LEARNING AND CONTROLLING THE SOURCE-FILTER REPRESENTATION OF SPEECH WITH A VARIATIONAL AUTOENCODER

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding and controlling latent representations in deep generative models is a challenging yet important problem for analyzing, transforming and generating various types of data. In speech processing, inspiring from the anatomical mechanisms of phonation, the source-filter model considers that speech signals are produced from a few independent and physically meaningful continuous latent factors, among which the fundamental frequency and the formants are of primary importance. In this work, we show that the source-filter model of speech production naturally arises in the latent space of a variational autoencoder (VAE) trained in an unsupervised fashion on a dataset of natural speech signals. Using an artificial speech synthesizer, we experimentally demonstrate that the fundamental frequency and formant frequencies are encoded in orthogonal subspaces of the VAE latent space. We propose a weakly-supervised method to accurately and independently control the source-filter latent factors of speech variation within the learned latent subspaces. Without requiring additional information such as text nor human-labeled data, it leads to a deep generative model of speech spectrograms which is conditioned on the fundamental frequency and formant frequencies.

1 INTRODUCTION

High-dimensional data such as natural images or speech signals exhibit some form of regularity which prevents their dimensions from varying freely. This suggests that there exists a latent representation of smaller dimension from which the high-dimensional observed data were generated. Discovering the hidden properties of complex data is the goal of representation learning, and deep latent-variable generative models have emerged as promising unsupervised approaches (Goodfellow et al., 2014; Kingma & Welling, 2014; Rezende et al., 2014; Chen et al., 2016; Higgins et al., 2017; Kim & Mnih, 2018; Chen et al., 2018).

The variational autoencoder (VAE) (Kingma & Welling, 2014; Rezende et al., 2014), which is equipped with both a generative and inference model, can be used not only for data generation but also for analysis and transformation. As an explicit model of a probability density function (pdf), the VAE can also be used as a learned prior for solving inverse problems such as compressed sensing (Bora et al., 2017), speech enhancement (Bando et al., 2018; Leglaive et al., 2018), or source separation (Jayaram & Thickstun, 2020). Making sense of the latent representation learned by a VAE and controlling the underlying continuous factors of variation in the data are important challenges to build more expressive and interpretable generative models and probabilistic priors.

Previous works on representation learning with deep generative models, in particular VAEs, have mostly focused on images (Higgins et al., 2017; Kim & Mnih, 2018; Chen et al., 2018; Locatello et al., 2019; 2020). Yet, it is not always easy to define the ground-truth latent factors of variation involved in the generation of natural images. For speech data, the latent factors of variation can be directly related to the anatomical mechanisms of speech production. A key concept for characterizing the structure of speech signals is deduced from the source-filter model proposed by Fant (1970). This model, described in more detail in Section 2.2, implies that a speech signal is mainly characterized by a few continuous latent factors of variation corresponding to the vibration of the vocal folds

(i.e., the source), which defines the fundamental frequency, and the vocal tract shape (i.e., the filter), which defines the formants.

The source-filter model is at the core of various fundamental speech processing techniques such as cepstral representations and linear predictive coding (LPC) (Rabiner & Schafer, 2010). LPC was recently combined with recurrent neural networks for efficient speech synthesis in LPCnet (Valin & Skoglund, 2019). This illustrates the interest of combining deep learning techniques with more traditional signal processing models and algorithms. In this sense, understanding VAEs from the perspective of the source-filter model of speech production can be beneficial for various applications in speech analysis, transformation, and synthesis. We also believe that speech data, as naturally characterized by a few physically meaningful latent factors of variation, are an interesting material for investigating the representation learning capabilities of VAEs, in a complementary manner to VAE studies dealing with image data.

In this work, we first train a VAE on a dataset of natural speech signals. Then, based only on a few examples of artificially-generated labeled speech data, we propose a method to analyze and control the fundamental frequency and the formant frequencies in the latent representation of the previously trained VAE. Our contributions are the following: (i) We experimentally demonstrate that the fundamental frequency and the frequency of the first three formants are encoded in orthogonal subspaces of the VAE latent space. This shows that a vanilla VAE trained in an unsupervised fashion is able to learn a representation that is compliant with the source-filter model of speech production. (ii) We propose a very weakly supervised method to precisely and independently control the source-filter continuous latent factors of speech variation within the learned subspaces. We put in evidence the orthogonality of these subspaces, which allows us to perform speech transformations in a disentangled manner (i.e., modifying one of the factors does not affect the others). (iii) Without requiring additional information such as text nor human-labeled data, we obtain a deep generative model that can generate speech spectrograms conditionally on the fundamental frequency and formant frequencies. To the best of our knowledge, this is the first study showing that the source-filter model of speech production naturally arises in the latent space of a VAE, and the first method for generating speech data with a VAE from fundamental frequency and formant frequency trajectories.

2 BACKGROUND

2.1 VARIATIONAL AUTOENCODER

Generative modeling consists in learning a probabilistic model of an observable variable $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^D$. Let $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_N \in \mathcal{X}\}$ be a dataset of $N = \#\mathcal{D}$ independent and identically distributed (i.i.d.) observations of $\mathbf{x}$. The empirical distribution of $\mathbf{x}$ is defined by $\hat{p}(\mathbf{x}) = \frac{1}{N}\sum_{\mathbf{x}_n \in \mathcal{D}} \delta(\mathbf{x} - \mathbf{x}_n)$, where δ is the Dirac delta function, which is null everywhere except in 0 where it takes the value 1.

The variational autoencoder (VAE) (Kingma & Welling, 2014; Rezende et al., 2014) attempts to approximate $\hat{p}(\mathbf{x})$ with a pdf $p_\theta(\mathbf{x})$ parametrized by θ. High-dimensional data such as natural images or speech signals exhibit some form of regularity which prevents the D dimensions of $\mathbf{x}$ from varying freely. We can thus assume that there exists a latent variable $\mathbf{z} \in \mathbb{R}^L$, with $L \ll D$, from which the observed data were generated. Accordingly, the model distribution in the VAE is defined by marginalizing the joint distribution of the latent and observed data, $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$.

In this work, the observed data vector $\mathbf{x} \in \mathbb{R}_+^D$ denotes the power spectrum of a short frame of speech signal (i.e., a column of the short-time Fourier transform power spectrogram). Its entries are non negative and its dimension D equals the number of frequency bins. We use the Itakura-Saito VAE (IS-VAE) (Bando et al., 2018; Leglaive et al., 2018; Girin et al., 2019) defined by

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}), \qquad p_\theta(\mathbf{x}|\mathbf{z}) = \prod_{d=1}^{D} \mathrm{Exp}\Big([\mathbf{x}]_d; [\mathbf{v}_\theta(\mathbf{z})]_d^{-1}\Big), \tag{1}$$

where $\mathcal{N}$ and Exp denote the densities of the multivariate Gaussian and univariate exponential distributions, respectively, and $[\mathbf{v}]_d$ denotes the d-th entry of $\mathbf{v}$. The inverse scale parameters of $p_\theta(\mathbf{x}|\mathbf{z})$ are provided by a neural network called the decoder, parametrized by θ and taking $\mathbf{z}$ as input.

