Les auto-encodeurs variationnels dynamiques et leur application à la modélisation de spectrogrammes de parole

Laurent Girin, Xiaoyu Bie, Simon Leglaive, Thomas Hueber, Xavier Alameda-Pineda

Les auto-encodeurs variationnels dynamiques et leur application à la modélisation de spectrogrammes de parole

Laurent Girin[1] Xiaoyu Bie[2] Simon Leglaive[3]
Thomas Hueber[1] Xavier Alameda-Pineda[2]
(1) Univ. Grenoble Alpes, CNRS, Grenoble-INP, GIPSA-lab, 38000 Grenoble, France
(2) Inria, Univ. Grenoble Alpes, CNRS, LJK, 38000 Grenoble, France
(3) CentraleSupélec, IETR, 35576 Cesson-Sévigné, France
{laurent.girin,thomas.hueber}@grenoble-inp.fr,
{xiaoyu.bie,xavier.alameda-pineda}@inria.fr,
simon.leglaive@centralesupelec.fr

RÉSUMÉ

L'auto-encodeur variationnel (AEV) est un modèle génératif profond permettant d'apprendre de façon auto-supervisé des représentations latentes compactes, à partir de données complexes de grande dimension. Dans le modèle AEV original, les vecteurs de données d'entrée sont traités indépendamment. Ces dernières années, plusieurs travaux ont proposé différentes extensions de l'AEV afin de traiter des données séquentielles (notamment temporelles). Ces modèles utilisent classiquement des réseaux de neurones récurrents pour tenir compte non seulement des dépendances entre les vecteurs d'une séquence d'entrée, mais également celles entre les représentations latentes correspondantes. Nous avons récemment effectué une revue complète de ces modèles et les avons unifiés en une classe générale appelée auto-encodeurs variationnels dynamiques (AEVDs). Dans le présent article, nous présentons cette classe de modèles et illustrons leur fort potentiel pour la modélisation des (spectrogrammes de) signaux de parole avec des expériences en analyse-resynthèse.

ABSTRACT

Dynamical variational autoencoders and their application to speech spectrogram modeling.

The Variational Autoencoder (VAE) is a powerful deep generative model that is now extensively used to represent high-dimensional complex data via a low-dimensional latent space learned in an unsupervised manner. In the original VAE model, input data vectors are processed independently. In recent years, a series of papers have presented different extensions of the VAE to process sequential data, that not only model the latent space, but also model the temporal dependencies within a sequence of data vectors and corresponding latent vectors, relying on recurrent neural networks. We recently performed a comprehensive review of those models and unified them into a general class called Dynamical Variational Autoencoders (DVAEs). In the present paper, we present this class of models and illustrate their high potential for modeling (spectrograms of) speech signals with speech analysis-resynthesis experiments.

MOTS-CLÉS : Modélisation des signaux de parole, auto-encodeurs variationnels dynamiques, spectrogrammes de parole, analyse-resynthèse de la parole.

KEYWORDS: Speech signals modeling, dynamical variational autoencoders, speech spectrograms, speech analysis-resynthesis.

1 Introduction

L'auto-encodeur variationnel (AEV) introduit dans (Kingma & Welling, 2014; Rezende *et al.*, 2014) est un modèle génératif profond maintenant largement utilisé pour représenter des données de grande dimension via un espace latent de faible dimension appris de manière non supervisée. Il a été utilisé pour la modélisation de la parole dans, par exemple, (Blaauw & Bonada, 2016; Hsu *et al.*, 2017a; Bando *et al.*, 2018; Leglaive *et al.*, 2018; Akuzawa *et al.*, 2018; Leglaive *et al.*, 2019). L'AEV original ne comprend pas de modélisation temporelle : chaque vecteur de données est traité indépendamment des autres vecteurs de données (et le vecteur latent correspondant est également traité indépendamment des autres vecteurs latents).

