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

Maxime Jacquelin, Maëva Garnier, Laurent Girin, Rémy Vincent, Olivier Perrotin

12th ISCA Speech Synthesis Workshop (SSW2023)
26-28 August 2023, Grenoble, France

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

Maxime Jacquelin[1,2], *Maëva Garnier*[1], *Laurent Girin*[1], *Rémy Vincent*[2], *Olivier Perrotin*[1]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, F-38000 Grenoble, France
[2]Vogo, F-38190 Bernin, France

maxime.jacquelin@grenoble-inp.fr, maeva.garnier@grenoble-inp.fr,
laurent.girin@grenoble-inp.fr, r.vincent@vogo-group.com, olivier.perrotin@grenoble-inp.fr

Abstract

Understanding latent representations of speech by unsupervised models enables powerful signal analysis, transformation, and generation. Numerous studies have identified directions of variation of acoustic features such as fundamental frequency or formants in unsupervised models latent spaces, but it is yet not well understood why the variation of such one-dimensional features is often explained by multiple latent dimensions. This paper proposes a methodology for interpreting these dimensions, in the latent space of a variational autoencoder trained on a multi-speaker database.

Index Terms: representation learning, speech encoding, variational autoencoder, source-filter model

1. Introduction

To model speech effectively, it is necessary to identify the dimensions, preferably limited in number and independent from each other, which best explain the wide acoustic variations in speech observed within individual's speech production modes, and from one individual to another.

While unsupervised models are increasingly powerful statistical tools to model speech signals, research on understanding their latent representations to highlight the full complexity of interactions between acoustic parameters is at its beginning. A recent study used the variational autoencoder (VAE) [1] to find a latent space that best represents acoustic variations [2]. They showed that, even with a unsupervised learning approach, the dimensions of the VAE latent space actually corresponded to the quasi-orthogonal encoding of the fundamental frequency f_0 and the frequency of the first three formants. An unexplained phenomenon, however, was that each acoustic parameter was encoded by several VAE dimensions. The question of what type of information or acoustic variability is captured by each of these latent dimensions remains open. Among the multiple interactions between acoustic parameters described earlier, our hypothesis, explored in this study, is that the different dimensions may reflect the different sources of inter- and intra-individual variability of each acoustic parameter.

We therefore introduce a methodology for analysing and interpreting the multi-dimensional aspect of the representation of single-dimension acoustic parameters, before testing our hypothesis.

2. Methodology

Data used : Two sets of tests were designed to evaluate the modeling capabilities of our VAE on the fundamental frequency (f_0) and the first three formants ($F_{1,2,3}$).

To demonstrate the multidimensional representation of individual acoustic parameters, we constructed a set of synthetic speech tests in which variations in acoustic parameters are isolated and controlled, following [2]. We generate four signals of $5\,\mathrm{second}$ duration called $D^{\mathsf{test}}_{\mathsf{SS}(\mathcal{F}),x}$, $\mathcal{F} \in \{f_0, F_1, F_2, F_3\}$, with variations of f_0, F_1, F_2 or F_3, while the other parameters remain constant. To observe the VAE's acoustic modeling capabilities by including natural covariations between acoustic parameters, a natural speech test set called $D^{\mathsf{test}}_{\mathsf{NS},x}$ was used, as a subset of VCTK dataset and comprising 3 hours of speech signals (approximately $10\,\%$ of the database) that were used neither for training nor in the validation set.

Model used : The VAE architecture used in this work is similar to that used in [2] and was trained on speech signals extracted from the VCTK dataset.

Analysis protocol : As each signal $D^{\mathsf{test}}_{\mathsf{SS}(\mathcal{F}),x}$ include the variation of an individual acoustic parameter $\mathcal{F} \in \{f_0, F_1, F_2, F_3\}$, it ensure that the variations in the VAE latent space corresponding to the encoded signal $D^{\mathsf{test}}_{\mathsf{SS}(\mathcal{F}),z}$ were representative only of that parameter. In this case, for each encoded synthetic test signal $D^{\mathsf{test}}_{\mathsf{SS}(\mathcal{F}),z}$, we applied a principal component analysis (PCA) in order to identify the directions in latent space (denoted $\mathsf{pca}_{\mathcal{F}}$) that were most relevant in explaining the variation of the corresponding acoustic parameter $\mathcal{F}$.

