



Méthodes neuronales pour le traitement de la parole : vers le démêlage des attributs de la voix

Olivier Zhang

► To cite this version:

Olivier Zhang. Méthodes neuronales pour le traitement de la parole : vers le démêlage des attributs de la voix. Intelligence artificielle [cs.AI]. Université de Rennes (2023-..), 2023. Français. NNT : . tel-04532447

HAL Id: tel-04532447

<https://hal.science/tel-04532447>

Submitted on 4 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES

ÉCOLE DOCTORALE N° 601

*Mathématiques, Télécommunications, Informatique, Signal, Systèmes,
Électronique*

Spécialité : *INFO*

Par

Olivier ZHANG

**Méthodes neuronales pour le traitement de la parole : vers le
démêlage des attributs de la voix**

Thèse présentée et soutenue à Rennes, le 21 décembre 2023

Rapporteurs avant soutenance :

Benoit FAVRE Professeur des universités, LIS, HDR, Université d'Aix-Marseille
Nicolas OBIN Maître de conférences, IRCAM, HDR, Sorbonne Université

Composition du Jury :

Président :

Examineurs :

Nicholas EVANS

Nicolas OBIN

Marie TAHON

Dir. de thèse :

Damien LOLIVE

Co-dir. de thèse :

Nicolas GENGEMBRE

Olivier LE BLOUCH

Professeur des universités, HDR, EURECOM

Maître de conférences, IRCAM, HDR, Sorbonne Université

Professeur des universités, LIUM, HDR, Université du Mans

Professeur des universités, IRISA, Université de Rennes

Chercheur, Orange Innovation

Chercheur, Orange Innovation

RÉSUMÉ

Des machines présentant une capacité de parler proche de celle de l'humain sont présentes dans beaucoup de fictions : HAL dans l'Odyssée de l'espace, C3PO dans Star Wars, ou plus généralement dans toute œuvre de science-fiction impliquant des interactions homme/machine. Mais contrairement à ce que ces références peuvent laisser penser, la synthèse de la parole, pleinement expressive et adaptée au contexte, n'est pas encore un problème totalement résolu.

La construction de machines générant de la parole à partir d'un texte, tâche désignée sous le nom de synthèse vocale, a une longue histoire. Mais durant ces dernières décennies, les progrès de l'informatique et des technologies numériques ont permis un très rapide progrès de ces méthodes de génération de la parole, notamment grâce à l'avènement des réseaux de neurones et de l'apprentissage profond.

De telles technologies facilitent la production de contenus audio, comme des livres audio, la vocalisation de contenus en ligne, ou le développement d'assistants virtuels. Plus généralement, la synthèse vocale améliore l'accessibilité aux contenus écrits pour les personnes ayant une déficience visuelle, et ouvre la voie à des expériences plus immersives et interactives. À une époque où il suffit de dire "Ok Google" ou "Hey Siri" pour entamer une conversation avec un appareil tenant dans une poche, la synthèse vocale presque indistinguishable de l'humain est reconnue comme étant atteinte. Cependant, le contrôle des attributs de la voix reste insuffisant, ce qui suscite un grand intérêt dans les récentes activités de recherche.

Le mécanisme de production de la voix ayant beau être bien compris, il est encore difficile d'en décrire précisément les caractéristiques subjectives telles que les attributs de la voix (claire, brillante, sombre, rauque...) ou la façon de parler (ton, émotion, souffle...). Cela constitue un obstacle de plus vers la synthèse vocale totalement contrôlable et adaptable à volonté. Un tel système est l'objectif ultime à atteindre pour réaliser une synthèse vocale personnalisable, la création de voix originales pour le doublage, un parfait clonage de la voix ou une synthèse vocale pleinement adaptée au contexte.

Aperçu bibliographique

Parole et synthèse

La parole est un moyen de communication séculaire, née du développement du conduit vocal de l'être humain il y a maintenant des milliers d'années. La parole est un moyen commode pour exprimer ses idées avec du son, permettant la transmission efficace de connaissances et d'idées. Aiguisée au fil des siècles, à travers des cultures nombreuses et variées, la parole est le moyen de communication privilégié pour un locuteur afin d'encoder des mots, transmettre ses émotions, ses intentions ou tout simplement son identité.

Ces grandes quantités et variétés d'information transitant dans la parole posent un véritable défi dans la description formelle de la parole et son analyse. Trouvant ses origines dans les poumons, un flux d'air devient audible une fois passé par les cordes vocales, et façonné par les résonateurs supraglottiques, comme la langue et les lèvres, pour produire un large éventail de sons possibles. Ce flux d'air subit des variations de pression et résonne le long du système articulatoire, lui conférant des propriétés acoustiques décrivant la forme d'onde perçue par un auditeur. La perception et l'interprétation de cette onde sonore par l'humain dessinent un sujet à part entière. Les mots prononcés, formant la partie linguistique de la parole, sont décodés sous forme de catégories (phonèmes, syllabes). L'état mental et les intentions d'un locuteur sont interprétés via des indications prosodiques, faisant partie de l'information paralinguistique. Les traits anatomiques d'un locuteur, englobés dans la part extralinguistique de l'information transportée par la parole, sont intuitivement assimilés, permettant aux individus de se reconnaître entre eux.

Les informations linguistiques, paralinguistiques et extralinguistiques, se distinguent donc par la nature de l'information encodée, et également par l'échelle temporelle des variations permettant de les transmettre, et de les comprendre. Les phonèmes et les syllabes sont formés de variations à très court terme, de l'ordre d'une dizaine de millisecondes. Les composants paralinguistiques, prosodie, émotion, style d'élocution, se transmettent par des variations plus long terme, couvrant plusieurs mots, ou des phrases entières. Enfin, les traits extralinguistiques, plus stables, liés à l'anatomie, aux spécificités et aux irrégularités de l'appareil phonatoire de chacun, peuvent être considéré constants, au moins le temps d'une étude conventionnelle.

De plus, une même phrase trouve une infinité de manières d'être prononcée, tout comme les émotions ou la prosodie n'ont pas une manière unique d'être exprimées. Ce manque d'invariance dans la parole est à la base de la complexité de sa modélisation : au-

cune caractéristique objective, acoustique ou articulatoire, n’a encore été identifiée comme suffisamment stable pour déterminer de manière exacte les attributs subjectifs de la parole. C’est cette intrication entre descripteurs objectifs et subjectifs qui est aujourd’hui au cœur des enjeux des travaux en cours et à venir sur le traitement automatique de la parole.

Ainsi, l’apprentissage automatique a montré un très grand potentiel pour modéliser les ondes sonores et gérer ce manque d’invariance dans la production de la parole. Avec l’arrivée en force des réseaux de neurones depuis ces dernières années, des systèmes extrêmement performants en termes de transcription, reconnaissance du locuteur, ou de synthèse vocale, ont fait leur apparition. Plusieurs solutions développées dans l’industrie ont atteint la maturité pour être déployées et proposées au grand public, rendant aujourd’hui le traitement de la parole augmentée par l’intelligence artificielle accessible à tous.

Bien que la génération de la parole pour la machine ait atteint des performances proches de l’humain, le contrôle fin des attributs de la voix reste encore limité. Grâce aux méthodes de conversion de la voix, des systèmes sont capables d’extraire une caractéristique donnée (identité du locuteur, prosodie) d’un énoncé, pour l’injecter dans un autre. Mais sans énoncé de référence, il reste compliqué de personnaliser chacun des attributs de la voix, de manière intuitive, et à volonté. Autrement dit, les réseaux de neurones actuels ne sont pas encore capables de comprendre les différents composants perceptuels de la parole, et leurs relations, ce qui limite leur usage dans les cas où il serait souhaitable d’avoir la main sur des aspects précis de l’énoncé à générer.

Démêler les intrications du monde

Le manque d’interprétabilité et de structure dans les abstractions apprises par les réseaux de neurones, appelées espaces latents, est un problème bien connu, faisant l’objet de beaucoup d’études. Il s’agit de comprendre les prédictions des modèles appris, pour en vérifier le bon fondement, éviter les biais liés aux données d’apprentissage, et rendre les modèles proposés robustes et fiables dans le cadre d’applications critiques. Comme avancé par Bengio et al. (2013) [13], une bonne représentation doit comporter certaines propriétés, dont celle d’être démêlées. En effet, il est supposé qu’étant donné un ensemble d’observations (un corpus d’images, d’énoncés), il existe un ensemble de facteurs génératifs, ne comportant pas de relations causales (voir Pearl [173]), à l’origine du processus de génération du “monde observable”, dont on ne dispose que d’échantillons. Une représenta-

tion dite démêlée, est alors capable d’extraire ces facteurs génératifs, et de les aligner avec certaines dimensions latentes. Ces représentations latentes deviennent alors interprétables, car il devient possible d’isoler certaines variations liées aux données dans un sous-ensemble de l’espace latent appris, et peuvent être employées pour diverses tâches sous-jacentes. Cela rend l’apprentissage de représentations démêlées une manière d’abstraire un ensemble d’observations dans un espace de représentation polyvalent et agnostique. Il s’agit d’un axe de recherche encore récent, mais très prometteur en matière d’interprétabilité, de généralisation, et de contrôle dans des contextes de génération de données, comme l’est la synthèse vocale.

Il convient de soulever un problème majeur de la notion de démêlage, qui est l’absence d’une définition formelle et consensuelle du démêlage, ce qui rend difficile l’apprentissage de représentations démêlées, et rend vague la manière de mesurer le degré de démêlage d’un modèle donné. Plusieurs pistes existent afin de donner un cadre plus formel au démêlage, via la notion de symétrie (Higgins et al (2018) [85]) ou de causalité (Suter et al. (2019) [213]), mais n’indiquent pas comment apprendre des représentations démêlées.

Un type de modèle ayant montré une bonne capacité de démêlage est l’autoencoder variationnel (VAE). Ce type de réseau de neurones a pour but de modéliser les facteurs génératifs via un espace latent, encodant les données observées sous une forme compacte, et optimisé pour être suffisamment informatif à propos des données pour les reconstruire dans une phase de décodage. Le pouvoir démêlant du VAE trouve ses origines dans la divergence de Kullback-Leibler de la fonction objectif du VAE, qui encourage l’indépendance statistique des composantes de l’espace latent. Partant de cette observation, de nombreuses extensions du VAE sont proposées (β -VAE, β -TCVAE, FactorVAE...), afin de renforcer davantage cette propriété démêlante.

Enfin, en l’absence de définition précise du démêlage, un grand nombre de métriques pour en mesurer le degré sont proposées par la littérature. Ces métriques ont des fonctionnements divers et variés, mais ont pour point commun l’inconvénient de nécessiter la connaissance des facteurs à démêler. Cela rend l’application de telles métriques non-viable dans des contextes totalement non-supervisés, ce qui est le cas du démêlage des attributs subjectifs de la voix : les corpus sont souvent annotés en identité, genre, parfois émotion, mais rarement avec des étiquettes plus précises et subjectives (voix rauque, sombre, claire, ou ton calme, stressé, ironique). C’est pourquoi la plupart des études sur l’apprentissage de représentations démêlées sont restreintes à des données synthétiques, où les données sont entièrement expliquées par des facteurs génératifs connus, étant très souvent des

images synthétiques.

