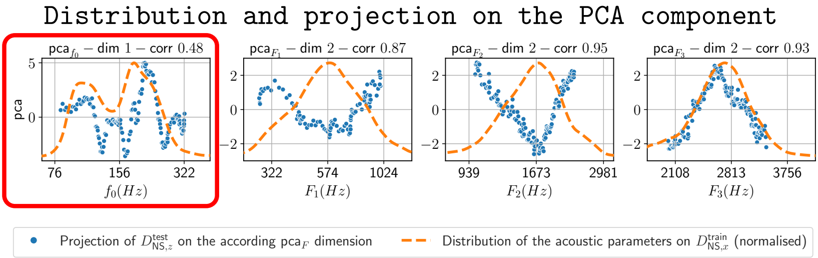


- For f_0 : each gender is encoded in a distinct component.
- For $F_{1,2,3}$: both genders are encoded in the same component.
- Why doesn't the model encode the fundamental frequency and the formants the same way?

Interpretation of the learnt dimensions

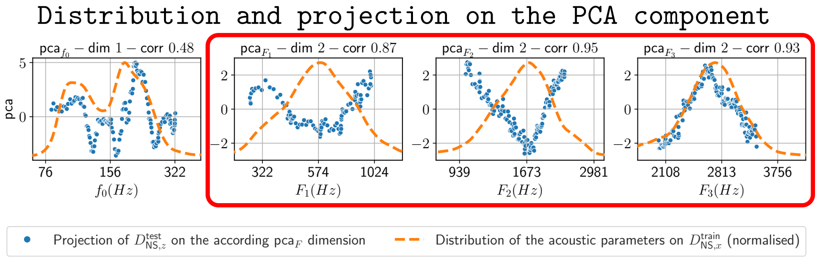


Interpretation of the learnt dimensions



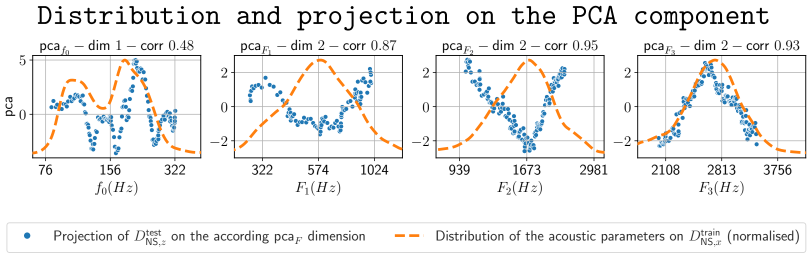
- For f_0 : the bimodal distribution is the most correlated with the first component

Interpretation of the learnt dimensions



- For f_0 : the bimodal distribution is the most correlated with the first component
- For $F_{1,2,3}$: the unimodal distribution is the most correlated with the second component.

Interpretation of the learnt dimensions



- For f_0 : the bimodal distribution is the most correlated with the first component
- For $F_{1,2,3}$: the unimodal distribution is the most correlated with the second component.
- The multidimensional representation of a single acoustic parameter is closely related to the multimodality of the parameter distribution.

Interpretation of the learnt dimensions

Objective

Identify a disentangled representation of inter- and intra-individual variability in the latent space

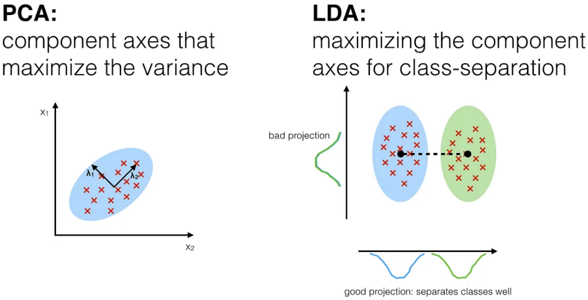
Hypothesis

A linear combination of latent dimensions that best discriminates the speakers, should display an inter-gender direction on its first component, and thus an intra-gender direction on remaining components

Universal vs. speaker-specific variations

Proposed method

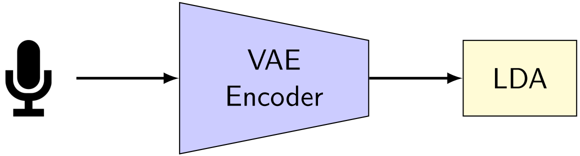
- VCTK [6] : multi-speaker dataset, english speakers not seen during training
- Linear Discriminant Analysis (LDA) : underline the model's ability to disentangle inter- and intra-individual variability