The marginal likelihood $p_\theta(\mathbf{x})$ and the posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ are intractable due to non linearities, so it is necessary to introduce an inference model $q_\phi(\mathbf{z}|\mathbf{x}) \approx p_\theta(\mathbf{z}|\mathbf{x})$ defined by

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}; \boldsymbol{\mu}_\phi(\mathbf{x}), \mathrm{diag}\{\mathbf{v}_\phi(\mathbf{x})\}\right), \tag{2}$$

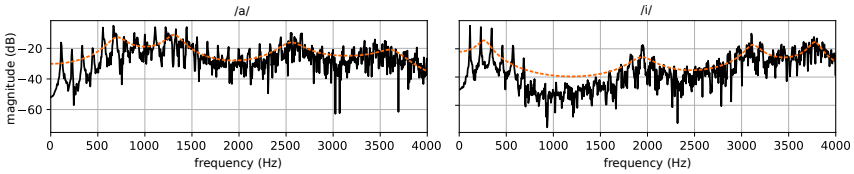


Figure 1: Power spectrum (solid black line) and spectral envelop (orange dashed line) for two vowels uttered by a male speaker.

where the mean and variance parameters are provided by a neural network called the encoder network, parametrized by ϕ and taking $\mathbf{x}$ as input. Then, the VAE training consists in maximizing a lower-bound of the intractable log-marginal likelihood, called the evidence lower-bound (ELBO) and defined by

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{\hat{p}(\mathbf{x})} \left[\ln p_\theta(\mathbf{x}) - D_{\mathrm{KL}} \left(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x}) \right) \right] \tag{3}$$

$$= \mathbb{E}_{\hat{p}(\mathbf{x})} \left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[p_\theta(\mathbf{x}|\mathbf{z}) \right] - D_{\mathrm{KL}} \left(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}) \right) \right]. \tag{4}$$

During training, the generative and inference model parameters θ and ϕ are jointly estimated by maximizing the ELBO in equation 4, using (variants of) stochastic gradient descent with the so-called reparameterization trick (Kingma & Welling, 2014; Rezende et al., 2014).

2.2 Source-filter model of speech production

The source-filter model of speech production (Fant, 1970) is at the basis of many speech processing systems. It considers that the production of speech results from the interaction of a source signal with a linear filter. In voiced speech, the source originates from the vibration of the vocal folds, which produces a quasi-periodic glottal sound wave whose fundamental frequency defines the *pitch*. In unvoiced speech, a noise source is produced by a turbulent airflow or an acoustic impulse. The source signal is modified by the vocal tract, which is assumed to act as a linear filter. The cavities of the vocal tract give rise to resonances, which are called the *formants* and are characterized by their frequency, amplitude and bandwidth. By moving the speech articulators such as the tongue, lips, and jaw, humans modify the shape of their vocal tract, which results in a change of the acoustic filter and the associated resonances. This is how the different elementary speech sounds called phonemes are produced to form syllables, words and sentences.

The power spectra and the spectral envelopes of two French vowels are displayed in Figure 1. The spectral envelopes show that the formant frequencies are different for the two vowels. In this example however, the harmonic structure of the spectra shows that the fundamental frequency is the same for the two vowels. Formant frequencies are important distinctive features of vowels. In a first approximation, they can be related to the opening of the mouth, the front/rear position of the tongue, and the rounding of the lips for the first, second, and third formant respectively. For voiced phonemes, humans are able to control the formants independently of the pitch (i.e., to change the filter independently of the source (Fant, 1970)). Globally, the formants are controlled independently of each other during speech production (MacDonald et al., 2011). The independence of the source and filter characteristics makes the speech signals an interesting material for representation learning methods, especially with deep generative latent-variable models.

In addition to the IS-VAE pre-trained speech spectrogram model, we also assume the availability of an artificial speech synthesizer allowing for an accurate and independent control of the fundamental frequency and formants. In this work, we use Soundgen (Anikin, 2019), a parametric synthesizer based on the source-filter model of speech production. For a given speech sound, the voiced component of the source signal is generated by a sum of sine waves, the noise component by a filtered white noise, and both components are then summed and passed through a linear filter simulating the effect of the human vocal tract. Importantly, this synthesizer allows us to easily generate artificial speech data labeled with the fundamental frequency and formant frequency values.

3 ANALYZING AND CONTROLLING SOURCE-FILTER FACTORS OF SPEECH VARIATION IN THE VAE

In this section, from a VAE trained on natural speech signals and a few artificially-generated labeled speech signals, we propose (i) a method to learn latent subspaces encoding source-filter factors of speech variation, (ii) a simple strategy to measure the disentanglement of the learned representation, and (iii) a weakly-supervised approach to control the continuous factors of variation in the learned subspaces and generate corresponding speech signals.

Let f_i denote the speech factor of variation (in Hz) corresponding to the fundamental frequency, for $i = 0$, and to the formant frequencies, for $i \in \{1, 2, ...\}$. Let $\mathcal{D}_i$ denote a dataset of artificially-generated speech vectors (more precisely short-term power spectra) synthesized by varying only f_i, all other factors $\{f_j, j \neq i\}$ being arbitrarily fixed. All examples in $\mathcal{D}_i$ are labeled with the factors of variation. It would be very difficult to build such a dataset from existing corpora of unlabeled natural speech. In contrast, it is an easy task using an artificial speech synthesizer such as Soundgen, which precisely takes the fundamental frequency and formant parameters as input and outputs waveforms from which we extract power spectra.

3.1 LEARNING LATENT SUBSPACES ENCODING SOURCE-FILTER FACTORS OF VARIATION

Let $\hat{p}^{(i)}(\mathbf{x})$ denote the empirical distribution associated with $\mathcal{D}_i$, defined similarly as $\hat{p}(\mathbf{x})$. We also introduce the following marginal distribution over the latent vectors:

$$\hat{q}_\phi^{(i)}(\mathbf{z}) = \int q_\phi(\mathbf{z}|\mathbf{x})\hat{p}^{(i)}(\mathbf{x})d\mathbf{x} = \frac{1}{\#\mathcal{D}_i} \sum_{\mathbf{x}_n \in \mathcal{D}_i} q_\phi(\mathbf{z}|\mathbf{x}_n). \tag{5}$$

This quantity is referred to as the aggregated posterior in the literature, where it is usually aggregated over the empirical distribution $\hat{p}(\mathbf{x})$ and is expected to match with the prior $p(\mathbf{z})$ (Makhzani et al., 2016; Chen et al., 2018; Dai & Wipf, 2018). In contrast, we aggregate over the "biased" data distribution $\hat{p}^{(i)}(\mathbf{x})$, where we know only one latent factor varies. This defines the explicit inductive bias (Locatello et al., 2019) that we exploit to learn the latent source-filter representation of speech in the VAE.

In the following, without loss of generality, we assume that, for each data vector in $\mathcal{D}_i$, the associated latent vector $\mathbf{z}$ has been centered by subtracting the mean vector

$$\boldsymbol{\mu}_\phi(\mathcal{D}_i) = \mathbb{E}_{\hat{q}_\phi^{(i)}(\mathbf{z})}[\mathbf{z}] = \frac{1}{\#\mathcal{D}_i} \sum_{\mathbf{x}_n \in \mathcal{D}_i} \boldsymbol{\mu}_\phi(\mathbf{x}_n). \tag{6}$$

Because only one factor varies in $\mathcal{D}_i$, we expect latent vectors drawn from the "biased" aggregated posterior in equation 5 to live on a lower-dimensional manifold embedded in the original latent space $\mathbb{R}^L$. We assume this manifold to be a subspace characterized by its semi-orthogonal basis matrix $\mathbf{U}_i \in \mathbb{R}^{L \times M_i}$, $1 \leq M_i < L$. This matrix is computed by solving the following optimization problem:

$$\min_{\mathbf{U} \in \mathbb{R}^{L \times M_i}} \mathbb{E}_{\hat{q}_\phi^{(i)}(\mathbf{z})} \left[\left\| \mathbf{z} - \mathbf{U}\mathbf{U}^\top \mathbf{z} \right\|_2^2 \right], \qquad s.t.\ \mathbf{U}^\top \mathbf{U} = \mathbf{I}. \tag{7}$$