Contributions en bref

Vers le démêlage de la parole

Afin d'éprouver les modèles de démêlage proposés par la littérature, des expérimentations préliminaires ont été menées sur des corpus d'images synthétiques. Plusieurs valeurs d'hyperparamètres et de taille de l'espace latent sont testés, pour donner un aperçu de leur effet, et valider les intuitions données par les différentes extensions du VAE. Un sous-ensemble des nombreuses métriques existantes est également éprouvé, pour avoir une idée de leurs comportements, et évaluer l'accord entre métriques. Il en ressort que les modèles appris sont capables, dans une certaine mesure, de démêler correctement certains facteurs génératifs, mais qu'il reste difficile de s'y retrouver parmi les nombreuses métriques, qui manifestent différents degrés d'optimisme sur le score à attribuer aux modèles, et qui parfois sont en désaccord.

Dans le but d'aller vers le démêlage des attributs de la parole, un corpus de voyelles synthétiques est ensuite introduit, *diSpeech* [247], permettant de combler le manque d'un jeu de données de "parole" analogue aux corpus synthétiques utilisés pour la vision par ordinateur. *diSpeech* est composé de cinq facteurs génératifs : les trois premiers formants F1, F2, F3, permettant de couvrir l'espace des voyelles françaises ; la fréquence fondamentale F0, pour contrôler la hauteur ; et le taux de décroissance de F0, qui introduit une variation temporelle et permet la génération de voyelles plus réalistes. Des valeurs sont fixées empiriquement pour chacun de ces facteurs, et l'ensemble des combinaisons possibles sont générées pour former le corpus complet. À noter que *diSpeech* est extensible en définissant d'autres valeurs pour ces facteurs, et même d'autres facteurs parmi les paramètres de Klatt, le synthétiseur utilisé pour générer les voyelles à partir de ces facteurs¹. Ainsi, le démêlage du corpus proposé est expérimenté, et un exemple d'application est décrit, où la capacité de démêlage des voyelles du corpus réaliste TIMIT est éprouvé à l'aide d'un modèle pré-entraîné sur *diSpeech*.

Pour mener cette entreprise de combler le fossé entre le démêlage et le traitement de la parole, des expériences sont réalisées sur les corpus de parole réaliste Bref120 et TIMIT,

1. Le code pour générer *diSpeech* est disponible sur GitHub : <https://github.com/Orange-OpenSource/diSpeech>

avec un autoencoder variationnel factorisé et hiérarchique (FHVAE). Ce modèle est également basé sur un VAE, mais plus avancé, car il factorise l’espace latent en deux représentations distinctes, et introduit une structure hiérarchique en conditionnant l’apprentissage de l’un par l’autre, afin de représenter la nature multi-échelle des attributs de la parole. Pour évaluer le démêlage de ce modèle, les annotations disponibles de Bref120 et TIMIT sont employés pour appliquer les métriques de démêlage, mais les résultats restent mitigés, et des biais dues aux corrélations entre facteurs faussent les conclusions de la métrique utilisée.

Mesurer le démêlage

Un problème régulièrement soulevé tout au long des expériences menées est la difficulté d’interpréter et d’estimer la fiabilité des métriques. La métrique DCI [55] est recommandée par la littérature, étant donné qu’elle mesure plusieurs propriétés du démêlage, et qu’elle permet une lecture détaillée pour chaque facteur et pour chaque latent.

Il apparaît alors crucial de s’assurer que les métriques utilisées sont fiables, et ne révèlent pas d’incohérences dans la pratique. Un processus est alors proposé, la “décimation de latents”, visant à vérifier que les scores annoncés par DCI sont bien cohérents. Des incohérences sont révélées sur le corpus diSpeech, et une nouvelle métrique, MIDCI, est proposée afin d’améliorer la cohérence sous la décimation de latents [246]. Cette nouvelle métrique se base sur la mesure d’information mutuelle entre latents et facteurs, et assigne des scores de démêlage moins optimistes que DCI, mais plus réalistes au regard du procédé de décimation de latents.

En creusant davantage cette idée de séparation de l’information des facteurs parmi les latents, une métrique basée sur la décomposition partielle de l’information (PID) est proposée. Cette mesure présente l’avantage de prendre en compte les corrélations entre facteurs et entre latents, en décomposant les interactions impliquant plus de deux variables en plusieurs morceaux élémentaires d’information, et en ne gardant que ceux identifiés comme relevant du démêlage. Ainsi, le degré de démêlage d’un facteur est défini comme étant la somme des quantités d’information uniquement capturées par chaque latent, en écartant les informations apprises de manière redondante ou générées par synergie entre latents. Cela laisse ainsi la possibilité aux facteurs complexes d’être appréhendés par plusieurs latents, tant que ces derniers transportent des variabilités distinctes. Bien que prometteuse, le calcul de cette mesure n’est pas trivial, car plusieurs manières de calculer le PID existent et ne donnent pas les mêmes résultats. Le calcul de mesure doit

également faire face à une complexité exponentielle au regard du nombre de variables et de réalisations possibles, comme la plupart des mesures émanant de la théorie de l'information. L'application de cette métrique, afin d'en montrer les vertus, est donc le sujet d'investigations à venir.

Conclusion

La parole est le moyen de communication le plus sophistiqué de l'être humain, et également le plus intriqué. Au delà du contenu linguistique, beaucoup plus d'information transite par la parole, à propos de l'état émotionnel du locuteur, ou de ses intentions, et à propos de lui-même, son identité, ses origines régionales, etc. Alors que l'apprentissage profond démontre d'impressionnants résultats sur des tâches spécifiques et bien cadrés grâce à l'apprentissage supervisé, la récente tendance générale est d'utiliser une grande quantité de données non-annotées, dans une approche auto-supervisée. Cette nouvelle ère, où règnent les modèles pré-entraînés puis adaptés sur des tâches subsidiaires, est témoin de grandes avancées dans l'efficacité du traitement de la parole par l'intelligence artificielle. Cependant, malgré tous ces progrès, l'"appréhension" des attributs subjectifs de la parole reste difficilement atteignable avec les systèmes actuels.

Bien que le système de production de la parole soit aujourd'hui bien compris, il reste difficile de cerner formellement les attributs subjectifs de la parole. Ce manque de taxonomie des caractéristiques de la voix rend difficile la modélisation des variations appréhendées de manière intuitive par l'humain (voix rauque ou claire, ton ironique ou sincère, état détendu ou stressé, etc.). D'un autre côté, des travaux de recherche visant à démêler de manière automatique les facteurs génératifs d'un ensemble d'observations parviennent à séparer des composantes indépendantes au sein de données, et de les contrôler dans un mécanisme de génération. Des résultats encourageants avec ce type d'approche sont démontrés par la littérature, mais leur capacité à démêler les attributs de la parole reste limité.

C'est pourquoi la thèse présentée dans ce manuscrit vise à combler le fossé subsistant entre les systèmes de traitement et de synthèse de la voix d'un côté, et les avancées en matière de démêlage de l'autre. Le corpus de voyelles synthétique diSpeech est ainsi proposé, afin de donner à la communauté scientifique un support pour repousser les limites des algorithmes de démêlage sur des signaux de parole. Le problème toujours ouvert sur la manière de mesurer le démêlage est également soulevé, et un procédé de décimation

de latents est introduit, pour assurer la cohérence des métriques existantes. La métrique MIDCI est alors développée, afin d’améliorer la consistance du DCI au regard de la décomposition de latents, tout en conservant les atouts du DCI.

Pour conclure, cette thèse s’inscrit dans la récente tendance à vouloir rendre plus structurés, explicables et contrôlables les réseaux de neurones, à la capacité et complexité toujours croissantes, alors que les bien nommés grands modèles de langage (LLMs) entrent dans le quotidien du grand public et sont, à l’heure de la rédaction de ce manuscrit, déjà en train de marquer un tournant dans le rapport de l’homme à l’intelligence artificielle. Le traitement de la parole par intelligence artificielle n’est pas en reste : de nombreuses solutions commencent à émerger dans l’industrie, proposant du clonage de la voix, de la traduction conservant l’empreinte vocale, de la synthèse de plus en plus contrôlable pour la création de contenus, etc. Des défis restent cependant à relever pour atteindre un véritable contrôle sur les attributs subjectifs de la parole, afin de pouvoir proposer une synthèse véritablement adaptée à différents contextes et environnements. Ainsi, l’apprentissage profond repousse toujours davantage les limites de la complexité et de la capacité à assimiler de l’information, mais l’enjeu des prochaines avancées majeures résidera dans la capacité des systèmes à démêler les intrications du monde observable, pour bâtir des modèles plus rationnels, fiables, robustes, et alignés avec l’intuition humaine.

ABSTRACT

The past few years’ advances in deep learning have brought unprecedented performances in a wide range of tasks and modalities. Among the factors of those breakthroughs is the learning of informative and contextual hidden representations within models, reached by means of well-defined architectures and training procedures.

Speech analysis and synthesis models are plainly concerned by neural network developments. An increasing number of close-to-human accuracy speech analysis (ASR, ASV, etc.) and near-natural speech generation (conversion, TTS, etc.) models are proposed by the research community, and multiple tools leveraging such technologies are emerging in the industry and are reaching the public.

Nevertheless, the increasing complexity and size of neural networks are causing a significant lack in their interpretability. Moreover, well-structured representations are not enforced by design in developed models. Hence, disentangled representations have emerged, which aim to prioritize representations structured by design related to data explanatory factors, which hopefully are aligned with human perceptions, i.e., interpretable. Such a paradigm to learn representations can be expected to properly recognize and split speech attributes (speaker identity, gender, emotion, expressivity, etc.), which may be leveraged for speech synthesis purposes. However, disentanglement learning is a research topic still in its early stages, needing simple and synthetic data to be developed. It is also lacking a clear and consensual definition and, consequently, a metric to quantify it.

Thus, this thesis endeavors to bridge the gap between speech processing and disentanglement, examining how state-of-the-art disentangling models can be employed to automatically recover speech attribute-related information and ultimately improve control over synthesized speech. To this end, a synthetic dataset of vowels is proposed to experiment and compare existing disentanglement models and metrics. Disentanglement of real speech is also experimented, and limitations are pointed out. Metrics to measure disentanglement are then studied, and unexpected metric behaviors are exposed. Furthermore, an information theoretic-based metric is introduced, which we believe to encourage beneficial properties for learned representations regarding the disclosed inconvenient behaviors and, in the long run, voice attributes.