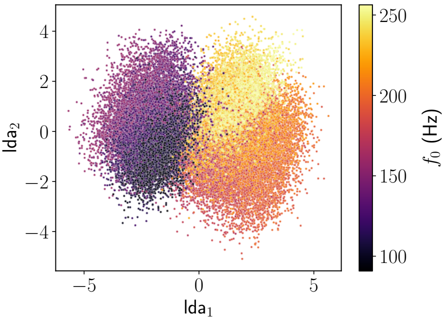
Universal vs. speaker-specific variations

Proposed method

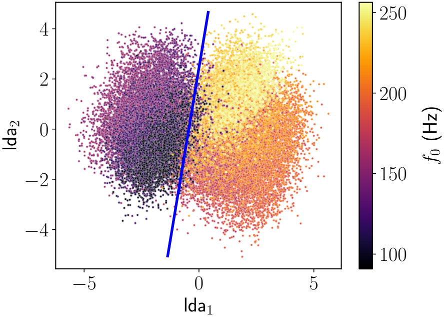
- VCTK [6] : multi-speaker dataset, english speakers not seen during training
- Linear Discriminant Analysis (LDA) : underline the model's ability to disentangle inter- and intra-individual variability



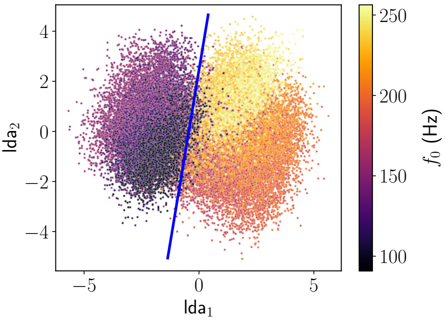
Universal vs. speaker-specific variations



Universal vs. speaker-specific variations

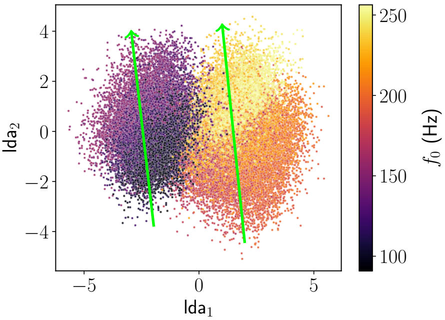


Universal vs. speaker-specific variations



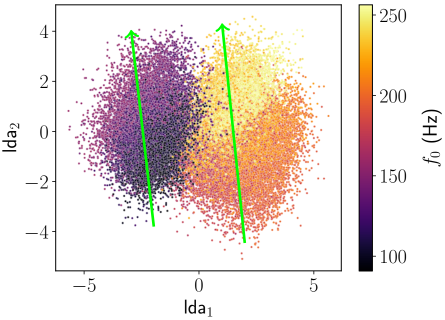
- The first component models the inter-gender variation of f_0.

Universal vs. speaker-specific variations



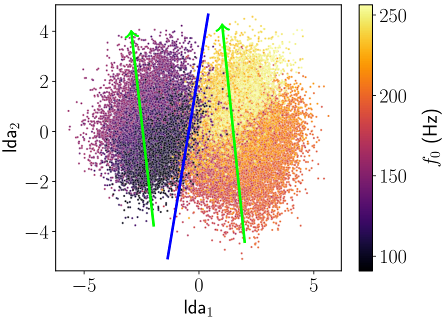
- The first component models the inter-gender variation of f_0.

Universal vs. speaker-specific variations



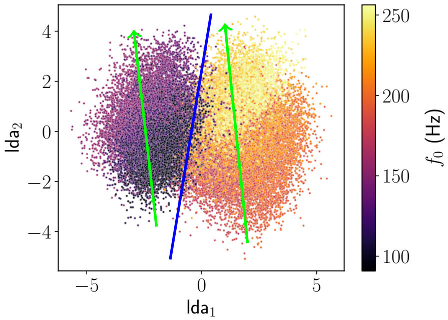
- The first component models the inter-gender variation of f_0.
- The second component models the intra-gender variation of f_0.