Applying a PCA to the encoding of the natural test set $D^{\mathsf{test}}_{\mathsf{NS},z}$ would provide the directions of maximum variation, but without any indication of the acoustic parameters he describe. To accurately analyze the variation of specific acoustic parameters in the natural test set, we therefore instead used linear regression (LR) to predict each acoustic parameter previously extracted with Praat on the test set, as a linear combination of the 16 dimensions of the latent space. We denote $\mathsf{m}_{\mathcal{F}}$ the directions of variation (DV) of each $\mathcal{F}$ parameter defined as the vector of LR coefficients. We also studied parameter variation within gender classes of speakers . In this case, we denote the DVs $\mathsf{m}_{\mathcal{F}|M}$ and $\mathsf{m}_{\mathcal{F}|F}$ for male and female, respectively.

Furthermore, to underline the model's ability to disentangle inter- and intra-individual variability, in our study we also performed a linear discriminant analysis (LDA) on $D^{\mathsf{test}}_{\mathsf{NS},z}$, noted lda_g. As gender is one of the most discriminating features of speech, we hypothesized that an LDA on speakers should identify an inter-gender direction on its first component, and therefore an intra-gender direction on the other components.

Finally, to assess whether the variation identified by PCA coincides with the relationships modeled by LR and LDA, we examined the collinearity of the directions extracted by PCA, LR or LDA, by calculating the cosine distance between $\mathsf{pca}_{\mathcal{F}}$, $\mathsf{m}_{\mathcal{F}}$, $\mathsf{m}_{\mathcal{F}|M}$, $\mathsf{m}_{\mathcal{F}|F}$, and lda_g.

This Late Breaking Report of the Speech Synthesis Workshop 2023 was not peer-reviewed

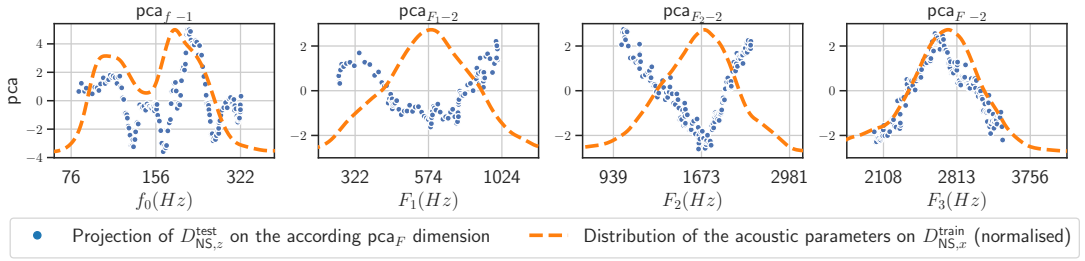

Figure 1: *Comparison of the shape of the acoustic parameter distributions on the test set and the most correlated PC component.*

3. Results

Multidimensional representation of acoustic parameters : Separate PCAs applied to the synthetic signals of the test set $D^{test}_{SS(\mathcal{F}),z}$ showed that all parameters require three principal components (PCs) to explain at least 80 % of the variance, with the exception of F_3 (two PCs). We observe that all the first PCs of each parameter are relatively orthogonal to each other, with a maximum value of 0.32 between pca_{f_0} and pca_{F_1}. In contrast, the correlation between parameters for the last PCs is of a higher order, with a maximum value of 0.68 between the first PC of pca_{f_0} and the third PC of pca_{F_2}.

Overall, these results highlight a multidimensional representation of acoustic parameters in the VAE latent space. Analysis of the collinearities between the PCs extracted through the parameters showed that VAE creates a balance between, on the one hand, the pseudo-independent source (f_0) and the filter ($F_{1,2,3}$) on their first PC, and on the other hand the modeling of well-known co-variations between acoustic parameters.