TABLE OF CONTENTS

Notations	17
Introduction	19
Speaking machines	19
Raise of deep learning for speech synthesis	20
Controllable speech synthesis	21
Problem statement	23
Key findings	24
Thesis outline	25
 I Bibliography	 27
1 Speech and synthesis backgrounds	28
1.1 Speech attributes	29
1.1.1 Speech breakdown	29
1.1.2 Concrete and perceptual entanglement	31
1.2 Speech synthesis	38
1.2.1 Statistical parametric TTS	39
1.2.2 Neural-based synthesis	41
1.2.3 Controlling non-verbal aspects	44
 2 Disentanglement learning background	 49
2.1 Disentangling disentanglement	51
2.1.1 Pink elephant	52
2.1.2 Discovering world's hidden mechanisms	54
2.1.3 Distinction from information factorization	58
2.2 Neural networks for generative factors discovery	61
2.2.1 Variational Autoencoder (VAE)	61
2.2.2 Information capacity: <i>A Latent Space Odyssey</i>	67

TABLE OF CONTENTS

2.2.3	Enforcing disentanglement: <i>The Way We Make Contact</i>	71
2.3	Synthetic corpora	74
2.4	Measuring disentanglement	77
2.4.1	Predictor-based	78
2.4.2	Information theory-based	80
2.4.3	Intervention-based	82
2.4.4	Unsupervised metrics	84
2.5	<i>Opening the pod bay doors</i>	85
II	Speech Attributes Disentanglement	87
3	Towards speech attributes disentanglement	88
3.1	Synthetic image datasets disentanglement	89
3.1.1	Setup and expectations	89
3.1.2	Results	91
3.1.3	Conclusions	102
3.2	diSpeech : a synthetic toy dataset for speech disentangling	104
3.2.1	Corpus description	104
3.2.2	Synthetic vowels disentanglement	107
3.2.3	Real vowels disentanglement	109
3.2.4	Discussions	112
3.2.5	Perspectives	116
3.2.6	Conclusion	117
3.3	Real speech disentanglement	117
3.3.1	Factorized Hierarchical VAE (FHVAE)	118
3.3.2	Training data	120
3.3.3	Training setup	122
3.3.4	Evaluations	123
3.4	Conclusions and discussions	126
4	Measuring disentanglement	129
4.1	Latent decimation for metric consistency	130
4.1.1	Metrics comparison	130
4.1.2	A closer look to metrics	131

4.1.3	Procedure description	133
4.1.4	Results on diSpeech	133
4.2	MIDCI	135
4.2.1	Definition	136
4.2.2	Consistency assertion	137
4.2.3	Conclusions and discussions	139
4.3	Decomposing information to quantify disentanglement	140
4.3.1	Disentangling pieces of information	140
4.3.2	Computing partial information pieces	142
4.4	Conclusions and discussions	143
Conclusion		145
	Key contributions	145
	Discussions and perspectives	146
Acronyms		149
Bibliography		151

NOTATIONS

Numbers and arrays

a	scalar value, random variable realisation
A	scalar matrix
$\text{tr}(A)$	trace of matrix A
$\det A$	determinant of matrix A

Sets

A	set of scalar values
\mathbf{A}	set of vector-valued random variables
\mathcal{A}	domain set

Dataset of observations

$x^{(i)}$	i -th observed data from a dataset
\mathbb{X}	a dataset of observations

Probability and information theory

a	scalar random variable
a_i	element i of random vector \mathbf{a}
\mathbf{a}	vector-valued random variable
$\text{Cov}(\mathbf{x})$	covariance matrix of \mathbf{x}
$\mathbb{E}[\mathbf{x}]$	expectation for \mathbf{x}
$\mathcal{N}(\mathbf{x}; \mu, \Sigma)$	normal distribution with mean μ and covariance matrix Σ
$\mathcal{H}(\mathbf{x})$	entropy of \mathbf{x}
$\mathcal{H}_k(\mathbf{x})$	entropy of \mathbf{x} with base k logarithm
$\mathcal{I}(\mathbf{x}; \mathbf{y})$	mutual information between \mathbf{x} and \mathbf{y}
$\mathcal{I}_p(\mathbf{x}; \mathbf{y})$	mutual information between \mathbf{x} and \mathbf{y} over distribution p

$\mathcal{TC}(x)$	total correlation over the components of x
$D_{KL}(p q)$	Kullback-Leibler divergence of p and q
$\mathcal{U}(t; s_1)$	unique information of source variable s_1 regarding target variable t
$\mathcal{R}(t; s_1 : s_2)$	redundant information between s_1 and s_2 regarding t
$\mathcal{S}(t; s_1 : s_2)$	synergistic information of s_1 and s_2 regarding t
$x \perp\!\!\!\perp y$	x and y are statistically independent
$x \not\perp\!\!\!\perp y$	x and y are statistically dependent
$\text{do}(x = x)$	interventional effect of setting x to the value x

Functions

$\ \cdot\ _p$	p -norm
$\log x$	natural logarithm of x
$f(x; \theta)$	function of x parameterized by θ

INTRODUCTION

Human-like talking machines have rooted references in widespread popular fictions: HAL in *Space Odyssey*, C3PO in *Star Wars*, or essentially any science fiction film with human-machine interactions. Unlike what those famous references might suggest, human-like expressive and contextually tailored speech synthesis is not (yet) a solved problem.

The building of machines that generate words and sentences from text, a task known as Text-to-Speech (TTS), has a long story. But recent advances in computer science and digital technologies have triggered a rapid advancement of speech synthesis methods. Such technologies facilitate the production of audio content, such as audio books, the vocalization of online content, and the development of virtual assistants. More broadly, TTS improves accessibility to written content for people with disabilities and inspires more engaging and interactive experiences. In an age where starting a conversation with a pocket-sized device is as simple as saying “Ok Google” or “Hey Siri”, near-natural speech generation is well acknowledged to be reached. But there is still a lack of control over voice attributes, which is of great interest in recent efforts.

The voice production mechanism is well understood, but it is still hard to precisely describe subjective features such as voice characteristics (bright, shining, dark, hoarse, etc.) or the way of speaking (tone, emotion, breathiness, etc.). This does not ease a fully tunable and controllable generation of speech. Such a system is the ultimate goal to reach in order to achieve customizable TTS, tunable voice dubbing, zero-shot voice cloning, or context-tailored speech synthesis.

Speaking machines

Speech synthesis is the process by which synthetic speech is generated, originally from a given text to enunciate. First attempts to build systems able to produce speech-like sounds were based on manually operated mechanical instruments. Wolfgang von Kempelen built in the late 18th century the first “speaking machine” [103] by modeling human’s vocal tract, with bellows as lungs to supply an airflow and pipes to produce vowels and consonants. Kempelen’s speaking machine paved the way for the development of more

advanced speech synthesizers, as The Voder [54], an electronic *formant synthesis* system i.e., based on vocal tract’s resonant frequencies modeling, with keys to control formants and a foot pedal to control pitch, developed in the late 30s. However, months of practice were needed to properly operate such a system.

Then computer-based systems were developed, enabling richer sound synthesis and whole sentence generation. Among them, Votrax’s speech synthesis system [61] is worth a mention. It is an *articulatory synthesizer*, i.e., vocal tract is modeled with a set of rules and algorithms to generate output articulations and transitions corresponding to an input phoneme sequence. Votrax is also known to have achieved the first computer-assisted pizza delivery ordering in 1974. Achieving more natural results, *Unit selection synthesis* is another type of system based on the concatenation of prerecorded sounds [95].

State-of-the-art synthesis is nowadays achieved with Statistical Parametric Speech Synthesis (SPSS) [243] models, which estimate acoustic features from linguistic features to generate audible waveforms with a vocoder, i.e., a component built to map acoustic features (spectrogram) to a sound wave. Such models were originally based on *Hidden Markov Models* (HMM), but neural networks progressively took over, currently being the core technology of modern speech synthesis systems [216].

As in many domains, artificial intelligence has recently shown very impressive results in speech generation. Research efforts in neural-based speech synthesis are increasingly dynamic, and recent breakthroughs in deep learning have made it possible to reach unprecedented results in TTS. Hence, the industry has already released a wide range of tools and solutions to generate speech, powered by neural networks.

Raise of deep learning for speech synthesis

Deep learning’s seminal works date back to the 1940s, designated as cybernetics. But it is only over the past decades that improvements in hardware, computation capacities, and data availability have led to deep learning’s groundbreaking performances in many research fields, such as computer vision, natural language processing, healthcare, and, at the heart of our interests, speech processing.

In “classical” machine learning approaches, handcrafting an efficient set of features for a given task is mandatory, but might take years of human expert effort. Neural networks avoid such obstacles, as their core concept is precisely based on the automatic learning of abstract representations from observed examples. Such models are called “deep” as

they build up an internal hierarchy of concepts, gaining in abstraction and complexity as a model grows deeper. In other words, neural networks have the ability to learn their own features from data samples and organize them into interconnected layers. Deeper layers capture increasingly abstract concepts out of shallower layers conveying simpler concepts [68].

With a growing community of researchers and industries over the past few years, deep learning has experienced a wide range of developments and improvements in model architectures (CNN [128], ResNet [79], LSTM [88]...), computational efficiency (regularization, hardware acceleration, parallelization...), optimization algorithms (SGD [131], Adam [116], Adagrad [53]...) and so on. Among the major breakthroughs, self-supervised representation learning paradigm advocates training procedures that can leverage a huge amount of unlabeled samples while extracting meaningful information by means of carefully designed objective functions. Self-attention introduced by Transformer model [224] is also worth to mention, as it brings contextually relevant token generation, enabling highly realistic sequence generation. The mentioned developments, and many others, have led to impressive results demonstrated by generative models since the achievement of highly natural multimedia content generation: image, audio, video, or text. Such models have been widely mediatized, namely GPT [166] or DALL-E 2 [183].

Neural networks have demonstrated a great potential for TTS, learning appropriate matching between linguistic and acoustic features, as Tacotron 2 [205]. Neural networks also exhibit strong capacities as vocoders, i.e., to predict waveforms from acoustic features, as WaveNet [165]. Fully end-to-end systems are also developed, as JETS [144], negating the need to articulate two components, unifying acoustic model and vocoder, in a standalone pipeline.

However, in an utterance, non-verbal features, e.g., speaker's own traits, prosody, or rhythm, are not to be neglected as they can convey much more besides linguistic content. Fortunately, deep learning has also shown remarkable capabilities to manipulate such complex variabilities, but still with limitations.

Controllable speech synthesis

The non-verbal characteristics of an utterance can tell us a lot about the speaker (e.g., sex, age, emotional state) and his intentions (e.g., statement, question, irony). Speaker identity can be characterized by its anatomical configuration, e.g., vocal folds, tongue, and

lip settings. Prosodic attributes can be described with acoustic feature variations, e.g., pitch, formants, rhythm, and tone. All in all, non-verbal information is tied to physical cues and can thus, in essence, be expressed through objective descriptors.

From a perceptual viewpoint, the description of speech with common and known to all terms, is mainly based on adjectives borrowed from other senses (e.g., bright, dark, warm, smooth) or tied to the perceived emotional state (happy, stressed, afraid) or intent (e.g., question, irony, command, hesitation). Hence, the subjective descriptors of speech are actually complex entanglements of objective features.

Consequently, achieving speech synthesis by controlling acoustic features to match a perceptual expectation requires profound expert knowledge of phonetics and is quite time-consuming. Even manually tricking the output of a text-to-speech model is a laborious effort. Thus, deep learning-based synthesis techniques handle such subjective concerns by adding non-verbal annotations (e.g., speaker identity, emotion). Then conditioning synthesis models on those labels enables control over perceptual characteristics in synthesis.

More recently, self-supervised models learned on huge amounts of unannotated speech segments are able to extract non-verbal attributes through meaningful representations. Speaker identity might be provided during training in order to discard it from learned representations and prioritize prosodic aspects. Then Voice Conversion (VC) is achieved when non-verbal features from a target segment are extracted and injected into a source segment.

Leveraging the high representational power of deep learning, non-verbal-related tasks have benefited from major advances. Speaker identity is successfully controlled by recent VC models, and expressivity is well addressed by *Expressive Voice Conversion* systems. Even *Cross-lingual Voice Conversion* is being achieved, allowing anyone to seemingly speak another language fluently. Such solutions are already being proposed by industrials such as Coqui², Elevenlabs³, or Respeecher⁴.

However, to reach context-adapted synthesis and natural vocal human-machine interactions, understanding and control over non-verbal features are required. With the principles outlined above, current deep learning models do not provide the means to easily tune fine-grained subjective characteristics. VC is achieved, but the speaker embedding invoked to convey the speaker identity (e.g., X-vector [208]) are only abstract numerical representations, which do not explicitly describe what constitutes speaker’s individuality.