Ces dernières années, une série d'articles a présenté différentes extensions séquentielles de l'AEV qui non seulement modélisent l'espace latent, mais modélisent également les dépendances temporelles au sein d'une séquence de vecteurs de données et de vecteurs latents correspondants, en s'appuyant sur des réseaux de neurones récurrents (RNR) (Bayer & Osendorfer, 2014; Fabius & van Amersfoort, 2014; Krishnan *et al.*, 2015; Chung *et al.*, 2015; Fraccaro *et al.*, 2016, 2017; Hsu *et al.*, 2017b; Li & Mandt, 2018; Leglaive *et al.*, 2020). En pratique, ces différents modèles varient au niveau des dépendances entre les variables observées et latentes, dans la définition et la paramétrisation de leurs distributions, et dans la façon d'implémenter ces éléments avec des RNRs. En revanche, la phase d'apprentissage est assez similaire entre les modèles puisqu'elle est systématiquement basée sur la méthodologie variationnelle employée pour l'AEV : enchaînement des modèles d'inférence et génératif (l'encodeur et le décodeur) et maximisation d'une borne inférieure de la vraisemblance des données sur un ensemble de données d'apprentissage.

Dans (Girin *et al.*, 2021), nous avons effectué une revue de littérature et une analyse approfondie de ces modèles. Nous avons introduit une classe générale de modèles appelés auto-encodeurs variationnels dynamiques (AEVD) qui englobe et unifie les modèles cités ci-dessus. Dans (Bie *et al.*, 2021), nous avons effectué une évaluation comparative de six AEVDs différents sur la tâche d'analyse-resynthèse de spectrogrammes de parole pour une base de données en Anglais.

Dans cet article, nous présentons cette classe de modèle ainsi que des illustrations de leur utilisation pour l'analyse-synthèse de (spectrogrammes de) signaux de parole en Français. Notre motivation est ici de promouvoir cette classe de modèles auprès de la communauté scientifique du traitement de la parole. En effet, jusqu'à présent, les AEVDs n'ont été utilisés que dans très peu d'études en parole, voir par exemple les travaux pionniers en débruitage (Leglaive *et al.*, 2020). Pourtant, nous sommes convaincus que leur puissance de modélisation pourraient être avantageusement exploitée dans de nombreuses applications telles que la synthèse et la transformation de la parole.

2 Auto-encodeurs variationels dynamiques

Le modèle AEV original introduit dans (Kingma & Welling, 2014; Rezende *et al.*, 2014) est défini par :

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z}), \tag{1}$$

où $p(\mathbf{z})$, la distribution a priori de la variable latente $\mathbf{z}$, est une distribution normale multivariée, $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ est la *fonction de vraisemblance* (conditionnelle) de la variable observée $\mathbf{x}$, et la dimension L de $\mathbf{z}$ est (beaucoup) plus faible que la dimension F de $\mathbf{x}$. Les paramètres de $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ sont fournis par

un réseau neuronal profond (RNP), appelé *réseau décodeur*, qui prend $\mathbf{z}$ comme entrée. $\boldsymbol{\theta}$ représente les paramètres de ce réseau décodeur (par exemple, les poids et les biais d'un perceptron multicouche).

Étant donné que la relation entre $\mathbf{z}$ et $\mathbf{x}$ est non linéaire, la distribution a posteriori $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ n'est pas calculable analytiquement. Elle est donc approximée par une distribution variationnelle paramétrique $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ appelée modèle d'inférence, dont les paramètres sont fournis par un autre RNP (appelé l'*encodeur*, avec des paramètres $\boldsymbol{\phi}$ et une entrée $\mathbf{x}$). Un choix habituel consiste à définir $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ comme une distribution gaussienne avec une matrice de covariance diagonale.

Les paramètres $\{\boldsymbol{\theta},\boldsymbol{\phi}\}$ sont ensuite estimés conjointement en maximisant une borne inférieure de la fonction de log-vraisemblance des données, appelée borne inférieure variationnelle (BIV), donnée par (pour un seul vecteur de données) :

$$\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta},\mathbf{x}) = E_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\big[\ln p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\big] - D_{\mathrm{KL}}\big[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\big], \tag{2}$$

et évaluée sur un jeu de données d'apprentissage (D_{KL} désigne la divergence de Kullback-Leibler). La maximisation de la BIV se fait en combinant descente de gradient (en général stochastique) et techniques d'échantillonnage. Un tel processus d'optimisation est désormais considéré comme une routine dans les boîtes à outils d'apprentissage profond telles que Keras et PyTorch.