The space spanned by the columns of $\mathbf{U}_i$ is a subspace of the original latent space $\mathbb{R}^L$ in which the latent vectors associated with the variation of the factor f_i in $\mathcal{D}_i$ are expected to live. In Appendix A, we show that, similarly to principal component analysis (PCA) (Pearson, 1901), the solution to the optimization problem 7 is given by the M_i eigenvectors corresponding to the M_i largest eigenvalues of

$$\mathbf{S}_\phi(\mathcal{D}_i) = \frac{1}{\#\mathcal{D}_i} \sum_{\mathbf{x}_n \in \mathcal{D}_i} \left[\boldsymbol{\mu}_\phi(\mathbf{x}_n)\boldsymbol{\mu}_\phi(\mathbf{x}_n)^\top + \mathrm{diag}\{\mathbf{v}_\phi(\mathbf{x}_n)\} \right] - \boldsymbol{\mu}_\phi(\mathcal{D}_i)\boldsymbol{\mu}_\phi(\mathcal{D}_i)^\top. \tag{8}$$

The dimension M_i of the subspace can be chosen such as to retain a certain percentage of the data variance in the latent space. Note that the only source of supervision used here is the knowledge that only the factor f_i varies in the dataset $\mathcal{D}_i$.

3.2 DISENTANGLEMENT ANALYSIS OF THE LATENT REPRESENTATION

As defined by Higgins et al. (2018), a representation is disentangled if it is possible to learn orthogonal latent subspaces associated with each factor of variation, whether they are single- or multi-dimensional. The approach presented in the previous subsection exactly follows this definition and offers a natural and straightforward way to objectively measure if the VAE, during its unsupervised training, managed to learn a disentangled representation of the factors of variation under consideration. First, by simply looking at the eigenvalues associated with the columns of $\mathbf{U}_i \in \mathbb{R}^{L \times M_i}$, we can measure the amount of variance that is retained by the projection $\mathbf{U}_i\mathbf{U}_i^\top$. If a small number of components M_i represents most of the variance, it indicates that only a few intrinsic dimensions of the latent space are dedicated to the factor of variation f_i and varying this factor can be done by affine transformations. Second, if for two different factors of variation f_i and f_j, with $i \neq j$, the columns of $\mathbf{U}_i$ are orthogonal to those of $\mathbf{U}_j$, this indicates that the two factors are encoded in orthogonal subspaces and therefore disentangled. It should however be verified experimentally that applying transformations by moving onto the subspace associated with f_i generalizes to values of $\{f_j, j \neq i\}$ different than the ones used in $\mathcal{D}_i$.

3.3 CONTROLLING THE FACTORS OF VARIATION FOR SPEECH TRANSFORMATION AND GENERATION

So far, for each factor f_i, we have defined a methodology to learn a latent subspace $\mathbf{U}_i \in \mathbb{R}^{L \times M_i}$ that encodes its variations in the dataset $\mathcal{D}_i$, containing a few examples of speech data generated by an artificial synthesizer. Making now use of the labels in $\mathcal{D}_i$, we want to learn a regression model $\mathbf{g}_{\eta_i} : \mathbb{R}_+ \mapsto \mathbb{R}^{M_i}$ from the factor f_i, whose value is denoted by $y \in \mathbb{R}_+$, to the data coordinates in the latent subspace defined by $\mathbf{U}_i$. The parameters η_i are defined as the solution of the following optimization problem:

$$\min_{\eta} \mathbb{E}_{\hat{q}_\phi^{(i)}(\mathbf{z},y)} \left[\left\| \mathbf{g}_\eta(y) - \mathbf{U}_i^\top \mathbf{z} \right\|_2^2 \right], \tag{9}$$

where $\hat{q}_\phi^{(i)}(\mathbf{z}, y) = \int q_\phi(\mathbf{z}|\mathbf{x})\hat{p}^{(i)}(\mathbf{x}, y)d\mathbf{x}$ and $\hat{p}^{(i)}(\mathbf{x}, y)$ is the empirical distribution associated with $\mathcal{D}_i$, considering now both the speech data vector $\mathbf{x}$ and the label y. It can be shown that the optimization problem 9 is equivalent to:

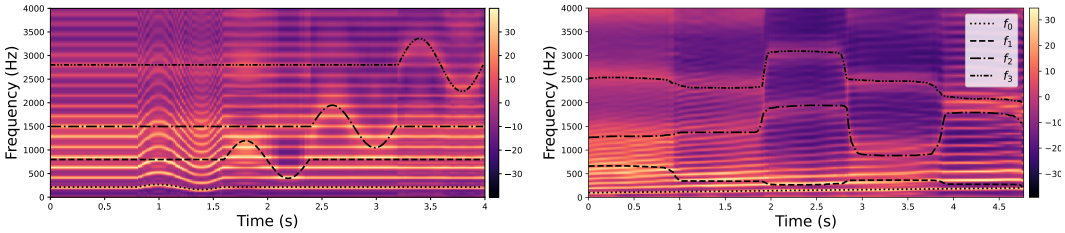
$$\min_{\eta} \left\{ \mathbb{E}_{\hat{p}^{(i)}(\mathbf{x},y)} \left[\| \, \mathbf{g}_\eta(y) - \mathbf{U}_i^\top \boldsymbol{\mu}_\phi(\mathbf{x}) \, \|_2^2 \right] = \frac{1}{\#\mathcal{D}_i} \sum_{(\mathbf{x}_n, y_n) \in \mathcal{D}_i} \left\| \mathbf{g}_\eta(y_n) - \mathbf{U}_i^\top \boldsymbol{\mu}_\phi(\mathbf{x}_n) \right\|_2^2 \right\}, \tag{10}$$

where we recall that $\mathbf{U}_i$ is the solution of equation 7. The dataset $\mathcal{D}_i$ is very small with only a few hundreds examples, and as it is synthetic and labels are not provided by human annotators, the problem can be considered very weakly supervised. For simplicity and because it revealed efficient for this task, $\mathbf{g}_{\eta_i}$ is chosen as a piece-wise linear regression model learned independently for each output coordinate $m \in \{1, ..., M_i\}$. This choice is supported by the fact that the semi-orthogonal matrix $\mathbf{U}_i$ decorrelates the data (Bengio et al., 2013). Solving the optimization problem 10 then basically consists in solving a linear system of equations (Jekel & Venter, 2019).

We can now transform a speech spectrogram by analyzing it with the VAE encoder, then linearly moving in the learned subspaces using the above regression model, and finally resynthesizing it with the VAE decoder. Given a source latent vector $\mathbf{z}$ and a target value y for the factor f_i, we apply the following linear transformation:

$$\tilde{\mathbf{z}} = \mathbf{z} - \mathbf{U}_i\mathbf{U}_i^\top \mathbf{z} + \mathbf{U}_i \mathbf{g}_{\eta_i}(y). \tag{11}$$

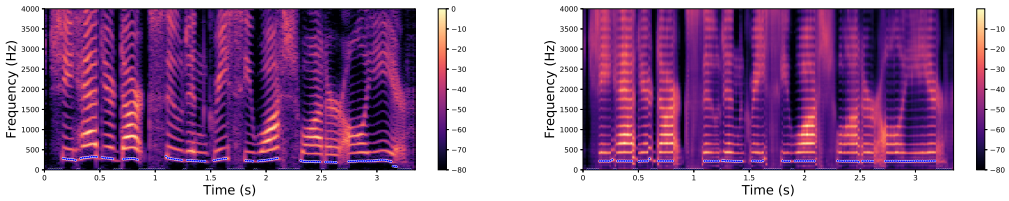
This transformation consists in (i) subtracting the projection of $\mathbf{z}$ onto the linear subspace associated with the factor of variation f_i; and (ii) adding the target component provided by the regression model $\mathbf{g}_{\eta_i}$ mapped from the learned subspace to the original latent space by the matrix $\mathbf{U}_i$. This process can be straightforwardly generalized to multiple factors, by subtracting and adding terms corresponding to each one of them. Contrary to the study of Hsu et al. (2017), the operation in equation 11 does not require the knowledge of the factor f_i associated with source vector $\mathbf{z}$, it only requires the value associated with the target vector $\tilde{\mathbf{z}}$.