2. <https://coqui.ai>

3. <https://elevenlabs.io>

4. <https://www.respeecher.com>

Such implicit speaker representations do not let one design a voice from scratch with the desired attributes. Similarly, expressive VC is possible only through implicit representations, which can hardly be manipulated to customize prosody for a specific context. The very recent advances in prompting techniques, arising from the fast-growing trend of Large Language Models (LLMs), are worth mentioning. Those approaches enable the very accurate synthesis from textual descriptions of non-verbal aspects and background noise, yet they do not provide fine-grained control over the generated speech.

Note that for supervision or evaluation purposes, annotating speech segments with perceptual labels is a hard and tedious task. Subjective descriptors do not dispose a formal definition, and each person has their own interpretation. The lack of an agreed-upon way to describe a voice and speech hence appears as an additional difficulty towards controlling non-verbal elements in synthesis.

Problem statement

As mentioned earlier, acoustic features are too punctual and entangled to properly synthesize subjective descriptors at will. Hence, one may expect neural networks to be able to learn intermediary representations, which might find correspondences with human intuitions. This leads to the quest for interpretable deep representations that disentangle non-verbal speech attributes.

Such a paradigm is emerging in recent studies, especially in image processing. Disentanglement learning refers to the automatic discovery of explanatory factors of variations within data. This suggests the explicit modeling and separation of hidden variables that “generate” observed data, leading in some sense to an understanding of world’s structure. The comprehension of generative factors in images, e.g., color, shape, or spatial position, is of particular interest for an agent exploring and interacting with a simulated or real world.

Speech is also concerned. As pointed out before, verbal, speaker, and prosodic information are non-trivial aspects often factorized to control one without affecting others. The troubles encountered when trying to distinguish perceptual characteristics of speech might also find answers in a learning procedure dedicated to the automated discernment of speech explanatory temporal patterns, which hopefully would let one control them in synthesis. For instance, if the aforementioned implicit speaker embeddings were disentangling speaker’s voice signature facets, they might become meaningful and efficiently

leveraged to control speaker identity in VC systems.

Besides a growing interest in disentangling model development and analysis, the depicted task is still lacking a clear and consensual definition. This arises from the difficulty of establishing a clear training objective and an appropriate quantitative measure of disentanglement. Until now, disentanglement has been an implicit target in model optimization schemes, not guaranteed to be reached or to match perceptual expectations. Disentanglement is mainly evaluated with visual and qualitative assessments. A wide range of quantitative metrics are also proposed, each measuring different characteristics, making it difficult to decide which one to use. Those metrics are also based on the knowledge of ground-truth factors to disentangle, limiting their usage to synthetic or simple cases.

Disentanglement as an objective to learn speech representations is at the core of some research efforts, but few of them are truly leveraging the theoretical endeavors supplied in disentanglement studies. This is disclosing a gap between speech attribute processing and disentanglement learning, which will be our main concern during the incoming parts of this manuscript, i.e., how to bridge the gap towards speech disentanglement.

Key findings

Among the contributions of the thesis portrayed in this manuscript, *diSpeech*, a corpus of synthetic vowels, lays the groundwork for the aforementioned endeavor. It constitutes an appropriate playground to analyze the behavior of existing disentanglement models and metrics regarding the acoustic features of vowels, i.e., formants and pitch. Experiments on *diSpeech* and real speech segments underline the difficulty of asserting disentanglement through both metrics and listening: metrics exhibit discrepancies between one another, and listening to every single synthesis when exploring directions of a learned latent space (traversals) is a tedious and highly subjective labor.

The complexity of disentanglement assessment being exposed, a *Latent Decimation* procedure is proposed in order to estimate consistency over metrics. We consider that a metric claims a degree of disentanglement for each representation dimension regarding a given factor of variation. If we iteratively remove dimensions and train a predictor each time, we can expect the predictor’s accuracy to have some correlation with the informativeness of the removed dimension. The procedure checks for the existence of such an agreement.

In light of the metrics comparison and lack of consistency, a new metric is coined,

MIDCI. It combines two existing metrics to keep their pros without their cons, and experiments show that it exhibits more consistent behaviors under the latent decimation procedure.

Finally, promising perspectives are developed towards the definition of an information theory-based metric, more precisely by leveraging on Partial Information Decomposition (PID) theory. So far, proposed metrics consider each dimension separately or are inherent to predictors that may be biased. Hence, there is a clear lack of consideration of latent dimension interactions and redundancies, which are precisely managed by PID measures.

Thesis outline

To properly go through the various findings and proposals related in this manuscript, Chapter 1 provides an overview of speech synthesis methods, starting from a broad description of speech and then digging into text-to-speech and non-verbal attributes control. Chapter 2 handles the other shore by surveying disentanglement objectives, methods, metrics, and challenges. In addition, it is believed that the unified discussion about speech attribute intricacies in Chapter 1 and the Variational Autoencoder (VAE) framework with its interpretations and its multiple disentanglement deviations provided in Chapter 2 are of great value as a standalone contribution in the quest of what constitutes voice, why disentanglement happens, and how to bring them together.

Contributions are then described, starting with Chapter 3, which introduces diSpeech [247] and speech disentanglement experiments and results. Chapter 4 relies on preceding findings to propose an in-depth metric analysis procedure, a.k.a., latent decimation, the metric MIDCI [246], and evidence of a PID-based metric relevance. Parts of the work described in this manuscript were published in the following articles:

- [247] Zhang, Olivier et al., « diSpeech: A Synthetic Toy Dataset for Speech Disentangling », *in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 8557–8561
- [246] Zhang, Olivier et al., « An extension of disentanglement metrics and its application to voice », *in: Proc. INTERSPEECH 2023*, 2023, pp. 2878–2882, DOI: 10.21437/Interspeech.2023-383

PART I

Bibliography

SPEECH AND SYNTHESIS BACKGROUNDS

The primary material of our concerns in this manuscript is speech. It is a secular means of communication dawned from human’s vocal tract development hundreds of thousands of years ago. Speech provides a convenient way to express thoughts with sounds, enabling the effective sharing of knowledge and ideas. Sharpened by centuries of development through sprawling cultures, speech is the privileged medium for a speaker to encode words, transmit its emotional state, its intentions, or merely its identity.

This wide range of information poses challenges in the formal description and analysis of speech. Originating from the lungs, an airflow becomes audible upon passing through vocal folds and shaped by the supraglottal vocal tract (e.g., tongue, lips) to produce a vast variety of sounds. As the airflow sustains pressure variations and resonances along the vocal tract, it acquires acoustic features that describe the sound wave frequencies received by the listener [132]. The perception and interpretation of such a sound wave by humans is a full-fledged topic. Spoken words are retrieved by decoding the heard signal into quantized categories (i.e., phonemes and syllables). Mental state and intents of a speaker are interpreted through prosodic cues. Anatomical traits of a speaker are perceptually assimilated, allowing individuals to recognize each other. In short, anatomy, acoustics, and perception are intricate but complementary descriptors of speech, as detailed in Section 1.1.

Furthermore, the lack of invariance [4] is an intrinsic complication of speech: no invariant feature (acoustic, anatomical) of speech has yet been identified to properly encode perceived features (phoneme, emotion). The high variability of a single speaker and between speakers to encode a same phoneme or emotion is the main challenge in achieving both recognition and synthesis based on perceptual features.

This manuscript is focused on non-linguistic characteristics in speech synthesis. Multiple methods were developed, but yet none of them have been able to suitably disentangle speech intricacies. In order to design new approaches able to find the parallels between a sound wave and human’s interpretation, speech synthesis systems and the involved challenges are described in Section 1.2.

1.1 Speech attributes

It is fascinating to see the number of domains involved in speech production and understanding, e.g., anatomy, acoustics, phonetics, auditory, or psychoacoustics. It is also interesting to notice that several layers of information are conveyed by speech. By “speech information” is meant any knowledge transmitted, intentionally or non-intentionally, by way of the sound wave and which can be deduced by a listener, e.g., speaker identity, textual content, or prosody. As one is speaking, all those pieces of information are diluted in the sound stream. And yet, it remains intuitive for a speaker to properly encode all parts together, as well as for a listener to correctly decode, distinguish, and interpret each part.

The encoding and decoding processes are the main concerns of speech synthesis and recognition systems, respectively. To clarify the points covered in this manuscript as well as the underlying pitfalls, it is necessary to decompose speech.

1.1.1 Speech breakdown

In order to suitably describe what is speech, a breakdown of speech components is proposed in Figure 1.1. Speech production is mainly driven by the spoken content, but the unspoken content should not be neglected to completely comprehend an uttered message. Berckmoes and Vingerhoets (2004) [14] consider two facets of the information conveyed by speech: verbal and vocal channels, i.e., verbal and non-verbal content. Diving further into non-verbal content, one may distinguish speaker’s own characteristics (e.g., sex, age, regional origins) [201] from the prosodic variations: state (e.g., health, sleepiness, intoxication) [200] and intentions (e.g., statement, question, apology) [199]. To summarize, Figure 1.1 refines speech information into 3 categories [134, 203, 161]: *Linguistic*: the verbatim, i.e., what is said; *Paralinguistic*: style, prosody, or tone-related variations, i.e., how is it said; *Extralinguistic*: speaker’s related characteristics, i.e., who is speaking.

Now, if we examine the produced sound wave, time scale is what can differentiate those information parts. Finer-grained pieces of information are linguistic cues, phonemes being the most punctual events lasting tens to hundreds of milliseconds. Syllables, words, and whole sentences are formed from sequences of such short-term variations. Zooming out to coarser and more transitory variations, paralinguistic cues become visible. They are observable as mid-term variations of sound waves. Two subcategories can be identified: speaker intents, i.e., the message that the speaker intends to convey (which might be

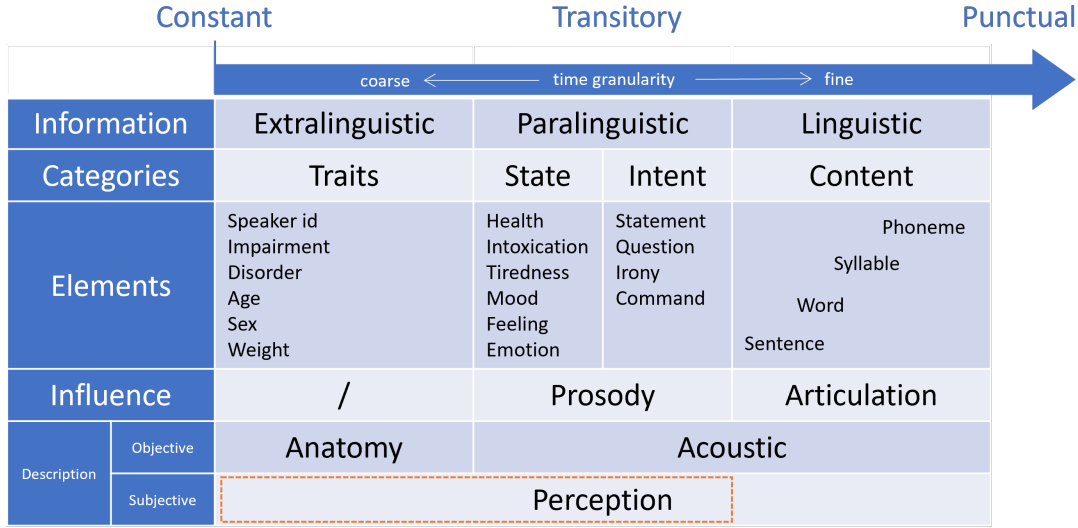


Figure 1.1 – Speech breakdown

different from the actual spoken message), and state, i.e., how the speaker is doing. The latter is considered temporally coarser than the former [199], as what nuances mood and emotion. In other words, expressed intentions are conveyed by shorter-term variations than psychological or physiological states. Coarsest variations lay when zooming out further, pertaining to speaker traits e.g., identity, age, sex, weight, impairment, or disorder. Such cues are considered near-constant, as they change during one’s lifetime but remain fairly stable within the scope of a conventional study.