Comme déjà mentionné dans l'introduction, l'AEVD est une classe de modèles génératifs profonds qui généralise l'AEV à la modélisation de données séquentielles (Girin *et al.*, 2021). Les modèles que nous considérons traitent une séquence temporelle ordonnée de données vectorielles $\mathbf{x}_{1:T} = \{\mathbf{x}_t\}_{t=1}^T$ et une séquence ordonnée correspondante de vecteurs latents $\mathbf{z}_{1:T} = \{\mathbf{z}_t\}_{t=1}^T$. Un AEVD est donc définie par la densité de probabilité (DDP) jointe suivante, qui est une généralisation de (1) :

$$p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T},\mathbf{z}_{1:T}) = p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|\mathbf{z}_{1:T})p_{\boldsymbol{\theta}}(\mathbf{z}_{1:T}). \tag{3}$$

Cependant, cette forme ne donne pas beaucoup d'informations sur le processus génératif, et nous préférons utiliser la règle de chaînage des probabilités pour reformuler (3) comme :

$$p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T},\mathbf{z}_{1:T}) = \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{1:t-1},\mathbf{z}_{1:t})p_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{x}_{1:t-1},\mathbf{z}_{1:t-1}). \tag{4}$$

Cette reformulation particulière (parmi d'autres possibilités) est une forme *causale* : $\mathbf{z}_t$ est d'abord généré à partir de $\mathbf{z}_{1:t-1}$ et $\mathbf{x}_{1:t-1}$, puis $\mathbf{x}_t$ est généré à partir de $\mathbf{x}_{1:t-1}$ et $\mathbf{z}_{1:t}$. Ces dépendances sont implémentées à l'aide de RNRs : les paramètres des distributions génératives de $\mathbf{z}_t$ et $\mathbf{x}_t$ sont les sorties des RNRs qui prennent $\mathbf{z}_{1:t-1}$ et $\mathbf{x}_{1:t-1}$ (et $\mathbf{z}_t$ pour la génération de $\mathbf{x}_t$) en entrée.

Les différents modèles AEVD que nous avons cités en introduction sont tous des cas particuliers de l'expression générale (4) où les dépendances sont éventuellement simplifiées. Dans la suite de l'article, nous considérons les six modèles suivants : le modèle *Deep Kalman Filter* (DKF) (Krishnan *et al.*, 2015), le *Stochastic Recurrent Neural Network* (STORN) (Bayer & Osendorfer, 2014), le *Variational Recurrent Neural Network* (VRNN) (Chung *et al.*, 2015), un autre type de réseau neuronal récurrent stochastique (SRNN) (Fraccaro *et al.*, 2016), le *Recurrent Variational Auto-Encodeur* (RVAE) (Leglaive *et al.*, 2020) et le *Disentangled Sequential Auto-Encodeur* (DSAE) (Li & Mandt, 2018). Les formes simplifiées correspondantes des distributions génératives sont données dans le tableau 1. On peut noter que VRNN est l'AEVD le plus riche possible en termes de dépendances des variables puisque toutes les dépendances dans (4) sont conservées, alors qu'en revanche, le VAE d'origine peut être vu comme un AEVD où toutes les dépendances temporelles ont été supprimées.

Modèle	Référence	Génération de $\mathbf{z}_t$	Génération de $\mathbf{x}_t$
VAE	(Kingma & Welling, 2014)	$p_{\boldsymbol{\theta}}(\mathbf{z}_t)$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t\|\mathbf{z}_t)$
RVAE	(Leglaive *et al.*, 2020)	$p_{\boldsymbol{\theta}}(\mathbf{z}_t)$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t\|\mathbf{z}_{1:t})$
STORN	(Bayer & Osendorfer, 2014)	$p_{\boldsymbol{\theta}}(\mathbf{z}_t)$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t\|\mathbf{x}_{1:t-1},\mathbf{z}_{1:t})$
DKF	(Krishnan *et al.*, 2015)	$p_{\boldsymbol{\theta}}(\mathbf{z}_t\|\mathbf{z}_{t-1})$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t\|\mathbf{z}_t)$
DSAE	(Li & Mandt, 2018)	$p_{\boldsymbol{\theta}}(\mathbf{z}_t\|\mathbf{z}_{1:t-1})$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t\|\mathbf{z}_t,\mathbf{v})$
VRNN	(Chung *et al.*, 2015)	$p_{\boldsymbol{\theta}}(\mathbf{z}_t\|\mathbf{x}_{1:t-1},\mathbf{z}_{1:t-1})$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t\|\mathbf{x}_{1:t-1},\mathbf{z}_{1:t})$
SRNN	(Fraccaro *et al.*, 2016)	$p_{\boldsymbol{\theta}}(\mathbf{z}_t\|\mathbf{x}_{1:t-1},\mathbf{z}_{t-1})$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t\|\mathbf{x}_{1:t-1},\mathbf{z}_t)$

TABLE 1 – Hypothèses d'indépendance conditionnelle pour divers modèles de la famille AEVD. Pour DSAE, $\mathbf{v}$ est une variable supplémentaire définie à l'échelle d'une séquence de données ; voir (Li & Mandt, 2018; Girin *et al.*, 2021) pour plus de détails.