(a) Fundamental frequency and formant transformations of the vowel /a/ uttered by a female speaker.

(b) Spectrogram generated from input trajectories of the pitch and formant frequencies.

Figure 2: Qualitative example of modified and generated spectrograms with the proposed method. The color bar indicates the power in dB.



(a)

(b) Spectrogram

Figure 3: Qualitative example

Equation 11 allows us to move in the latent subspace associated with the factor f_i. If this subspace is orthogonal to the latent subspaces associated with the other factors $\{f_j, j \neq i\}$, the latter should remain the same between $\mathbf{z}$ and $\tilde{\mathbf{z}}$, only f_i should be modified.

Finally, as the prior $p(\mathbf{z})$ and inference model $q_\phi(\mathbf{z}|\mathbf{x})$ are Gaussian (see equations 1 and 2), the transformation in equation 11 has the following probabilistic formulation:

$$p(\tilde{\mathbf{z}}; f_i = y) = \mathcal{N}\Big(\tilde{\mathbf{z}}; \mathbf{U}_i g_{\eta_i}(y), \mathbf{I} - \mathbf{U}_i \mathbf{U}_i^\top\Big) \tag{12}$$

$$q_\phi(\tilde{\mathbf{z}}|\mathbf{x}; f_i = y) = \mathcal{N}\Big(\tilde{\mathbf{z}}; \mathbf{U}_i g_{\eta_i}(y) + (\mathbf{I} - \mathbf{U}_i \mathbf{U}_i^\top)\boldsymbol{\mu}_\phi(\mathbf{x}), (\mathbf{I} - \mathbf{U}_i \mathbf{U}_i^\top)\,\mathrm{diag}\{\mathbf{v}_\phi(\mathbf{x})\}\Big). \tag{13}$$

The prior in equation 12 is now conditioned on the factor f_i and can be used to generate speech data given input trajectories of fundamental frequency and formant frequencies. As we assumed centered latent data, the mean vector $\boldsymbol{\mu}_\phi(\mathcal{D}_i)$ defined in equation 6 must be added to $\tilde{\mathbf{z}}$ before mapping $\tilde{\mathbf{z}}$ through the generative model $p_\theta(\mathbf{x}|\mathbf{z})$.

4 Experiments

This section presents qualitative and quantitative experimental results of the proposed method for analyzing and controlling a VAE-based speech model in terms of source-filter factors of variation. The VAE is trained on about 25 hours of unlabeled read speech data from the Wall Street Journal (WSJ0) dataset (Garofolo et al., 1993). For a given factor of variation, the corresponding latent subspace is learned (see Section 3.1) using short trajectories of speech power spectra (typically a few hundred vectors) generated with Soundgen (Anikin, 2019), all other factors being arbitrarily fixed. The regression models used to control the speech factors of variation in the latent space (see Section 3.3) are learned on the same trajectories, but using the labels that correspond to the input control parameters of Soundgen. More details on the experimental set-up can be found in Appendix B. Given a generated or transformed spectrogram at the output of the VAE, we use Waveglow (Prenger et al., 2019) to reconstruct the phase and then obtain a time-domain signal by inverse STFT.

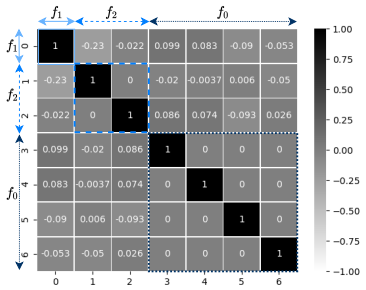


Figure 4: Correlation matrix of the learned latent subspaces basis vectors.

Factor	Method	NISQA (↑)	MOSnet (↑)	δf_0 (%, ↓)	δf_1 (%, ↓)	δf_2 (%, ↓)
f_0	TD-PSOLA	2.32 ± 0.55	2.60 ± 0.36	3.8 ± 2.5	6.3 ± 2.8	3.7 ± 0.9
	WORLD	2.49 ± 0.60	2.46 ± 0.46	4.5 ± 0.6	3.7 ± 1.8	2.3 ± 0.7
	VAE baseline	1.94 ± 0.43	2.44 ± 0.28	6.2 ± 2.8	10.4 ± 2.4	6.2 ± 0.9
	Proposed	2.08 ± 0.48	2.55 ± 0.40	0.8 ± 0.2	7.2 ± 1.3	3.6 ± 1.2
f_1	VAE baseline	1.84 ± 0.50	2.40 ± 0.10	11.3 ± 4.2	15.1 ± 3.5	6.0 ± 1.2
	Proposed	1.85 ± 0.40	2.45 ± 0.12	6.0 ± 1.6	8.4 ± 3.2	5.7 ± 0.4
f_2	VAE baseline	2.01 ± 0.40	2.55 ± 0.08	19.5 ± 3.2	10.7 ± 0.5	10.9 ± 1.9
	Proposed	2.03 ± 0.43	2.58 ± 0.41	8.5 ± 1.1	8.7 ± 1.1	6.2 ± 1.5

Table 1: Performance (mean and std) for the pitch (f_0) and formant (f_1, f_2) transformations.

4.1 Qualitative results

In Figure 2a, we illustrate the ability of the proposed method to modify the fundamental frequency and formant frequencies in an accurate and independent manner. The spectrogram contains five segments of equal length. The first segment corresponds to the original spectrogram of the steady vowel /a/ uttered by a female speaker. In the following segments, using equation 11, we vary each individual factor f_i, for $i = 0$ to 3, as indicated by the black lines in the figure. Variations of f_0 modify the harmonic structure of the signal,[2] whereas variations of f_i, $i \in \{1, 2, 3\}$, modify the resonance frequencies (i.e. the formants) as indicated by the color map while keeping the fundamental frequency unaltered. It can be seen that modifying one factor does not affect the others.

Figure 2b represents a spectrogram that is generated by first sampling $\mathbf{z}$ from the $p(\mathbf{z})$ in equation 1, and then applying equation 11 successively for each factor f_i, $i \in \{1, 2, 3\}$, according to the input trajectories represented by the lines in the figure. We can observe that the properties of the generated speech spectrogram match well with the input trajectories.

Other qualitative results can be found in Appendix D and at `https://sites.google.com/view/sturnus`. It includes audio examples and other representations of speech spectra that make formant modifications easier to observe. In particular, we illustrate the ability of the proposed model to suppress the pitch, that is to transform voiced speech into whispered speech. This is achieved by simply removing to $\mathbf{z}$ its projection onto the latent subspace corresponding to the pitch, i.e. by considering only the two first operations in the right-hand side of equation 11. This is remarkable considering that the VAE was not trained on whispered speech signals, and it further confirms that the proposed method dissociates the filter and the source contributions in the VAE latent space.

4.2 Quantitative results

Due to space limitations, we provide in this section quantitative results related to the pitch (f_0) and the first two formant frequencies (f_1 and f_2) only. Additional results can be found in Appendix D.