Regarding the perception of speech (i.e., from a decoding perspective), coarse-to-fine informational influences lay in the proposed time granularity-based ordering. We define “perception” as the procedure by which a human mentally interprets, represents, and organizes information from a heard speech. Extralinguistic influences linguistic cues via intrinsic physiological attributes (e.g., speaker’s identity or sex define pitch range) or personal elocution preferences (e.g., regional or social origins may influence pronunciation and linguistic variations). Paralinguistic impacts linguistic (e.g., tone¹, pitch or amplitude variations influence the phoneme cues). Personality and propensities may differ significantly from speaker to speaker, affecting paralinguistic cues.

Considering how speech is produced (i.e., from an encoding viewpoint), linguistic content is accomplished via articulation, i.e., the orchestration of vocal tract organs to shape airflow to produce consonants and vowels. Paralinguistic content is then expressed (inten-

1. One may rightly notice that in tonal languages, tone is a linguistic factor as well [141].

tionally or not) through prosodic variables, e.g., pitch, amplitude, and rhythm. Physical configuration determines extralinguistic features and, by their very nature, cannot be manipulated by the speaker. Impressionists might trick this statement, but are out of the concerns of this manuscript, yet it would not be without interest to investigate which techniques are using impressionists to mimic, and question them about how they apprehend and analyze, a target voice.

With this insight of speech's nature, we can dive into the main interest of the work described in this manuscript, which pertains to the description of speech. While physical attributes and phenomena are well defined, formalizing perceptual descriptors remains a challenge.

1.1.2 Concrete and perceptual entanglement

The natures of information encoded by a speaker are depicted in Figure 1.1: extralinguistic, paralinguistic, and linguistic. In order to analyze and understand a heard sound wave, i.e., deduce the underlying information from the signal, a proper description is essential. As speech is by essence a subjective means of communication, relations between measurable physical phenomena and perception are highly convoluted.

Objective description

Speech in its conventional form intrinsically conveys linguistic content, which in an utterance can be retrieved through a segmental [34] analysis, i.e., the identification of bounded units within the sound wave. As linguistic content is produced by very short-term air pressure variations, acoustic features of the resulting wave are accurate descriptors to identify the encoded phonemes. Indeed, a time-frequency analysis can reveal the high energy frequencies, i.e., the formants of a speech signal. Formants are well-known and used acoustic features, easily distinguishable in time-frequency analysis (spectrogram), which can be used to determine a pronounced vowel or consonant. Physiologically, phoneme production is well detailed by articulatory depictions: vowels are categorized following tongue (height, backness) and lips (roundness) positioning, and consonants by the constriction place (e.g., labial, coronal), manner (e.g., plosive, fricative), or phonation (voiced/voiceless) [132], and all these variations yield modifications of the formants. However, the encoding of a sequence of phonemes in the sound wave implies phonemic variations due to coarticulation phenomena. Phonemes are intermixed, preventing a triv-

ial mapping between phonemes and speech segments, and giving rise to the full-fledged psycholinguistics question of how a listener comes to accurately decode such a complex sound stream [142].

Coarser speech variations (suprasegmental [34]) are non-trivial to hold. Prosody does not enjoy a formal depiction to express how a sentence has to be said among the numerous possibilities. Punctuation does offer some hints, but is definitely too restricted to cover the whole “shades of meaning conveyed by prosody” [76]. The suprasegmental nature of paralinguistic content leads to complications in the proposal of an abstract representation to transcribe prosody, whereas the linguistic content does own such representations, as the *International Phonetic Alphabet* (IPA) to transcribe phonemes into phones, due to its segmental encoding. Hence, prosody is best described through acoustic variations, as pitch, loudness, or duration [81]. Hirst and Di Cristo [87] express the difference between lexical and non-lexical cues, further studied by Gussenhoven [73]: lexical cues (tone) are affecting word parts (e.g., stress, duration); non-lexical cues (intonation) are utterance-scaled modulations (e.g., pitch contour, rhythm). Lieberman and Michaels (1962) [143], and Huttar (1968) [97] experimentally showed the importance of pitch, amplitude, or duration in the recognition of emotional content. Prosody is quite harder to describe through articulatory measures, but is in the interest of some studies [29, 60, 127].

The main physiological characteristic that differentiates one speaker from another is the vocal folds. Their elasticity, mass, shape, asymmetry, or wetness might affect formants’ bandwidth and frequency. But other parts of the speech production system are involved in the specificity of an individual’s voice. In the respiratory system, lung size and elasticity, and airway dimensions may vary from one to another, leading indirectly to variations in speech sound production, e.g., subglottal pressure fluctuations impacting pitch range, distribution, and contour. Vocal tract configuration is also an essential source of variations among speakers. For instance, vowel formants are altered by sizes, proportions, width of pharynx and mouth, as illustrated in Figure 1.2 from Stevens (1971) [210]. Fricative consonants rely on hard palate and teeth shape. The nasal cavity configuration is also a source of fluctuation in the production of nasal consonants. Overall, an individual’s physiological configuration leaves measurable acoustic traces. The third formant provides a clue of the vocal tract length, and the fourth and fifth formants are relatively stable across vowels, originating from resonances occurring in nearly static regions of the vocal tract [210]. Morphology and body size (height, weight, BMI) have also influences on vocal tract configuration [59] and thus on acoustic cues [32]. Speaker’s preferences of elocution

(e.g., rhythm [40]) or dialect [39] are also relevant clues. For speaker recognition and identification purposes, leveraging on such acoustic features may lead to valuable results [194], but higher-level features such as *Mel-Frequency Cepstral Coefficients* (MFCC), *Perceptual Linear Prediction* (PLP) [83], and *Linear Prediction Cepstrum Coefficients* (LPCC) [154] were developed and demonstrated better performances [136].

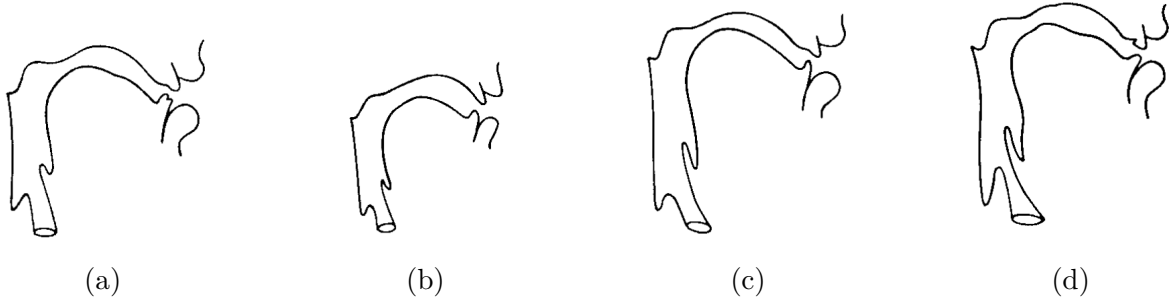


Figure 1.2 – Examples of vocal tract configuration variations, from Stevens (1971) [210]: (a) Reference configuration; (b) Dimensions scaled down; (c) Pharyngeal portion ratio increased; (d) Longer and narrower larynx tube

Since everything in speech arises from the articulatory system, speech can be undertaken in terms of physiological characteristics. Time-dependent variations are, however, tedious to quantify through vocal tract organ movements. Hence, acoustic features are more relevant descriptors and easier to acquire and analyze. But as a listener, one may have their own perception and apprehension of the speaker’s traits, state, and intents.

Subjective description

Despite its technical sophistication, speech is undoubtedly human’s most efficient means of communication. Beyond the linguistic content, a listener can instantly recognize a known speaker or distinguish between unknown ones. He can instinctively discern untold emotions and intents, even from other languages. In addition, when hearing an utterance, one has the innate intuition, akin to a sense of “physical empathy”, of how it was produced [33]. All in all, speech is intuitive to produce, understand and learn, therefore inherently subjective.

Nevertheless, a formal characterization of speech may be necessary for various purposes. Forensic speaker identification relies on an expert’s perception to judge whether multiple recordings are of the same person, sometimes drawing conclusions that contradict the assumptions of a naive listener [190]. Clinical diagnosis of vocal pathology and dys-

phonic patients is processed with listener ratings. Speech synthesis is evaluated through subjective tests to assess heard characteristics, e.g., naturalness, similarity with a target voice, or emotion.

The challenging task of perceptual assessment of voice characteristics has extensively been studied, through what phoneticians use to call *Voice Quality*. No formal definition of voice quality is acknowledged in the concerned works, thence we can consider Trask's (1996) [222] definition:

voice quality *n.* The characteristic auditory coloring of an individual's voice, derived from a variety of laryngeal and supralaryngeal features and running continuously through the individual's speech.

By the term “coloring”, one can promptly notice the subjective character of the task. Voice quality further appears to be the perceived aftermath of extralinguistic features in produced speech. It seems, therefore, appropriate to investigate how phoneticians are addressing this very topic.

Substantial efforts have been carried out to perceptually assess voice quality, especially to establish suitable protocols for detecting speech disorders. The protocol *GRBAS* [86] stands among the most used. It comprises five characteristics to be rated: Grade, i.e., the overall voice quality alteration; Roughness, i.e., vocal fold perturbations; Breathiness, i.e., air leakage from glottal closure; Aesthenia, i.e., speaking weakness; Strain, i.e., laryngeal tension Dejonckere et al. (1996) [38] proposed an extension, *GRBASI*, with the additional characteristic Instability, i.e., voice quality fluctuation over time. *Vocal Profile Analysis* (VPA) [135] is also a widely used protocol, including between 30 and 40 articulatory settings (supralaryngeal, laryngeal, prosodic, and optionally muscular tension features) following the considered version [196]. Another popular protocol is *Buffalo III Voice Profile* [235], which rates nine characteristics relating to tone, pitch, loudness, resonance, tension, and speaking rate. The *Stockholm Voice Evaluation Approach* (SVEA) [75] and the *Consensus Auditory Perceptual Evaluation* (CAPE-V) [104] are also noteworthy protocols from the literature, rating similar aspects of voice quality.

The numerous proposed schemes of perceptual assessment of voice characteristics reflect the complexity of such a task. The first obstacle is merely the definition of aspects to rate. Voice quality is, by nature, multidimensional. Thereby, the considered aspects must cover all relevant articulatory settings while remaining comprehensible and clear enough to avoid understanding mismatches between listeners and perceptually distinguishable from other co-occurring phenomena to be isolated and rated properly. Defined aspects must

also be as independent as possible in order to avoid redundancies, cumbersome rating processes, and evidence overweighting [125, 106, 196].

The scaling scheme of rating is also a decisive choice [126]. Most of the mentioned protocols rely on *Equal-Appearing Interval* (EAI), which defines a discrete scale of n (generally ≤ 10) points to grade the presence/severity of a characteristic. *Visual Analog Scale* (VAS) provides a simple line of 100 or 200mm for each aspect, extremities representing its absence and extreme presence. It has been argued to be more reliable than EAI, as it matches listener’s continuum perception of vocal characteristics [12]. *Direct Magnitude Estimation* (DME) enables listeners to assign arbitrary numbers, often with reference to a particular magnitude (e.g., 100), thereby providing additional freedom of scaling. Conversely, a bipolar rating reduces the rating of each aspect to two opposites (high-low, weak-strong, rough-smooth), making the choice easier. Finally, paired comparison relies on the relative ranking of two recordings with respect to a given aspect, avoiding the formalization of a ground scale [65].