La méthodologie d'inférence et d'apprentissage d'un modèle AEVD suit celle de l'AEV : Définition d'un modèle d'inférence $q_{\boldsymbol{\phi}}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$ (puisque la distribution a posteriori exacte $p_{\boldsymbol{\theta}}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$ est non calculable analytiquement), chaînage de l'encodeur et du décodeur, et apprentissage en maximisant la BIV sur les données d'apprentissage. Comme pour le modèle génératif, on peut écrire le modèle d'inférence sous la forme générale suivante, en utilisant la règle de chaînage des probabilités :

$$q_{\boldsymbol{\phi}}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}) = \prod_{t=1}^{T} q_{\boldsymbol{\phi}}(\mathbf{z}_t|\mathbf{z}_{1:t-1},\mathbf{x}_{1:T}). \tag{5}$$

Nous pouvons voir que l'expression (5) est causale concernant les vecteurs latents passés $\mathbf{z}_{1:t-1}$ mais pas concernant la séquence complète des vecteurs observés $\mathbf{x}_{1:T}$. Comme pour le modèle génératif, les dépendances dans (5) peuvent être simplifiées (ou non) si on veut par exemple que le modèle d'inférence ait la même structure de dépendances que la distribution a posteriori exacte (Bishop, 2006, Chapitre 8), ou si on veut avoir une implémentation causale pour permettre l'inférence en ligne.

La BIV pour la classe AEVD est définie de façon générale par (pour une séquence de données) (Girin *et al.*, 2021) :

$$\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta},\mathbf{x}_{1:T}) = E_{q_{\boldsymbol{\phi}}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})}\big[\ln p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T},\mathbf{z}_{1:T}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})\big]. \tag{6}$$

La forme développée, obtenue en réinjectant (4) et (5) dans (6) et en utilisant un calcul en "cascade", est donnée dans (Girin *et al.*, 2021). En fonction des modèles génératifs et d'inférence spécifiques (choisis par l'utilisateur), cette forme développée peut être simplifiée. De manière générale, exprimer la BIV sous une forme différentiable par rapport à $\boldsymbol{\phi}$ et $\boldsymbol{\theta}$ nécessite un échantillonnage de $\mathbf{z}_{1:T}$, qui se fait ici de manière récursive. Cet échantillonnage est alterné avec le calcul du gradient de la BIV sur un ensemble de données d'apprentissage et la mise à jour des paramètres (voir (Girin *et al.*, 2021) pour plus de détails).

3 Application à la modélisation de spectrogrammes de parole

Dans cette section, nous illustrons l'utilisation des AEVDs pour la modélisation des spectrogrammes de signaux de parole. Plus précisément, une séquence de données $\mathbf{x}_{1:T}$ traitée par un AEVD est le spectrogramme de puissance d'un signal de parole obtenu en prenant le module carré de sa

Transformée de Fourier à Court Terme (TFCT). Chaque vecteur $\mathbf{x}_t = \{x_{t,f}\}_{f=0}^{F-1}$ de la séquence est le spectre de puissance à court terme du signal à la trame t (et f est le canal fréquentiel). Chaque coefficient $x_{t,f}$ est supposé suivre une distribution Gamma avec le paramètre de forme égal à 1 et un paramètre d'échelle noté $\sigma_{\mathbf{x},t,f}$ qui peut être interprété comme la densité spectrale de puissance du signal (Févotte *et al.*, 2009; Girin *et al.*, 2019). Dans le contexte AEVD le plus général, le vecteur de paramètres $\boldsymbol{\sigma}_{\mathbf{x},t}(\cdot) = \{\sigma_{\mathbf{x},t,f}(\cdot)\}_{f=0}^{F-1}$ dépend de $\mathbf{x}_{1:t-1}$ et $\mathbf{z}_{1:t}$ et est fourni par le RNR décodeur prenant $\mathbf{x}_{1:t-1}$ et $\mathbf{z}_{1:t}$ en entrée. La distribution générative du vecteur latent $p_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1})$ est définie comme une gaussienne avec un vecteur moyen $\boldsymbol{\mu}_{\mathbf{z},t}(\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1})$ et une matrice de covariance diagonale définie par les entrées du vecteur $\boldsymbol{\sigma}^2_{\mathbf{z},t}(\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1})$. Ces deux vecteurs sont fournis par un RNR prenant ici $\mathbf{x}_{1:t-1}$ et $\mathbf{z}_{1:t-1}$ en entrée. Ces dépendances sont simplifiées selon le tableau 1 en fonction du modèle AEVD utilisé. L'encodeur suit l'équation (5) où $q_{\boldsymbol{\phi}}(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{x}_{1:T})$ est une gaussienne avec un vecteur moyen $\boldsymbol{\mu}_{\boldsymbol{\phi},t}(\mathbf{z}_{1:t-1}, \mathbf{x}_{1:T})$ et une matrice de covariance diagonale définie par les entrées du vecteur $\boldsymbol{\sigma}^2_{\boldsymbol{\phi},t}(\mathbf{z}_{1:t-1}, \mathbf{x}_{1:T})$. Ces deux vecteurs sont fournis par le RNR encodeur. Ici aussi, les dépendances peuvent être simplifiées, et dans nos expériences, pour chaque modèle AEVD, nous avons utilisé le modèle d'inférence décrit dans l'article original correspondant.