4.2.1 Orthogonality of the latent subspaces

Following the discussion in Section 3.2 regarding the disentanglement of the learned representation, we compute the dot product between the all unit vectors in the matrices $\mathbf{U}_i \in \mathbb{R}^{L \times M_i}$, $i \in \{0, 1, 2\}$. Note that for a given factor, the latent subspace dimension M_i is chosen such that 80% of the data variance is retained, when solving the optimization problem 7. It can be seen in Figure 4 that the resulting correlation matrix is mainly diagonal, confirming the orthogonality of the learned subspaces and thus the disentanglement of the learned source-filter representation of speech.

4.2.2 Pitch and formant transformations

In this experiment, we evaluate the performance of the proposed method regarding the modification of the pitch and formant frequencies in speech signals (see Section 3.3).

Dataset For evaluation, we use a corpus of English vowels (Hillenbrand et al., 1995), including 50 male and 50 female speakers pronouncing each 12 vowels.

Metrics For the modification of a given factor f_i, we measure the performance regarding three aspects: First, in terms of *precision*, by computing δf_i defined as the relative absolute error (in percent, as used e.g. in (Alku et al., 2013), the lower the better) between the target value y for the factor f_i (see equation 11) and its estimation $\hat{y}$ computed from the modified speech signal: $100\% \times |\hat{y} - y|/y$. Second, in terms of *disentanglement*, by computing again the relative absolute error but this time between the estimated values of f_j, for all $j \neq i$, before and after modification of the factor f_i in the speech signal. We denote by δf_j, $j \neq i$, these measures which aim to assess if the other factors of variation remained unchanged after modifying f_i. Third, in terms of *naturalness* of the transformed speech, using the objective measures provided by MOSnet (Lo et al., 2019) and NISQA Mittag & Möller (2020). These metrics (in between 1 and 5, higher is better) were developed in the context of speech transformation algorithms and they were shown to highly correlate with subjective mean opinion score (MOS) (i.e. human ratings). Given a speech signal, we use CREPE (Kim et al., 2018) to estimate the pitch (f_0) and Parselmouth (Jadoul et al., 2018) which is based on PRAAT (Boersma & Weenink, 2021) to estimate the formant frequencies ($f_i, i \neq 0$).

Reference methods We compare the proposed approach with several reference methods: (i) Time-domain pitch-synchronous overlap-and-add (TD-PSOLA) (Moulines & Charpentier, 1990) which performs pitch modification exploiting a decomposition of the signal into pitch-synchronized overlapping frames. (ii) WORLD (Morise et al., 2016), a vocoder which transforms pitch by exploiting on a decomposition of the speech signal into three components characterizing the fundamental frequency, the aperiodicity and the spectral envelope. (iii) The method proposed in Hsu et al. (2017) (here referred to as "VAE baseline") which consists in directly performing arithmetic operations in the latent space of the VAE, and not latent subspaces. Unlike the proposed approach, this method requires predefined latent representations in order to carry out the arithmetic transformation: $\mathbf{z}_{\text{trgt}} = \mathbf{z}_{\text{src}} - \boldsymbol{\mu}_{\text{src}} + \boldsymbol{\mu}_{\text{trgt}}$ where $\boldsymbol{\mu}_{\text{src/trgt}}$ is the so-called latent attribute representation associated with the source/target values of the factor to be modified. Unlike the proposed method, this VAE baseline requires analyzing the input speech signal to be modified in order to choose the proper source latent attribute representation. To ensure fair comparison, we compute these predefined latent attribute representations using the same artificially-generated labeled speech data as in the proposed method. TD-PSOLA also requires the pitch contour of the input speech signal.

Discussion Experimental results are provided in Table 1. Compared with the VAE baseline, the proposed method obtains better performance in terms of precision, disentanglement, and naturalness. It confirms the effectiveness of performing the transformations in latent subspaces and not directly in the latent space, and the advantage of using regression models instead of dictionaries of predefined latent attribute representations. Regarding transformations of the pitch (f_0), WORLD obtains the best performance in terms of disentanglement, which is due to the fact that source and filter contributions are decoupled in the architecture of the vocoder. In terms of naturalness, WORLD or TD-PSOLA obtain the best performance depending on the metric. It may be explained by the fact that these method operate directly in the time domain, therefore they do not suffer from phase reconstruction artifacts, unlike the proposed and VAE baseline methods. The naturalness is indeed greatly affected by the phase reconstruction algorithm, even from an original speech spectrogram (i.e. without transformations). Phase reconstruction is still an open problem in speech processing.

5 Related work

Modifying characteristics of speech signals such as the pitch or timbre is a highly covered research topic where deep generative modeling has recently emerged as a promising approach. Voice conversion consists for instance in modifying speaker characteristics while keeping the linguistic information unchanged. Most voice conversion methods are text-to-speech systems (Wang et al., 2018; Zhang et al., 2019; Kulkarni et al., 2020; Li et al., 2021). Using text as an input greatly helps for transferring non-verbal vocal attributes from one speech signal to another, but it requires datasets labeled with text.

Recently, Qian et al. (2020) proposed a method for decomposing a speech signal into pitch, timbre and rhythm latent representations, without resorting to textual information. The disentanglement of the representations is enforced by the design of the model which comprises three independent encoder networks taking as input different representations of the speech signal. This method is also speaker dependent, as it generates speech signals conditionally on the speaker identity. It contrasts with the proposed method which is speaker independent and where the source-filter representation of speech naturally emerges in the latent space of a simple VAE, without requiring a specific encoder for each component.

Variants of the VAE (Kingma & Welling, 2014; Rezende et al., 2014) have recently seen considerable progress in disentangled representation learning (Kim & Mnih, 2018; Higgins et al., 2017; Chen et al., 2018). From experimental analyses mostly based on image data (often synthetic), these methods suggest that a vanilla VAE cannot learn a disentangled representation. In the present study, we experimentally show that learning a disentangled source-filter representation of speech using a simple VAE is possible, complying with the definition of disentanglement proposed in (Higgins et al., 2018). Locatello et al. (2019; 2020) have recently shown both theoretically and from a large-scale experimental study that the unsupervised learning of disentangled representations is fundamentally impossible without inductive biases on both the models and the data. In this work, we precisely employ a few examples of artificially-generated labeled speech data in order to disentangle the latent representation of a simple VAE, in terms of source-filter factors of speech variation.

The method closest to our work is probably (Hsu et al., 2017). It proposes to modify the speaker identity and the phonemic content of speech signals using arithmetic operations in the latent space of a VAE. This is the baseline VAE approach we presented and compared with in Section 4. It requires to know upstream predefined latent representations associated with the source/target speech attributes (e.g. male/female speaker identity) to be modified. This differs from the proposed method which relies on projection onto the latent subspace associated with a given attribute, and only requires the target value for this attribute. Moreover, (Hsu et al., 2017) does not address the control of continuous factors of speech variation in the VAE latent space, contrary to the present work.

Several methods have been recently proposed to control continuous factors of variation in deep generative models (Jahanian et al., 2019; Plumerault et al., 2020; Goetschalckx et al., 2019; Härkönen et al., 2020), focusing essentially on GANs. They consist in identifying and then moving onto semantically meaningful directions in the latent space of the model. The present work is inspired by (Plumerault et al., 2020), which assumes that a factor of variation can be predicted from the projection of the latent vector along a specific axis, that is learned from artificially generated trajectories. The proposed method is however more generic, thanks to the learning of latent subspaces associated to the latent factors, and to the introduction of a general formalism based on the use of "biased" aggregated posteriors. Moreover, these previous works on controlling deep generative models only allow for moving "blindly" onto semantically meaningful directions in the latent space. In this work, we are able to generate data conditioned on a precise desired value for a given factor of variation. This is achieved by exploiting a few automatically-labeled data to learn a regression model from the value of the factor to the coordinates in the latent subspace it is associated with. Finally, all the above-mentioned previous works focused on image data, and to the best of our knowledge, this is the first approach to control source-filter factors of speech variation in deep generative models.