The intrinsic subjectivity of the perceptual assessment of speech characteristics inevitably implies intrarater and interrater agreement and reliability issues to be considered. Interrater agreement quantifies the degree of consensus among the raters regarding a particular speaker. Interrater reliability ensures that ratings are consistent from rater to rater, i.e., raters exhibit similar/proportional variations across voices [218]. For intrarater agreement and reliability, analogous properties are measured, but regarding a single rater and multiple recordings of a same speaker. According to Kreiman et al. (1993) [126], agreement and reliability can vary from an experiment to another, depending on the context. Besides the choice of protocol and scale scheme, raters’ background can have an influence: experienced listeners appear more reliable [41]. Training sessions prior the true rating are recommended, with discussions among listeners to reach an agreed definition of the characteristics to rate [12], or by providing anchor references [126, 49]. However, no consensual rater training procedure was found yet [172].

Heretofore, voice quality was considered with clinical concerns. But this manuscript is interested in voice quality in its broad sense, not specifically focused on pathological voices. The assessment of characteristics of nondysphonic voices with the depicted protocols has been conducted. For instance, San Segundo and Mompean (2017) [195] showed that the voice similarity of monozygotic twins can be on average established with a simplified version of VPA protocol. GRBASI protocol was used by Delvaux and Pillot-Loiseau (2020) [41] to assess the characteristics of healthy voices. They observed that the raters

did not find the rating scheme non-adapted to normal population.

The depicted approaches are focused on the speaker, i.e., on the description of voice regarding perception. But the listener's perception is also of great interest, i.e., the description of the perception regarding a voice. The auditory process is out of the scope of the present work, but the way one interprets voice aspects, what makes the perceptual similarity or dissimilarity of voices, what makes a voice pleasant or unpleasant, are points that have to be considered to achieve context-tailored voice synthesis. Towards this approach, Voiers (1964) [225] attempts to determine the number and nature of perceptual dimensions needed to distinguish voices for non-expert listeners. The rating scale is composed of 49 bipolar characteristics, and the rating form was derived into eight versions with variations in characteristic order and orientation (e.g., cool-warm, warm-cool). After analysis, four orthogonal factors are found to carry 88% of the total rating variance and are interpreted as: clarity, roughness, magnitude, and animation. The effect of characteristic order and orientation was not significant, but speaker order does exhibit some influences on four characteristics: loud-soft, large-small, soft-hard, and repeated-varied. Gelfer (1988) [65] studies the agreement among expert and non-expert listeners over a developed and proposed bipolar rating scales of 17 items (e.g., low-high, loud-soft). Gallardo and Weiss (2018) [62] propose Nautilus Speaker Characterization (NSC), a corpus of 300 German speakers collected, labeled with interpersonal characteristics and naïve descriptors. Resulting most informative dimensions are provided for males and females. Weiss and Estival (2018) [232] also investigated the assessment of vocal perceptual dimensions of non-expert listeners. Direct comparison of three voices is performed: the listener chooses two similar voices and advises their perceived similar characteristics. The rater also reports its perceived dissimilar characteristic with the third voice. After analysis of similarities and dissimilarities following the raters, perceptual dimensions are proposed, e.g., calmness-activity, factual-emotional, maturity-immaturity. One dimension remained unnamed, too hard to be interpreted. It is worth noting that dimensions were found to be different depending on the language (German/English) and the speaker gender.

Besides finding the most informative characteristics of voice, the inferred speaker state and traits from voice are also relevant to investigate. Kramer (1963) [123] reviews perceptual assessments from voice of physical characteristics (age, height, photographs, body description), aptitudes and interests (intelligence, leadership, political preferences), and personality traits (dominance, intro-extroversion, sociability). Results are reported as mainly mitigated, and the significant interrater agreement but low accuracy of characteristic in-

ference points out the prominence of stereotypes in judgments. Scherer (1978) [197] also examines personality inference from voice quality. Extroversion is found to be correlated with voice energy cues such as vocal effort and dynamic range. Other characteristics could not be accurately inferred from voice quality. Krauss et al. (2002) [124] study listeners' capacity to infer anatomical traits from voice: age, height, weight, and the guessed corresponding photo from a pair of photos.

Altogether, the definition of a consensual list of perceptual characteristics to completely (but not redundantly) describe a speaker's state, traits, and intents remains an elusive effort. Attempts towards such a purpose may leverage Poyatos (1991)[178]'s description of paralinguistic aspects through a list of 10 qualifiers: breathing, laryngeal, esophageal, pharyngeal, velopharyngeal, lingual, labial, mandibular, articulatory, and tension. Corresponding types of perceived characteristics are also depicted. Schultz (2007) [202] proposes a list of speaker characteristics relevant for human-machine vocal interactions and a taxonomy distinguishing physiological from psychological aspects.

Objective and subjective descriptors intricacy

A great deal of effort has been made to describe speech using objective (articulatory, acoustic) and subjective (perceptual) cues. However, the relationship between them is still opaque [80]. Acoustic-perception and physiology-perception correlations have been explored but remain hard to understand and interpret [125]. Moreover, Kent (1996) [106] points out that a listener might miss some cues that are differentiable acoustically (auditory illusion) or, conversely, might hear non-existing speech variations (verbal transformation), further complicating the endeavor. A substantial list of voice qualifiers with their acoustic correlates is provided by Memon (2020) [158], which definitely illustrates the entanglement between perceptual descriptors and acoustic cues.

Hopefully, one may expect recent advances in machine learning methods enabling the automatic learning of acoustic-perception mapping. Obin et al. (2014) [163] and Obin and Roebel (2016) [162] investigated this clue by designing an automatic voice casting system based on a multi-label classification system to learn voice signatures and find perceptually similar acted voices. Subjective classes are defined, assumed to represent the perceptual clues in voice similarity assessment, and the probabilities inferred by distinct classifiers, one for each class, are concatenated to form a vocal signature. The learned paralinguistic space is successfully leveraged to find similar voices among a database of actors, and outperforms speaker verification-based representations. Dealing with percep-

tual characteristics seems feasible for speech synthesis. The following Section 1.2 will review the multiple approaches to generate speech, in order to portray how non-verbal characteristics are controlled by State-of-the-art methods.

1.2 Speech synthesis

Speech is the most natural, common, and efficient means of communication used by humans to express their ideas and thoughts. It is less formal than text, but definitely more personal, persuasive, and impactful. The synthesis of speech thus appears as a praiseworthy goal. Vocalization of written content improves accessibility for people with visual impairments, and inclusivity for persons with speech disorders. Speech synthesis can also be used for entertainment purposes, e.g., to create unique and deep voices for characters or bring back to life a famous personality. Moreover, speech generation leads to more engaging human-machine interactions through conversational agents, personal assistants, or chatbots.

But speech is also human’s most complex means of expression, as far more information than the linguistic content can be conveyed. Even humans sometimes hardly understand unspoken intents. A speaker may also disclose unintentionally his emotional or psychological state. As described in Section 1.1, speech carries a lot of information, of multiple natures, with different time scale influences. It is therefore a true challenge to successfully and independently handle all aspects during speech synthesis.

Thanks to recent advances in artificial intelligence, the intelligibility and quality of speech synthesis have achieved near-natural performances. The huge modeling capacity of Deep Neural Networks (DNNs) has enabled the assimilation of voice acoustic characteristics and variations pertaining to a textual transcription, and the generation of the corresponding utterance. The rise of self-supervised learning paradigm, and end-to-end model architectures, allow the use of vast amounts of unlabeled data and negate the painful crafting of efficient features (linguistic, acoustic). Therefore, verbal content is acknowledged to be synthesized in a near-natural quality.

Despite the accelerating progress of artificial intelligence, truly natural interaction with machines has not been achieved yet. While computers speak with a great naturalness, they still struggle to properly adapt non-verbal attributes. To address this issue, Voice Conversion (VC) systems rely on a reference utterance from which the extralinguistic (e.g., speaker identity) or paralinguistic (e.g., prosody, emotion) contents are extracted

and adjusted with the linguistic content of a target utterance. Apart from this, some methods do allow the finer-grained control over some aspects, but still in a limited range.

All in all, the achievement of speech synthesis is essentially a matter of information transmission. As stated by Claude Shannon:

“Information is the resolution of uncertainty”

Accordingly, the aim of TTS is to properly transmit text information to the output speech. But as described in Section 1.1, speech conveys much more information than its linguistic content. Along with some unpredictable noise, a same sentence has an infinite number of possible realizations, an issue referred to as the one-to-many issue, or lack of invariance [4]. The main purpose of TTS systems is thus twofold: to ensure the transmission of input information (text, speaker identity, prosody), and to deal with the remaining uncertainty to produce speech, as natural as possible.

To cover the synthesis of the various information in speech, Subsection 1.2.1 describes statistical parametric speech synthesis models. Subsection 1.2.2 further introduces how neural networks achieve near-natural text-based speech synthesis. Finally, Subsection 1.2.3 depicts how non-verbal information is handled, with VC or finer control over voice attributes.

1.2.1 Statistical parametric TTS

At the very least, linguistic information is required to deduce the short-time variations encoding the phoneme sequence in the resulting waveform. Among the early approaches in TTS, concatenative speech synthesis systems rely on a database of prerecorded pieces of speech, and analyze a text input to concatenate the suited sequence of units. The pieces of speech might be of different sizes (e.g., phones², diphones, words, sentences), which affects the synthesis procedure: large units lead to natural synthesis but provide limited flexibility, and small units are more flexible but lead to degraded naturalness [215]. One of the most common size of unit is diphone. Diphones [45] are units joining the middles of two phonemes and have the advantage of properly conveying phonemic coarticulation phenomena and allophone variations following the context. All existing diphones must be recorded to form a diphone inventory, and signal processing algorithms must be applied to ensure smooth transitions between units and control the resulting prosody, i.e., pitch and

2. The phonemes, i.e., abstract linguistic units, are distinguished from phones, i.e., their actual realization in a sound wave.

duration (PSOLA [24], TD-PSOLA, and FD-PSOLA [159]). However, storing a single realization of each diphone cannot cover all contextual variabilities, and pitch and duration manipulations are not sufficient to generate those variants. Resorting to signal processing techniques is also prone to signal degradation and distortion. It sounds hence natural to extend the unit database to allow non-uniform-sized units [192] to be considered at a time, e.g., sentences, words, syllables, or phonemes. By leveraging a large amount of annotated data, each phoneme can find multiple corresponding units. Unit selection [96] methodology provides an optimal combination of (non-uniform) units, which jointly minimizes the distance with the target diphones (target cost) and the acoustic mismatch between concatenated units (concatenation cost). Paralinguistic features may also be annotated or predicted (stress, Part of Speech (POS) tagging), which can be leveraged to help during the selection process [217].

Concatenative speech synthesis, however, exhibits significant drawbacks. Such systems are usually limited to single-speaker recordings, and one may build a whole new units database to generate speech with another voice. Concatenative synthesis is recognized for its lack of flexibility and difficulty in adjusting to different speaking styles and producing new or unusual sounds.