Non seulement les AEVDs diffèrent par les dépendances entre variables, mais un même AEVD peut avoir de nombreuses implémentations différentes. Nous avons utilisé ici l'implémentation des six AEVDs de la table 1 décrite dans (Girin *et al.*, 2021, chapitre 13). Nous ne décrivons pas cette implémentation en détails ici par souci de concision. Mentionnons simplement que d'une manière générale, nous avons essayé de trouver un bon compromis entre le respect de l'architecture du modèle tel que décrit dans l'article original et la garantie d'une comparaison équitable entre les différents modèles pour la tâche d'analyse-resynthèse de la parole. Des modules similaires sur les différents modèles AEVD sont donc implémentés avec le même nombre de couches, d'unités par couches et les mêmes fonction d'activation. De plus, les spécifications suivantes sont communes à tous les modèles : La dimension du vecteur d'observation $\mathbf{x}_t$ et du paramètre de sortie $\boldsymbol{\sigma}_{\mathbf{x},t}(\cdot)$ vaut $F = 513$ (la taille de la fenêtre d'analyse TFCT étant fixée à $N = 1024$ avec un recouvrement de 50%) ; La dimension du vecteur latent $\mathbf{z}_t$ est fixée à $L = 16$; La dimension des vecteurs d'états internes cachés des RNRs est fixée à 128 ; Tous les RNRs sont des réseaux LSTM. Les modèles sont entrainés et évalués sur la base multilocuteurs *Wall Street Journal dataset* (WSJ0) (Garofolo *et al.*, 1993) contenant environ 10 heures de parole (16kHz) en langue Anglaise (1.5h est utilisée comme ensemble de test). La taille de chaque spectrogramme d'apprentissage est fixée à $T = 50$ trames (soit $0{,}8$s de parole). Nous invitons le lecteur à se reporter à (Girin *et al.*, 2021) pour plus de détails sur le protocole expérimental.

Tout d'abord, nous présentons une version synthétique des résultats de l'évaluation comparative quantitative des différents modèles AEVD pour la tâche d'analyse-resynthèse (chaînage de l'encodeur et du décodeur), menée sur cette base de données en langue anglaise, et rapportée initialement dans (Girin *et al.*, 2021). Cette évaluation est menée en considérant les trois métriques suivantes : le *Scale-Invariant Signal-to-Distortion Ratio* (SI-SDR) entre le spectrogramme original et le spectrogramme reconstruit, la *Perceptual Evaluation of Speech Quality* (PESQ, valeurs comprises dans $[-0.5, 4.5]$) (Rix *et al.*, 2001) et la *Extended Short-Time Objective Intelligibility* (ESTOI, valeurs comprises dans $[0, 1]$) (Taal *et al.*, 2011). Les résultats sont présentés à la table 2. Nous constatons tout d'abord que les différents modèles de AEVD permettent une bonne reconstruction du spectrogramme avec un SI-SDR allant de 6.9 à 11 dB, des scores PESQ allant de "acceptable" à "bon", et des scores ESTOI la plupart du temps supérieurs à 0.90, ce qui suggère une bonne intelligibilité de la parole reconstruite. De plus, tous les modèles AEVD considérés présentent de meilleures performances que l'AEV standard, ce qui confirme l'intérêt (attendu) de modéliser explicitement la dynamique du signal de parole. Parmi les différents AEVD, les meilleurs performances sont obtenues avec SRNN puis avec VRNN qui sont