Interpretation of learned dimensions : To interpret the significance of each dimension, we then studied the variation of each parameter on the natural test set $D^{test}_{NS,z}$, and observed the correlations of these directions with those observed on the synthetic test set. These directions, denoted $\mathsf{m}_{\mathcal{F}}$, are extracted from the DV LR of each parameter on $D^{test}_{NS,z}$and calculated by gender. For all formants, the first PC is most correlated with LRs of both genders of the parameter. Whereas for f_0, the DV of $\mathsf{m}_{f_0|M}$ and $\mathsf{m}_{f_0|F}$ correlates with the second and third PCs of pca_{f_0}, respectively.

Let's recall that in the computation of the PCs from the synthetic test set $D^{test}_{SS(f_0),z}$, no parameters other than f_0 vary in the input signal. Yet, analysis of correlations between $\mathsf{m}_{\mathcal{F}|M}$, $\mathsf{m}_{\mathcal{F}|F}$ and pca_{f_0} showed that the DV of f_0 for male and female correlate with the second and third PCs of pca_{f_0}, respectively. These results mean that the VAE is able to disentangle f_0 values that are more likely to belong to male or female speakers from $D^{test}_{SS(f_0),z}$ alone. We hypothesise that the model has learned the bimodal distribution of f_0 values encountered in the training set, and is able to sort synthetic frames based on this distribution. To test this hypothesis, we investigate whether the distribution is encoded in the latent space by computing the correlation between the f_0 distribution measured on the training set, and the $D^{test}_{SS(f_0),z}$ values taken on each pca_{f_0} PC. We obtained the highest correlation with the first PC of pca_{f_0} (0.46), that is shown on the left of Fig. 1.

We find that the two main peaks of $D^{test}_{SS(f_0),z}$ on the first PC of pca_{f_0} are close to the median of the f_0 distribution of both genders. Furthermore, pca_{f_0} values are high when the f_0 distribution is high, while they are close to 0 when the two modes merge, thus modeling the uncertainty of classification between male or female groups. The same behavior is observed for the three formants. For each of them, the second PC represented in Fig. 1 is the most correlated with the distribution of formants, reaching correlation values above 0.9 for each of them. In this case, the distribution is uni-modal, which coherently explains the correlation of a single PC with the formant value.

Overall, we have shown that the multidimensional representation of single acoustic parameters is closely related to the multimodality of the parameter distribution. For each parameter, one PC encodes the parameter distribution that is learned from the training set, and acoustic parameter values that belong to different modes are encoded by a few other distinct PCs.

Universal or speaker-specific variations : The repreentation of each mode of the distribution of a single parameter (in our case f_0) into distinct directions of the latent space raises the question of control: should each mode (each genre) be treated independently, or can we find new directions that model intra- and inter-genre variation, allowing for inter- and intra-class control? We calculated the mean value of f_0 per gender on the test set $D^{test}_{NS,z}$, and derived a gender-centered value of f_0, called $< f_0 >_{\mathcal{C}}$. $\mathsf{m}_{<f_0>_{\mathcal{C}}}$ is the DV of the LR of this parameter on $D^{test}_{NS,z}$We also performed an LDA on $D^{test}_{NS,z}$to find the linear combination of parameters that best discriminates the speaker.

Cosine similarities between the regressions and the first two components of lda_g, highlight the good correlation (0.86) between the first component of lda_g and m_{f_0}, which includes gender information. Meanwhile, the second component lda_g correlates well with $< f_0 >_{\mathcal{C}}$ (cosine similarity of 0.77), in which gender bias is removed. Consistently, we also observe that the DVs calculated by gender ($\mathsf{m}_{f_0|M}$ and $\mathsf{m}_{f_0|F}$) are more correlated with the second component lda_g. These results converge with our hypothesis that intra- and inter-generic information are encoded along distinct LDA directions.

4. Conclusions

After showing that the variation of each parameter is encoded by multiple dimensions in the latent space of a VAE, we demonstrated that one of these dimensions encodes the global shape of the distribution of each acoustic parameter over the training set. Then, if the distribution is multimodal, we have identified the directions in latent space that explain the between-mode and within-mode variation of the acoustic parameter.

We believe that this methodology could easily be applied to other types of unsupervised models, and thus is a step forward in understanding the latent representation of unsupervised models in order to build powerful tools for the analysis and the controllable generation of speech signals.

5. References

[1] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 2014.

[2] 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*, 2023, pp. 53–65.