6 Conclusion

Understanding the latent space means controlling the generative model. We proposed a very weak supervised method to control cautiously attributes of speech (ff, formant) via a variational autoencoder trained on real speech spectrograms. From the automatically labeled trajectories associated with each variation factor, the method relies on linear transformations and projections between the VAE latent space and the factor subspace. We succeeded in generating or controlling accurately the ff and formants without any human intervention conditioned only by the value of the target factor, this thanks to the regression that allows us to move perfectly in the associated subspace. We have shown experimentally that these variation factors are strongly decorrelated in the latent space. This is reinforced by the quantitative results where the source and filter are independently controlled. For future work, we plan to study other higher-level factors such as more or less complex emotions by

controlling the latent space of the VAE in a hierarchical manner, by including other modalities such as image and text and by modeling the temporal aspect of the data.

REFERENCES

Paavo Alku, Jouni Pohjalainen, Martti Vainio, Anne-Maria Laukkanen, and Brad H Story. Formant frequency estimation of high-pitched vowels using weighted linear prediction. *The Journal of the Acoustical Society of America*, 134(2):1295–1313, 2013.

Andrey Anikin. Soundgen: An open-source tool for synthesizing nonverbal vocalizations. *Behavior Research Methods*, 51(2):778–792, 2019.

Yoshiaki Bando, Masato Mimura, Katsutoshi Itoyama, Kazuyoshi Yoshii, and Tatsuya Kawahara. Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 716–720, 2018.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, 2006. ISBN 9780387310732.

Paul Boersma and David Weenink. Praat: doing phonetics by computer [Computer program]. Version 6.1.38, retrieved 2 January 2021 `http://www.praat.org/`, 2021.

Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *International Conference on Machine Learning (ICML)*, pp. 537–546, 2017.

Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Bin Dai and David Wipf. Diagnosing and enhancing vae models. In *International Conference on Learning Representations (ICLR)*, 2018.

Gunnar Fant. *Acoustic theory of speech production*. Number 2. Walter de Gruyter, 1970.

John S. Garofolo, David Graff, Doug Paul, and David Pallett. CSR-I (WSJ0) Sennheiser LDC93S6B. `https://catalog.ldc.upenn.edu/LDC93S6B`, 1993. Philadelphia: Linguistic Data Consortium.

Laurent Girin, Fanny Roche, Thomas Hueber, and Simon Leglaive. Notes on the use of variational autoencoders for speech and audio spectrogram modeling. In *International Conference on Digital Audio Effects (DAFx)*, pp. 1–8, 2019.

Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5744–5753, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

James Hillenbrand, Laura A Getty, Michael J Clark, and Kimberlee Wheeler. Acoustic characteristics of american english vowels. *The Journal of the Acoustical society of America*, 97(5): 3099–3111, 1995.

Wei-Ning Hsu, Yu Zhang, and James Glass. Learning latent representations for speech generation and transformation. In *Interspeech*, pp. 1273–1277, 2017.

Yannick Jadoul, Bill Thompson, and Bart de Boer. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15, 2018.

Ali Jahanian, Lucy Chai, and Phillip Isola. On the "steerability" of generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Vivek Jayaram and John Thickstun. Source separation with deep generative priors. In *International Conference on Machine Learning (ICML)*, pp. 4724–4735, 2020.

Charles F Jekel and Gerhard Venter. Pwlf: a python library for fitting 1D continuous piecewise linear functions. URL: `https://github.com/cjekel/piecewise_linear_fit_py/raw/master/paper/pwlf_Jekel_Venter_v2.pdf`, 2019.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning (ICML)*, pp. 2649–2658, 2018.

Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. Crepe: A convolutional representation for pitch estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 161–165, 2018.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Ajinkya Kulkarni, Vincent Colotte, and Denis Jouvet. Improving latent representation for end to end multispeaker expressive text to speech system. 2020.

Simon Leglaive, Laurent Girin, and Radu Horaud. A variance modeling framework based on variational autoencoders for speech enhancement. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2018.

Tao Li, Shan Yang, Liumeng Xue, and Lei Xie. Controllable emotion transfer for end-to-end speech synthesis. In *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 1–5, 2021.

Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. Mosnet: Deep learning based objective assessment for voice conversion. *arXiv preprint arXiv:1904.08352*, 2019.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning (ICML)*, pp. 4114–4124, 2019.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. A sober look at the unsupervised learning of disentangled representations and their evaluation. *Journal of Machine Learning Research*, 21:1–62, 2020.

Ewen N MacDonald, David W Purcell, and Kevin G Munhall. Probing the independence of formant control using altered auditory feedback. *The Journal of the Acoustical Society of America*, 129 (2):955–965, 2011.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. In *ICLR 2016 Workshop, International Conference on Learning Representations (ICLR)*, 2016.

Gabriel Mittag and Sebastian Möller. Deep learning based assessment of synthetic speech naturalness. In *Interspeech*, 2020.

Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884, 2016.

Eric Moulines and Francis Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, 9(5-6):453–467, 1990.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

Antoine Plumerault, Hervé Le Borgne, and Céline Hudelot. Controlling generative models with continuous factors of variations. In *International Conference on Learning Representations (ICLR)*, 2020.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3617–3621, 2019.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox. Unsupervised speech decomposition via triple information bottleneck. In *International Conference on Machine Learning (ICML)*, pp. 7836–7846, 2020.

Lawrence Rabiner and Ronald Schafer. *Theory and applications of digital speech processing*. Prentice Hall Press, 2010.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *International Conference on Machine Learning (ICML)*, 2014.

Jean-Marc Valin and Jan Skoglund. Lpcnet: Improving neural speech synthesis through linear prediction. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5891–5895, 2019.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning (ICML)*, pp. 5180–5189, 2018.

Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6945–6949, 2019.

A SOLUTION TO THE LATENT SUBSPACE LEARNING PROBLEM

In this Appendix, we show that the solution to the optimization problem 7 is given by the principal eigenvectors of $\mathbf{S}_\phi(\mathcal{D}_i)$ in equation 8.

Without loss of generality, we formulate the problem from a centered version of the latent data:

$$\mathbf{z} \leftarrow \mathbf{z} - \boldsymbol{\mu}_\phi(\mathcal{D}_i), \tag{14}$$

where $\boldsymbol{\mu}_\phi(\mathcal{D}_i)$ is defined in equation 6. This centering also affects the inference model originally defined in equation 2, as follows:

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}; \boldsymbol{\mu}_\phi(\mathbf{x}) - \boldsymbol{\mu}_\phi(\mathcal{D}_i), \mathrm{diag}\{\mathbf{v}_\phi(\mathbf{x})\}\right). \tag{15}$$