AEVD	SI-SDR (dB)	PESQ	ESTOI
VAE	5.3	2.97	0.83
DKF	9.3	3.53	0.91
STORN	6.9	3.42	0.90
VRNN	10.0	3.61	0.92
SRNN	11.0	3.68	0.93
RVAE	9.0	3.49	0.90
DSAE	9.2	3.55	0.91

TABLE 2 – Performances des modèles AEVD testés dans notre expérience d'analyse-resynthèse de la parole. Les scores SI-SDR, PESQ et ESTOI sont moyennés sur le sous-ensemble de test de WSJ0.

les modèles avec de nombreuses dépendances entre variables observables et latentes. Ces modèles sont donc a priori les mieux "armés" pour capturer et restituer la structure complexe du signal de parole. Les performances des modèles DKF et DSAE sont proches, et légèrement moins bonnes que celles obtenues avec le modèle VRNN. Rappelons que ces modèles ne supposent pas de dépendance explicite entre $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$, mais seulement entre $\mathbf{z}_{t-1}$ (ou $\mathbf{z}_{1:t-1}$ pour le DSAE) et $\mathbf{z}_t$. Ils n'ont donc pas la même "puissance explicative" que les modèles SRNN et VRNN. Enfin, les performances les moins bonnes sont systématiquement obtenues avec le modèle STORN. Une des raisons pouvant expliquer ce résultat est que le modèle d'inférence de STORN ne respecte pas la structure exacte de la distribution a posteriori (il n'utilise ni $\mathbf{z}_{1:t-1}$, ni $\mathbf{x}_{t+1:T}$).

Nous présentons à présent une évaluation qualitative des différentes modèles de AEVD. Ces derniers sont ici utilisés pour reconstruire le spectrogramme d'amplitude d'un enregistrement d'une phrase en Français, et donc dans une autre langue que celle considérée lors de l'apprentissage, et prononcée par une locutrice n'étant pas dans la base d'apprentissage. Les modèles sont tout d'abord utilisés en analyse-resynthèse en chaînant l'encodeur et le décodeur (comme pour l'évaluation quantitative présentée précédemment), puis en génération pure. L'encodeur n'est alors plus utilisé, et seul le décodeur est utilisé pour générer le spectrogramme, en étant initialisé par le contexte gauche du signal (i.e., la première seconde de l'enregistrement). Les résultats sont illustrés à la figure 1. Dans le mode analyse-resynthèse, on observe que les spectrogrammes reconstruits sont tous relativement proches du spectrogramme original. Les trajectoires formantiques sont globalement correctement restituées, mais la structure fine du spectre est légèrement lissée, ce qui est typique d'un effet de compression des données (rappelons que la dimension du vecteur d'observation est 513, la dimension du vecteur latent est 16, et celle des vecteurs d'état internes cachés des RNRs est 128).

En mode génération pure (à partir de 1s), l'AEV n'est pas capable de générer un spectrogramme réaliste. Les trames successives n'ont pas de cohérence temporelle et présentent des transitions abruptes. Les résultats sont bien meilleurs avec les modèles AEVD. Par exemple, le spectrogramme généré par DKF présente une structure harmonique évoluant de façon cohérente au cours du temps, ainsi que des trajectoires formantiques compatibles avec un signal de parole. Cette cohérence temporelle est ici assurée par le processus de Markov qui sous-tend l'évolution des dimensions latentes dans ce modèle. En ce qui concerne RVAE, même si les vecteurs latents sont échantillonnés indépendamment à partir d'une distribution gaussienne standard, nous observons une forme de cohérence temporelle dans le spectrogramme généré. Le modèle RVAE est donc capable de "recréer" cette cohérence en combinant les vecteurs d'une séquence non corrélée. Pour VRNN, nous pouvons observer des segments qui ressemblent à différents phonèmes, avec des trajectoires formantiques compatibles avec des voyelles et des segments de bruit large-bande compatibles avec des fricatives.