Universal vs. speaker-specific variations



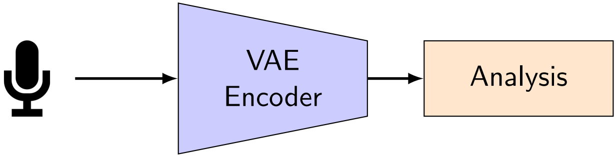
- The first component models the inter-gender variation of f_0.
- The second component models the intra-gender variation of f_0.

Universal vs. speaker-specific variations

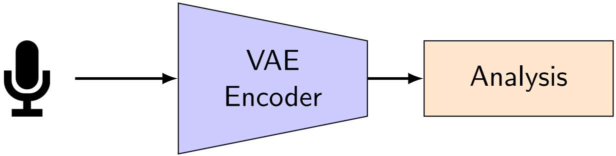


- The first component models the inter-gender variation of f_0.
- The second component models the intra-gender variation of f_0.
- The model is able to disentangle inter- and intra-gender variations along two distinct directions.

Control of the acoustic parameters

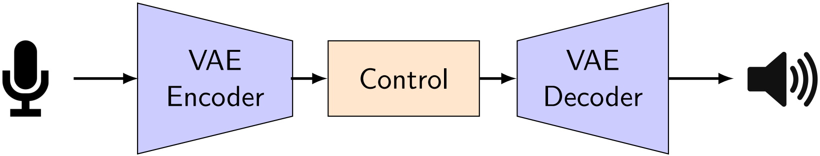


Control of the acoustic parameters



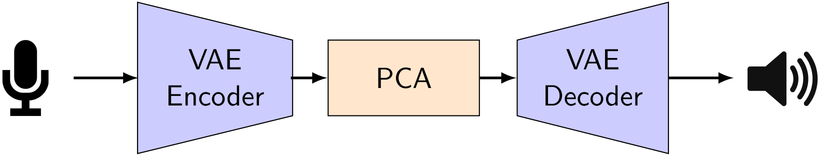
Can we use those methods to control the acoustic parameters values ?

Control of the acoustic parameters

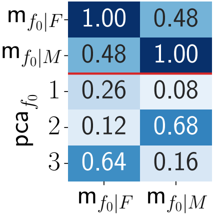


Can we use those methods to control the acoustic parameters values ?