Statistical Parametric Speech Synthesis (SPSS) systems address these issues: they strongly rely on an intermediate information representation: acoustic features. Linguistic features (phoneme sequence, POS tagging) are extracted in the same way as concatenative approach, but an acoustic model is employed to compute acoustic features. A third component, a vocoder, then predicts the resulting waveform to be heard. Until a decade ago, the most commonly used approach for such a purpose was to combine a *Hidden Markov Model* (HMM) with a decision tree. The HMM models the acoustic features sequencing (states) and their duration (transitions) [220, 155]. Duration is proposed to be modeled with multidimensional Gaussian distributions, clustered with a decision tree [238] based on linguistic characteristics and contextual information (e.g., surrounding phonemes or POS). This approach enhances the naturalness and controllability of the speaking rate. Acoustic features might also be modeled through Gaussian distributions and clustered with decision trees [239] in order to reduce the number of states, and provide flexibility in that new sounds and contexts can be addressed if properly clustered by the decision tree. In this context, the generated acoustic feature sequences are usually mel-cepstral coefficients, synthesized with Mel-Log Spectrum Approximation (MLSA) filter [98]. Hybrid approaches jointly leveraging unit selection and HMMs/decision trees are also worth to

mention [243].

This parametric setting greatly enhances the coherence of the synthesis, and negates the need to store a lot of prerecorded speech fragments. It is, however, prone to noise and artifacts. HMMs and decision trees may struggle to model context dependencies, and the averaging of multiple HMM states leads to the over-smoothing of generated acoustic features [113]. To remedy those issues, components were progressively replaced by DNNs.

1.2.2 Neural-based synthesis

Among the major advancements that allowed DNNs predominance for speech synthesis, Recurrent Neural Networks (RNNs) can model the underlying dynamic patterns within sequential data, Long-Short-Term-Memory (LSTM) [88] and *Gated Recurrent Unit* (GRU) [31] being the most used recurrent models. Sequence-to-sequence (seq2seq) [214] paradigm enables the generation of a sequence from another sequence, without making assumptions about the input sequence length. From image processing efforts have emerged Convolutional Neural Networks (CNNs) [128], which efficiently models neighborhood patterns, which was found to be a relevant way to deal with speech. Attention mechanism [10] is also a key concept, which introduces the notion of weighted context when considering an element of a sequence. It has been further extended to Transformer [224], more complex and non-autoregressive (hence parallelizable and faster to compute, but requires more memory and computing resources). Generative Adversarial Network (GAN) [69] is another powerful generative model, comprising two components: a generator and a discriminator. The former is trained to generate fake data, and the latter is optimized to discriminate real data from generated fake ones. This adversarial scheme leads to high-quality data generation. Variational Auto-Encoder [115] is also a well-used generative model, able to approximate with an inference model the distribution of a real dataset through a latent space with a simple (Gaussian) prior distribution, and generate highly realistic data with a neural decoder by sampling from the learned distribution. A last generative model worth to mention is Inverse Autoregressive Flow [117], consisting of a series of invertible transformations modeled by autoregressive neural networks over a simple distribution to build a more complex one.

Acoustic feature generation supplied by HMM-based models, as depicted in Subsection 1.2.1, was demonstrated to be efficiently handled with DNNs [242, 77] with similar complexity. Better performances are achieved for multi-speaker speech synthesis with a single encoding module [57]. Considering contextual information to leverage the sequential

nature of speech is demonstrated to further improve naturalness, with bidirectional [58] or unidirectional [241] LSTMs.

Later on, DNNs were further leveraged to entirely build synthesis systems. Following the seq-2-seq paradigm, Tacotron [230] and Tacotron 2 [205] are popular acoustic models, directly predicting (mel-)spectrogram from characters, and integrating an attention-based alignment procedure. DeepVoice [6] relies on Connectionist Temporal Classification (CTC) [71] to directly predict phoneme labels from unsegmented speech waveforms during training in order to optimize duration and pitch prediction components. DeepVoice 2 [5] extends DeepVoice 1 and Tacotron to multi-speaker synthesis by conditioning duration, pitch prediction and vocoder components with speaker embeddings. DeepVoice 3 [176] is fully-convolutional and attention-based rather than RNN-based, hence parallelizable and faster for training and inference. To further speed up computation, FastSpeech [186] is only composed of non-autoregressive procedures, achieving up to 38x faster waveform generation than autoregressive models. It introduces Feed-Forward Transformer (FFT) block, which generates a mel-spectrogram from a phoneme sequence based on multi-head attention and 1D convolution (Conv1D), and has a controllable length regulator, i.e., the voice speech can be modified. FastSpeech, however, is trained with sequence-level knowledge distillation [112], i.e., the mel-spectrogram prediction and the duration predictor are optimized regarding a pre-trained autoregressive teacher model (Transformer TTS [139]) outputs. But relying on a teacher model makes the training a complicated procedure, and some information might be lost by the simplified teacher outputs. Thus, FastSpeech 2 [185] alleviates such issues by directly using ground-truth data for optimization.

Once the acoustic features (i.e., mel-spectrogram) are generated from linguistic features (i.e., character or phoneme sequence), neural vocoder models are employed to convert them into audible waveforms. Relying on dilated causal convolutions to deal with temporal context dependencies and get a wide receptive field, WaveNet [165] stands among the first neural-based vocoders. WaveRNN [102] introduces several acceleration procedures to reach real-time inference speed with a dual softmax layer, weight sparsification and sequence subscaling, without degrading performances. However, the autoregressive scheme involves slow inference speed. Hence, Parallel WaveNet [164] tries to speed up inference while preserving the same performances, by using a pre-trained WaveNet as a teacher to learn through distillation an IAF-based student architecture. Furthermore, GANs were successfully leveraged to speed up inference while synthesizing high-quality speech. WaveGAN [47] stands as the first approach to generate speech with an unsu-

pervised GAN architecture. It uses 1D transposed convolution and phase shuffling operations to properly model speech variations, and can generate realistic sounds (speech, birds, piano) in a fully parallel way. Other GAN-based vocoders were further developed and widely used, as Parallel WaveGAN [236], based on a non-autoregressive version of WaveNet (non-causal convolutions) as a generator and a multi-resolution *Short Time Fourier Transform* (STFT) analysis to model the time-frequency patterns of speech and deal with the time/frequency resolution trade-off of STFT. HiFi-GAN [121] is also a very popular vocoder, with convolution-based generator and discriminators using multiple dilation rates, periods and scales to learn speech patterns at several resolutions.

In summary, substantial efforts were conducted to build two-staged synthesis models: acoustic models to convert linguistic to acoustic features, and vocoders to generate waveforms from acoustic features. Furthermore, autoregressive models do not need explicit prediction of the alignment between linguistic and acoustic features, but are, as mentioned, computationally slow. Hence, the described parallel models also require an external module to deduce the alignment. Each component is developed and learned separately, which requires supervision and annotations to train each stage. In order to simplify the training process, and leverage the modeling power of data-driven learned hidden representations, fully end-to-end models have been of great interest in recent studies to directly generate waveforms from character and phoneme sequences. For instance, ClariNet [175], similarly to Parallel WaveNet, learns an IAF-based student model with a pre-trained teacher WaveNet, but which is directly conditioned on hidden states, leading to better performances than with acoustic features. However, alignment is still done autoregressively. Hence EATS [48] is a feed-forward GAN-based end-to-end model, trained from scratch to predict in parallel its own hidden features, their alignment and the output speech waveform from text or phoneme sequences. Alignment is learned with the help of a soft Dynamic Time Warping (DTW) loss, to mimic human’s variations of speaking rate over utterances. VITS [110] is also employing an adversarial training scheme, with a Variational Autoencoder (VAE) conditioned on input phoneme sequence and predicted alignment, followed by a flow-based model to transform the simple Gaussian prior to a more complex distribution. The durations are stochastically predicted with a Monotonic Alignment Search (MAS) [111] algorithm and a flow-based model. VITS2 [122] further improves naturalness and multi-speaker setting handling with adversarial training of the stochastic duration predictor, a Transformer block added to the normalizing flow models, and a speaker-conditioned text encoder. Finally, EdenTTS [152] implicitly models align-

ment with a guided aligner module, i.e., the scaled-dot attention terms are weighted by an energy matrix which enforces diagonal alignment, to help the training of the duration predictor module. It therefore does not require ground-truth alignment, at the expense of quality degradation.

Altogether, parallel and end-to-end models are faster and less complex, but are hard to properly train, i.e., jointly predicting in parallel and handling the mismatch between character sequence and waveform sample scales [216]. However, TTS models have reached near-natural speech synthesis quality. In a sense, sound waves are successfully generated from linguistic content (i.e., characters/phoneme sequences). The challenge now resides on how to deal with the remaining non-verbal attributes of speech. Multi-speaker TTS systems [230] deal with this concern by injecting paralinguistic and extralinguistic information in the synthesis procedure, for instance with speaker representations extracted from Automatic Speaker Recognition (ASV) models. On the other hand, VC aims to replace the non-verbal information of a reference utterance into a target one, without altering the linguistic information. Essentially, VC is a problem of filtering and replacement, whereas TTS is a matter of information forecasting. However, both approaches tend to converge towards similar architectures, and the sole remaining boundary between TTS and VC is whether text is provided as input to the model or not.

1.2.3 Controlling non-verbal aspects

Broadly speaking, non-verbal speech aspects can be controlled implicitly or explicitly. Implicit control refers to expressive TTS and VC techniques, where target speech recordings are used to extract non-verbal attributes (e.g., speaker identity, prosody, speaking rate) to be injected into a source utterance. On the other hand, attributes can be explicitly controlled in a more fine-grained manner. It is merely referred to as controllable speech synthesis.

A straightforward solution is to consider speaker identity, i.e., extra-linguistic features. This aspect can be addressed in multi-speaker TTS systems by conditioning on speaker embeddings, as DeepVoice 2, which learns its own set of speaker embeddings to be incorporated in several layers of a Tacotron [230] acoustic model. Such methods limit the synthesis to speakers seen during training. Hence, speaker embeddings are usually extracted from pretrained Automatic Speaker Recognition (ASV) models to perform multi-speaker TTS for seen and unseen speakers during training, viz. zero-shot TTS. For instance, Jia et al. (2018) [101] employ d-vectors [223] to be concatenated with text embeddings in a

Tacotron 2 [205] model. Among the very recent advances, VALL-E [226] extracts audio codes from an EnCodec [37] model, to perform SOTA zero-shot TTS with only three seconds of enrollment recording of an unseen voice. To gain control over para-linguistic aspects in TTS, style transfer approaches are introduced, leveraging learned style/prosody embeddings, as GST [229], Skerry-Ryan et al. (2018) [207], or An et al. (2022) [3].

With the advantage of leveraging large amounts of untranscribed speech data, Voice Conversion (VC) aims to convert a given voice to sound like another, without altering the linguistic content. Hsu et al. (2016) [91] leverage a VAE to learn speaker-independent phonetic representations from a source utterance and a one-hot vector to specify the targeted speaker identity. To improve the preservation of the content targeted, Saito et al. (2018) [193] leverage Phonetic Posteriorgrams (PPGs) [212], extracted from an Automatic Speech Recognition (ASR) model. Towards conversion from and to unseen speakers during training, viz. any-to-any or zero-shot VC, AutoVC [179] advocates the importance of tuning the information bottleneck by finding the suited content embedding size, to reduce the leakage of speaker-related information, while preserving content-related information. VQMIVC [227] further improves the factorization of information pertaining to content, speaker, and also pitch variations through Mutual Information (MI) minimization and vector quantization.

As one may notice, throughout the very recent advancements, the boundary between TTS and VC approaches is becoming increasingly blurred, especially in zero-shot TTS and any-to-any VC. Both tasks tend to be accomplished by similar model architectures, leading to an unification of the methodologies, as pointed out by Zhang et al. (2019) [245]. Roughly speaking, both tasks lean on the factorization of the various sources of variations, mainly being restricted to linguistic content, speaker identity, and style/prosody/expressivity, embodied by deep embeddings learned by proper encoders. Hence, the distinction between TTS and VC only lies in the modality from which linguistic content comes from: text for TTS or source utterance for VC. According to this idea, VITS [110] can perform VC through inversion of the flow-based decoder. More recently, YourTTS [22] extends VITS to jointly achieve zero-shot multi-speaker TTS and any-to-any VC with an H/ASP [82] speaker encoder.