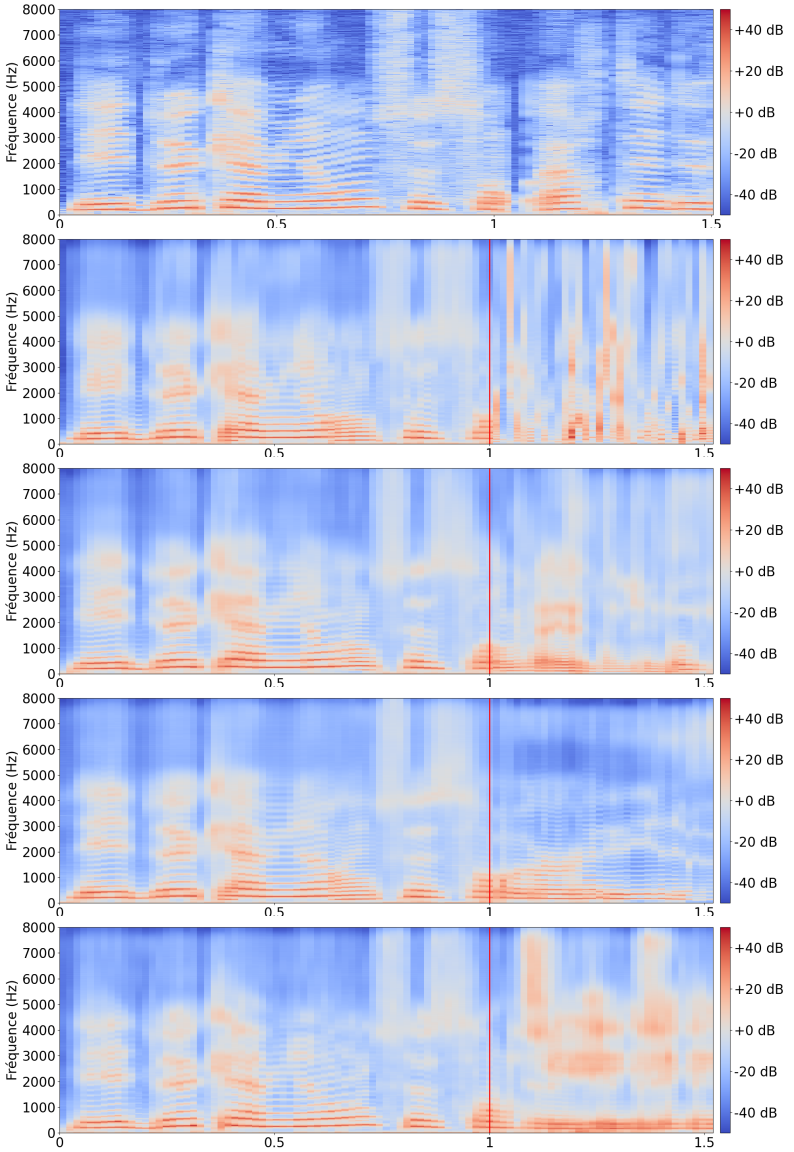

FIGURE 1 – Exemples de spectrogrammes de puissance pour une phrase prononcée par une locutrice (la phrase est : “les deux camions se sont heurtés de face”), reconstruits (de 0 à 1s, avant la ligne rouge verticale) puis générés (de 1 à 1,5s, après la ligne rouge) par différents modèles AEVD. De haut en bas : Spectrogramme original, spectrogrammes reconstruits par AEV, RVAE, DKF, VRNN.

4 Conclusion

Nous avons proposé d’unifier différents modèles génératifs profonds à variables latentes dans une classe générique appelée auto-encodeurs variationnels dynamiques (AEVD). Différents modèles tels que DKF, DSAE, RVAE, SRNN, VRNN et STORN peuvent être vu comme des instances particulières

d'AEVD qui diffèrent dans les dépendances temporelles entre variables observées et variables latentes. Nous avons réalisé une évaluation comparative de ces modèles sur une tache d'analyse-synthèse de spectrogrammes de parole et nous avons illustré qualitativement leur capacité à générer des portions complètes de signal vocal, l'initialisation étant faite sur une portion de parole existante. L'implémentation unifiée de ces différents modèles d'AEVD (basée sur *PyTorch*) est disponible sur `https://team.inria.fr/robotlearn/dvae/`. Nous pensons que cette boite à outils sera utile à la communauté du traitement automatique de la parole, par exemple pour des applications de débruitage, de synthèse, de conversion, ou bien pour l'apprentissage auto-supervisé de représentations.

Remerciements

Ce travail a été soutenu par MIAI@Grenoble Alpes (ANR-19-P3IA-0003) et par la Commission Européenne (projet H2020 SPRING sous GA #871245).