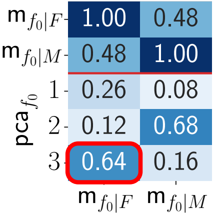
Control intra

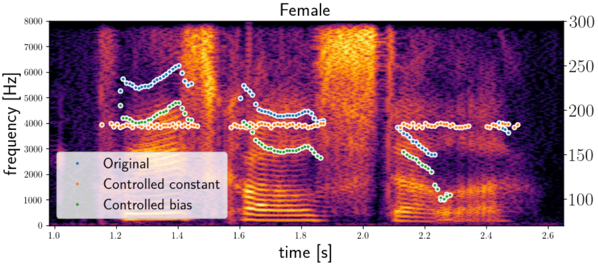


Control intra

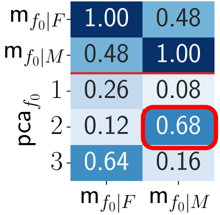


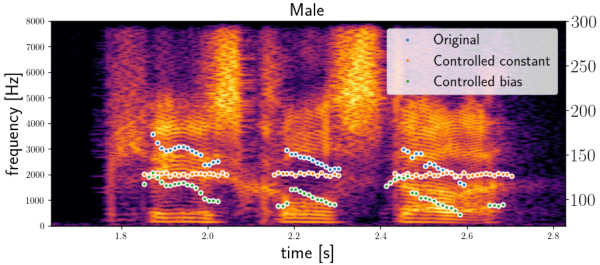
Control intra



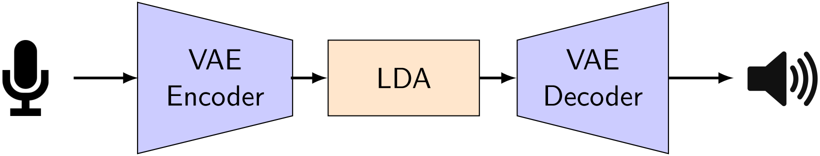


Control intra

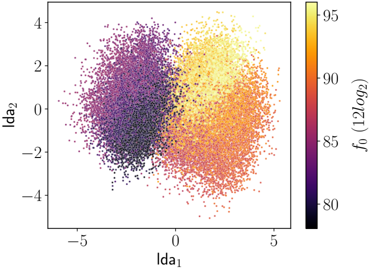




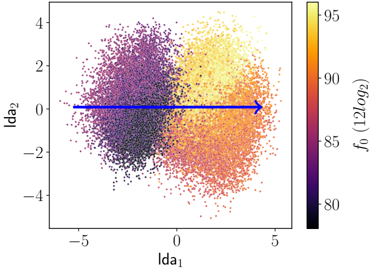
Control inter

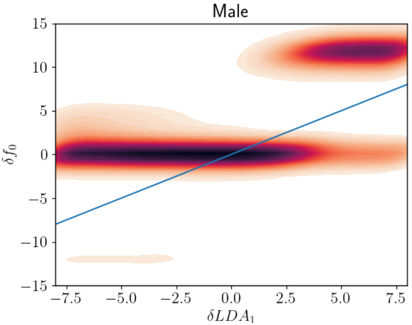


Control inter

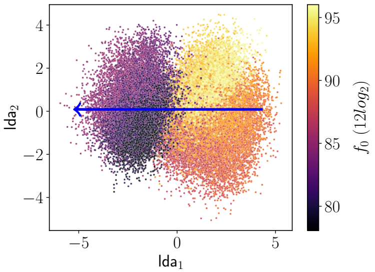


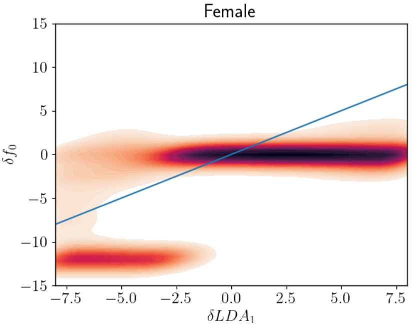
Control inter





Control inter





Conclusion

- We proposed a method for interpreting the latent space of a variational autoencoder trained on a multi-speaker database.

Conclusion

- We proposed a method for interpreting the latent space of a variational autoencoder trained on a multi-speaker database.
- We demonstrated that one of these dimensions encodes the global shape of the distribution of each acoustic parameter over the training set.

Conclusion

- We proposed a method for interpreting the latent space of a variational autoencoder trained on a multi-speaker database.
- We demonstrated that one of these dimensions encodes the global shape of the distribution of each acoustic parameter over the training set.
- We identified the directions in latent space that explain the between-mode and within-mode variation of the acoustic parameter.

Conclusion

- We proposed a method for interpreting the latent space of a variational autoencoder trained on a multi-speaker database.
- We demonstrated that one of these dimensions encodes the global shape of the distribution of each acoustic parameter over the training set.
- We identified the directions in latent space that explain the between-mode and within-mode variation of the acoustic parameter.
- We controlled the variation of fundamental frequency between-mode and within-mode.

Future works

- Increase the number of features of interest and propose a disentangled representation.

Future works

- Increase the number of features of interest and propose a disentangled representation.
- Evaluate the effect of a more expressive training dataset on the observed results.

Future works

- Increase the number of features of interest and propose a disentangled representation.
- Evaluate the effect of a more expressive training dataset on the observed results.
- Apply this method to other types of unsupervised or self-supervised models.

Thank you for your attention

1. Introduction
2. Analysis & Results
   - Multidimensional representation of acoustic features
   - Interpretation of the learnt dimensions
   - Universal vs. speaker-specific variations
   - Control of the acoustic parameters
3. Conclusion
4. References

References I

[1] M. Schroeder and B. Atal, "Code-excited linear prediction(CELP): High-quality speech at very low bit rates," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Tampa, FL, USA, 1985, pp. 937–940.

[2] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[3] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," in *IEICE Transactions on Information and Systems*, vol. E99.D, 2016, pp. 1877–1884.

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

[5] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, Banff, Canada, 2014.

[6] Y. Junichi, V. Christophe, and M. Kirsten, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

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