Using equation 5, the fact that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$, and equation 15, the cost function in the optimization problem 7 can be rewritten as follows:

$$\begin{aligned}
\mathbb{E}_{\hat{q}_\phi^{(i)}(\mathbf{z})}\left[\left\|\mathbf{z} - \mathbf{U}\mathbf{U}^\top\mathbf{z}\right\|_2^2\right] &= \frac{1}{\#\mathcal{D}_i}\sum_{\mathbf{x}_n\in\mathcal{D}_i}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_n)}\left[\left\|\mathbf{z} - \mathbf{U}\mathbf{U}^\top\mathbf{z}\right\|_2^2\right] \\
&= \mathrm{tr}\left\{(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\frac{1}{\#\mathcal{D}_i}\sum_{\mathbf{x}_n\in\mathcal{D}_i}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_n)}[\mathbf{z}\mathbf{z}^\top]\right\} \\
&= \mathrm{tr}\left\{(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\mathbf{S}_\phi(\mathcal{D}_i)\right\},
\end{aligned} \tag{16}$$

where $\mathbf{S}_\phi(\mathcal{D}_i)$ is defined in equation 8. From this last equality, we see that the optimization problem 7 is equivalent to

$$\max_{\mathbf{U}\in\mathbb{R}^{L\times M_i}} \ \mathrm{tr}\left\{\mathbf{U}^\top\mathbf{S}_\phi(\mathcal{D}_i)\mathbf{U}\right\}, \qquad s.t.\ \mathbf{U}^\top\mathbf{U} = \mathbf{I}. \tag{17}$$

Very similarly to PCA (Pearson, 1901), it can be shown (see e.g. (Bishop, 2006, Section 12.1)) that the solution is given by the M_i dominant eigenvectors of $\mathbf{S}_\phi(\mathcal{D}_i)$ (i.e. associated to the M largest eigenvalues).

B EXPERIMENTAL SETUP

VAE training For training the IS-VAE, we used the Wall Street Journal (WSJ0) dataset Garofolo et al. (1993) which represents 25 hours of speech recordings, including 52 female and 49 male speakers.

The architecture of the VAE is quite simple with two dense layers for the encoder and three dense layers for the decoder as shown in Figure 5. The "tanh" activation function is used at each layer, except for the output layers where the identity function is used. The latent space is 16-dimensional. The VAE is trained using the Adam optimizer with a learning rate equal to 0.001.

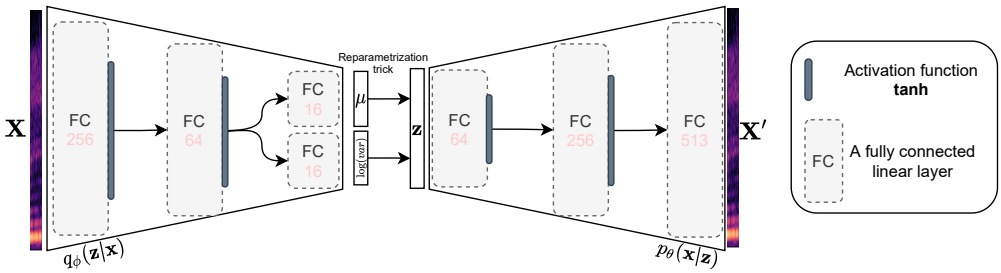


Figure 5: the architecture of the VAE

Analyzing and controlling source-filter factors of speech variation in the VAE For a given factor of variation, the corresponding latent subspace is learned (see Section 3.1) using trajectories of speech power spectra generated with Soundgen (Anikin, 2019), all other factors being arbitrarily fixed. For f_0, the trajectory contains 226 points evenly spaced in the range $[85, 310]$ Hz, f_1 and f_2 are fixed at 600 Hz and 1500 Hz, respectively. For f_1, the trajectory contains 401 points evenly spaced in the range $[200, 1000]$ Hz, f_0 and f_2 are fixed at 140 Hz and 1600 Hz, respectively. For f_2, the trajectory contains 401 points evenly spaced in the range $[800, 2800]$ Hz, f_0 and f_1 are fixed at 140 Hz and 500 Hz, respectively. The regression models used to control the speech factors of variation in the latent space (see Sections 3.3) are learned on the same trajectories, but using the labels that correspond to the input control parameters of Soundgen.

C WHAT DO THE VARIATION FACTORS LOOK LIKE IN LATENT SPACE?

As explained previously, the dimension of the latent space does not make it possible to visualise the distribution of the variation factors. But using the linear transformation $\mathbf{U}_i$, we succeed in reducing the dimension of the latent space while preserving the volume intact; in other words, given that the determinant of $\mathbf{U}_i$ is equal to 1, this means that $\mathbf{U}_i$ transforms the latent space without contracting or expanding it. The results of these visualizations are given in figure 6, where we see the three trajectories for the pitch, formant 1 and formant 2. The complexity of the latter is not the same. The trajectory of the pitch (multitude of spirals) has a more complex shape than formant 1 where the shape is almost affine. This complexity of trajectories requires a greater number of dimensions (directions) to represent them without losing information.

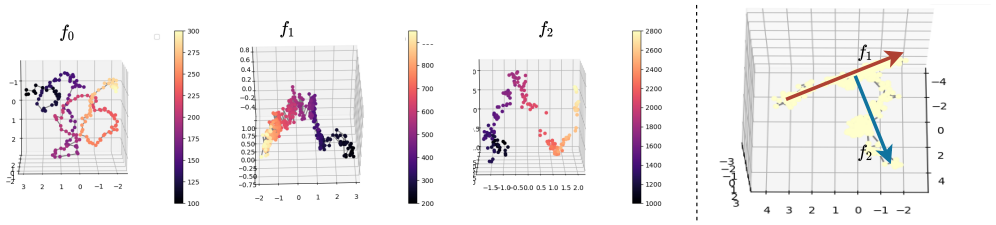
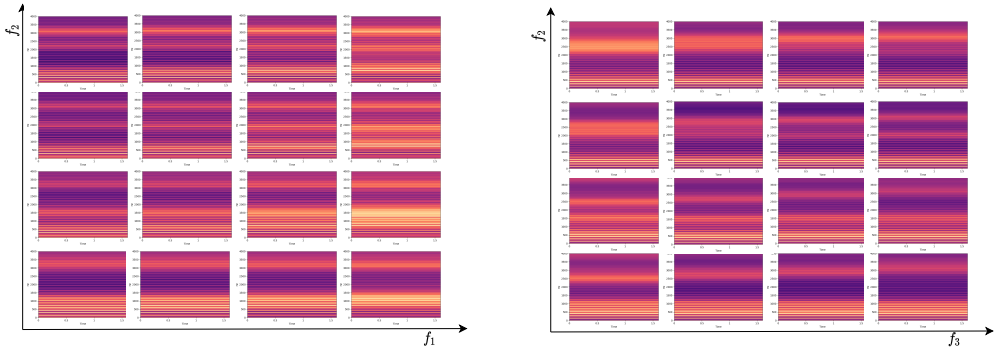


Figure 6: the figure shows the trajectory of the pitch, formant 1 and formant 2 respectively (from left to right) projected in the space of the pca.