Références

AKUZAWA K., IWASAWA Y. & MATSUO Y. (2018). Expressive speech synthesis via modeling expressions with variational autoencoder. In *Conference of the International Speech Communication Association (INTERSPEECH)*, Hyderabad, India.

BANDO Y., MIMURA M., ITOYAMA K., YOSHII K. & KAWAHARA T. (2018). 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)*, Calgary, Canada.

BAYER J. & OSENDORFER C. (2014). Learning stochastic recurrent networks. *arXiv preprint arXiv :1411.7610*.

BIE X., GIRIN L., LEGLAIVE S., HUEBER T. & ALAMEDA-PINEDA X. (2021). A benchmark of dynamical variational autoencoders applied to speech spectrogram modeling. In *Conference of the International Speech Communication Association (INTERSPEECH)*, Brno, Czech Republic.

BISHOP C. (2006). *Pattern Recognition and Machine Learning*. Springer.

BLAAUW M. & BONADA J. (2016). Modeling and transforming speech using variational autoencoders. In *Conference of the International Speech Communication Association (INTERSPEECH)*, San Francisco, CA.

CHUNG J., KASTNER K., DINH L., GOEL K., COURVILLE A. & BENGIO Y. (2015). A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, Montreal, Canada.

FABIUS O. & VAN AMERSFOORT J. R. (2014). Variational recurrent auto-encoders. *arXiv preprint arXiv :1412.6581*.

FÉVOTTE C., BERTIN N. & DURRIEU J.-L. (2009). Nonnegative matrix factorization with the itakura-saito divergence : With application to music analysis. *Neural Comp.*, **21**(3), 793–830.

FRACCARO M., KAMRONN S., PAQUET U. & WINTHER O. (2017). A disentangled recognition and nonlinear dynamics model for unsupervised learning. In *Advances in Neural Information Processing Systems*, Long Beach, CA.

FRACCARO M., SØNDERBY S. K., PAQUET U. & WINTHER O. (2016). Sequential neural models with stochastic layers. In *Advances in Neural Information Processing Systems*, Barcelona, Spain.

GAROFOLO J., GRAFF D., PAUL D. & PALLETT D. (1993). Csr-i (wsj0) sennheiser ldc93s6b. https ://catalog.ldc.upenn.edu/ldc93s6b. *Philadelphia : Linguistic Data Consortium*.

GIRIN L., LEGLAIVE S., BIE X., DIARD J., HUEBER T. & ALAMEDA-PINEDA X. (2021). Dynamical variational autoencoders : A comprehensive review. *Foundations and Trends in Machine Learning*, **15**(1–2), 1–175.

GIRIN L., ROCHE F., HUEBER T. & LEGLAIVE S. (2019). Notes on the use of variational autoencoders for speech and audio spectrogram modeling. In *Digital Audio Effects Conference (DAFx)*, Birmingham, UK.

HSU W.-N., ZHANG Y. & GLASS J. (2017a). Learning latent representations for speech generation and transformation. *arXiv preprint arXiv :1704.04222*.

HSU W.-N., ZHANG Y. & GLASS J. (2017b). Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in Neural Information Processing Systems*, Long Beach, CA.

KINGMA D. P. & WELLING M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, Banff, Canada.

KRISHNAN R., SHALIT U. & SONTAG D. (2015). Deep Kalman filters. In *arXiv preprint arXiv :1511.05121*.

LEGLAIVE S., ALAMEDA-PINEDA X., GIRIN L. & HORAUD R. (2020). A recurrent variational autoencoder for speech enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain.

LEGLAIVE S., GIRIN L. & HORAUD R. (2018). A variance modeling framework based on variational autoencoders for speech enhancement. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Aalborg, Denmark.

LEGLAIVE S., GIRIN L. & HORAUD R. (2019). Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK.

LI Y. & MANDT S. (2018). Disentangled sequential autoencoder. In *International Conference on Machine Learning (ICML)*, Stockholm, Sweden.

REZENDE D. J., MOHAMED S. & WIERSTRA D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning (ICML)*, Beijing, China.

RIX A., BEERENDS J., HOLLIER M. & HEKSTRA A. (2001). Perceptual evaluation of speech quality (PESQ) : A new method for speech quality assessment of telephone networks and codecs. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, Utah.

TAAL C. H., HENDRIKS R. C., HEUSDENS R. & JENSEN J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, **19**(7), 2125–2136.