D ADDITIONAL EXPERIMENTAL RESULTS



(a) f1 f2 (b) f2 f3

Figure 7: f2 f3

Factor	Method	NISQA (↑)	MOSnet (↑)	δf_0 (%, ↓)	δf_1 (%, ↓)	δf_2 (%, ↓)	δf_3 (%, ↓)
f_0	TD-PSOLA	2.32 ± 0.55	2.60 ± 0.36	3.8 ± 2.5	6.3 ± 2.8	3.7 ± 0.9	2.1 ± 0.5
	WORLD	2.49 ± 0.60	2.46 ± 0.46	4.5 ± 0.6	3.7 ± 1.8	2.3 ± 0.7	1.2 ± 0.2
	VAE baseline	1.94 ± 0.43	2.44 ± 0.28	6.21 ±2.8	10.4 ± 2.4	6.2± 0.9	4.5 ± 0.2
	Propo. (WSJ)	2.08 ± 0.48	2.55 ± 0.40	0.8 ± 0.2	7.2 ± 1.3	3.6 ± 1.2	3.8 ± 0.3
	Propo. (Fr)	1.93 ± 0.43	2.53 ± 0.38	1.2 ± 0.5	10 ± 4.2	8.3 ± 1.1	14 ± 0.2
	Propo. (TESS)	1.98 ± 0.5	2.53 ± 0.48	2.7 ± 2.3	9.3 ± 3.5	9 ± 0.8	7 ± 0.2
	Propo. (LJ)	1.96 ± 0.4	2.45 ± 0.52	1.2 ± 0.6	9.3 ± 1.2	5.6 ± 0.6	4.6 ± 0.1
f_1	VAE baseline	1.84± 0.5	2.40 ± 0.10	11.3 ± 4.2	15.1 ± 3.5	6.0 ± 1.2	4.2 ± 0.4
	Proposed	1.85 ± 0.4	2.45 ± 0.12	6.0 ± 1.6	8.4 ± 3.2	5.7 ± 0.4	4.4 ± 0.3
f_2	VAE baseline	2.01± 0.4	2.55 ± 0.08	19.5 ± 3.2	10.7 ± 0.5	10.9 ± 1.9	5.8 ± 0.6
	Proposed	2.03 ± 0.43	2.58 ± 0.41	8.5 ± 1.1	8.7 ± 1.1	6.2 ± 1.5	5.8 ± 0.2

Table 2: Performance (mean and std) for the pitch (f0) and formant (f1,f2) transformations.

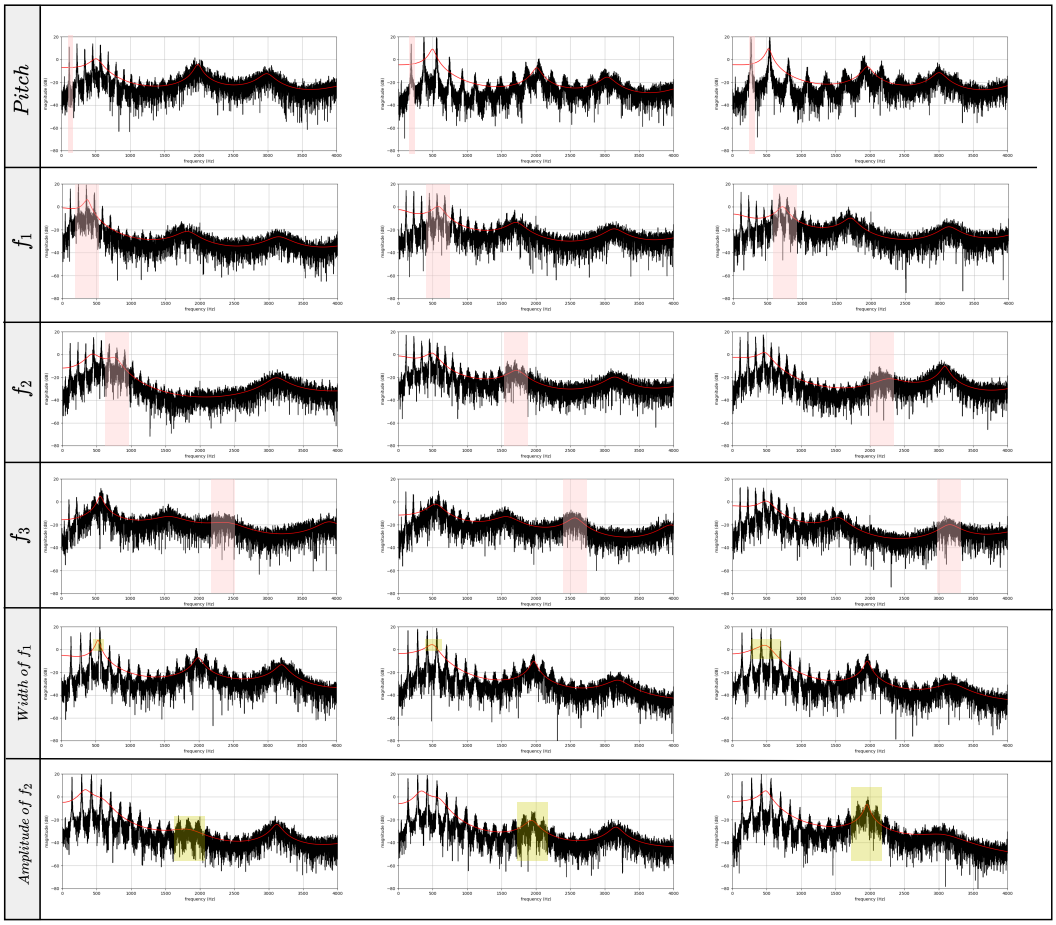


Figure 8: Caption

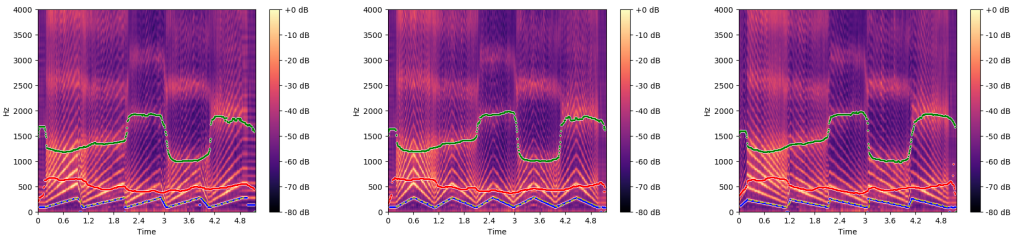


Figure 9: Pitch transformations

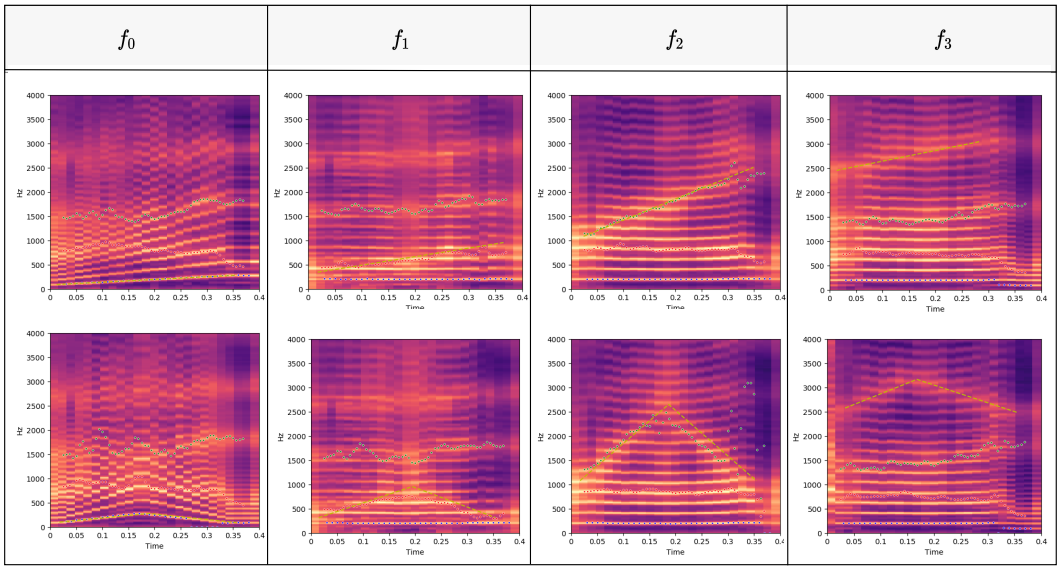


Figure 10: Pitch and formant transformations of the spectrum of the vowel ‘ah’ spoken by a woman