While the growing and sharp interest in TTS and VC leads to ever-improving naturalness and speaker similarity in synthesized speech, the non-verbal facets are addressed in an implicit manner. Therefore, encoded features are informative, but unstructured and not interpretable. To tackle this lack of transparency, models able to explicitly control

such aspects are proposed. FastSpeech 2 [185] can control, in addition to speaker identity and duration in FastSpeech [186], pitch and energy, through predictor modules, in order to better model speech variations and deal with the one-to-many issue. During training, reference duration, pitch and energy values are extracted from ground-truth data, and are made controllable during inference. Based on the same endeavor, SpeechSplit [180] and SpeechSplit 2.0 [23] are VC models, which unsupervisedly factorize content, rhythm and pitch by means of signal processing operations. Raitio et al. (2020) [182] propose a mono-speaker TTS model which explicitly and supervisedly models four prosodic features: pitch, phone duration, energy and spectral tilt, to control them in synthesis. Raitio et al. (2022) [181] extended their work by adding utterance-wise prosodic features, in a hierarchical way, to provide global and local control over para-linguistic aspects. Nansy++ [30] separately encodes pitch, periodic amplitude, aperiodic amplitude, timbre and linguistic features to supply a self-supervised framework able to perform TTS, VC and voice design through an ECAPA-TDNN [42]-based age and gender extraction. Finally, ControlVC [25] enables the temporal control of rhythm by a speed curve and Time-Domain Pitch-Synchronous Overlap and Add (TD-PSOLA), and pitch by pitch contour manipulation.

Despite the very promising results, the explicit control of non-verbal attributes through supervised or self-supervised TTS/VC frameworks remains sophisticated and restricted to raw acoustic features, far from one’s intuitive representation of speech. Speaker identity is not yet further decomposed beyond gender and age, but extralinguistic attributes might be way more refined, as discussed in Subsection 1.1.1. Carried by recent breakthroughs in text-guided generation of text (e.g., GPT-3 [17]) and images (e.g., DALL-E-2 [183]), controlling speech non-verbal attributes from textual description is an emerging trend. It supplies a convenient way to intuitively describe the desired voice, with one’s own words. At the outset of this research direction, PromptTTS [72] employs a style encoder containing a BERT [43] model, pre-trained to predict gender, pitch, rhythm, energy and emotion from a style prompt. Informative style embeddings are thus extracted, to be coupled with the content encoding and synthesize expressive speech. PromptStyle [146] further improves style modeling by guiding the prompt-based style encoder with a speech-based style encoder. Pertaining to a Voice Conversion desideratum, Kuan et al. (2023) [129] leverage an EnCodec [37] in a VALL-E [226]-like fashion to use the textual prompt as style guidance to infer the desired style and emotion in the source utterance.

All in all, non-verbal attributes can be controlled in an implicit way in TTS and VC

by mimicking and replacing in a source utterance the speaker identity, prosody or emotion transmitted by a target utterance, but without deeper knowledge and control over the converted aspect. In this regard, methods are advanced to acquire finer control over paralinguistic and extralinguistic cues in TTS and VC, but are tied to acoustic features, hence necessitating signal processing skills to be manipulated, besides being cumbersome to adjust. Oppositely, text prompt-guided synthesis techniques arising from recent progresses in Large Language Models (LLMs) are auspicious methods towards user-friendly speech customization and voice design procedures. But in the middle, few efforts have been conducted to build intermediate models, with fine-grained control over intuitive speech facets, not limited to knotty acoustic features. Such intermediate approaches are expected to disentangle speech attributes, in a way less restrictive than described controllable TTS and VC models, but far more detailed than text-guided ones. Amongst the few approaches standing in between, FHVAE [92], Capacitron [11], and GMVAE-Tacotron [93] leverage disentanglement learning techniques and hierarchical modeling to self-supervisedly extract and control non-verbal characteristics.

To summarize, implicitly handling non-verbal speech attributes in TTS or VC has the advantage of enabling a close to natural transfer of attributes (e.g., speaker identity, prosody), but do not let one easily control them. Conversely, controllable speech synthesis methods are able to modify fine-grained speech characteristics, but might be tedious to be properly used. The recent trend of prompt-based methods provides easy-to-use controllable synthesis through natural language, but does not let one cautiously shape the produced speech. Overall, designing and training such models presents a challenge that lies at the heart of recent research and industry concerns. A promising hint to automatically model paralinguistic and extralinguistic aspects and gain better control over them in synthesis is disentanglement learning, a growing research field covered in Chapter 2.

DISENTANGLEMENT LEARNING

BACKGROUND

Reality favors symmetry.

Jorge Luis Borges

Among recent deep learning-related investigations, disentanglement has emerged with the very appealing ambition to address the lack of structure, and thus interpretability, within learned neural network representations. While still in its early stages, disentanglement learning forms a full-fledged research area.

In principle, every task is about discerning relevant information within a complex and noisy environment. The more complex the environment and the task, the more diluted the useful patterns among the overwhelming variations and distractions. Empowered by neural networks' growing modeling capacity, artificial intelligence excels in the discovery of such hidden underlying patterns. While supervised learning relies on ground-truth annotations to retrieve the relevant information, unsupervised learning alleviates this constraint by allowing the processing of massive amounts of unlabeled data. Self-supervised learning is a very popular paradigm to train a model without supervision, by deducing the label directly from the raw data, or employing the raw data itself as the label, typically in an autoencoding scheme, i.e., encoding an input data in a compact format, and trying to reconstruct the original input with a decoding stage. Representation learning, which aims to project observations of a given environment in an abstract representation space, greatly benefits from self-supervised approaches, enabling the learning of more agnostic insights, before doing any specific task.

Learning insightful and agnostic representations, while alleviating the resort to annotations, has been a trendy paradigm in recent research efforts. By way of proof, the *International Conference on Learning Representations* (ICLR) has become one of the quickest-growing artificial intelligence conferences since its inception in 2013, with almost

5000 submissions in 2023. It is nowadays an unmissable appointment for deep learning enthusiasts, focused on representation learning advancements. Eminent conferences in machine learning such as *International Conference on Machine Learning* (ICML), *Neural Information Processing Systems* (NeurIPS) or *International Joint Conference on Artificial Intelligence* (IJCAI) also exhibit sessions dedicated to representations. Speech processing also benefits from such investigations, gathered in dedicated sessions in conferences such as INTERSPEECH or *International Conference on Acoustics, Speech, and Signal Processing* (ICASSP).

With such recent advances in mind, Bengio et al. (2014) [13] stated that learned representations are still not able to efficiently organize relevant information about data, and should “identify and disentangle the underlying explanatory factors hidden in the observed milieu of low-level sensory data”. It is hence suggested that a “good representation” should disentangle factors of variations and exhibit the hierarchical structure of explanatory factors. Such a representation may thus provide improved robustness to small data variations and enhanced transfer to multi-task settings, as “general-purpose” relevant information is appropriately disentangled from noise and from each other.

Learning disentangled representations remains a recent undertaking. By way of proof, this paradigm is still lacking a formal definition, hence leaving unclear how to measure the degree of disentanglement. Multiple works are proposing metrics, assessing their own definition and properties, relying on synthetic data, as the ground truth factors still have to be known to measure their disentanglement. It is thus still troublesome to decide which metric to use, and investigations upon real data are still challenging.

Since Bengio et al.’s seminal work [13], one may observe an increasing interest towards this endeavor. Figure 2.1 shows the number of publications with the terms “disentangle”, “disentanglement”, or “disentangling” in their title published each year in some leading conferences in Artificial Intelligence (AI). Conferences about AI advances in general: ICLR, ICML, NeurIPS, and IJCAI, are marked with circles. Speech-related conferences: ICASSP and INTERSPEECH with crosses. As many efforts towards disentanglement are carried by image processing research efforts, computer vision conferences are also reported in Figure 2.1 with triangle markers, namely *International Conference on Computer Vision* (ICCV) and *Computer Vision and Pattern Recognition* (CVPR). Globally, the number of concerned articles has clearly expanded over the years. It is, however, worth to mention that Figure 2.1 only provides a clue of the number of publications concerned with disentanglement, as some may actually study disentangled representations while not explicitly

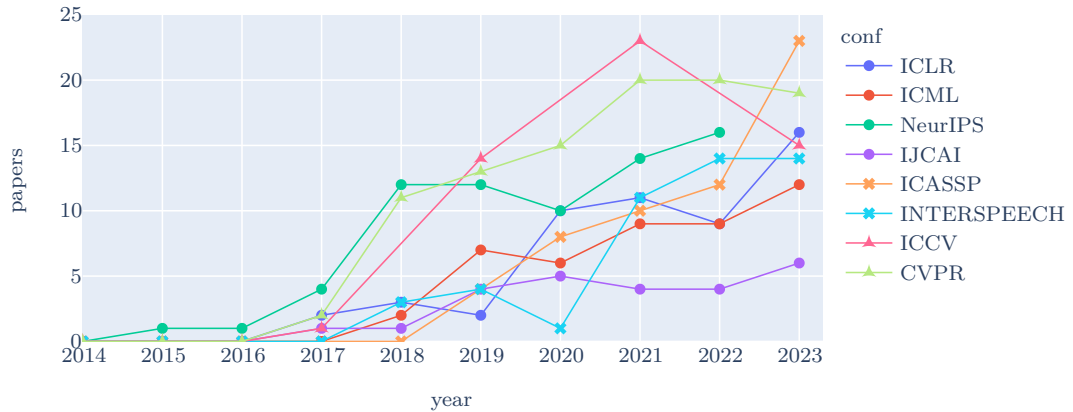


Figure 2.1 – Number of papers with “disentanglement” in title per conference

mentioning this term in their title, and the mentioned conferences are far to exhaustively cover all research efforts.

In order to properly go through disentanglement learning, Section 2.1 draws the global picture of what disentanglement is and which benefits one may expect from it. Thereafter, Section 2.2 overviews the various deep learning approaches leveraged to disentangle. Section 2.3 introduces the synthetic corpora typically used in disentanglement studies. Finally, Section 2.4 describes some metrics proposed to objectively measure the degree of disentanglement, and conclusions are drawn in Section 2.5 about disentanglement learning and how it can help speech understanding and synthesis.

2.1 Disentangling disentanglement

As previously mentioned, learning disentangled representation is a recent endeavor, explicitly incepted by Bengio et al. (2014) [13]. They disclose the limitations of traditional learned representations and highlight the advantages of disentanglement principles. However, they leave open the issues of formally defining disentanglement and developing models that can effectively learn to disentangle.

Subsection 2.1.1 outlines the concept of representation learning and its shortcomings, hopefully filled with disentanglement principles. Subsection 2.1.2 attempts to outline the contours of which properties should be expected from a disentangled representation. Sub-

section 2.1.3 discloses the ambiguity often encountered between what can be referred to as “information factorization” and disentanglement as it is actually considered in this manuscript.

2.1.1 Pink elephant

Prior insight into disentangled representation involves acknowledging the principles of conventional neural-based representations. Let $X = \{x^i\}_{i \in \mathbb{N}}$ a set of observed data, lying in a high-dimensional space \mathcal{X} . **Observations** X are populating \mathcal{X} following a distribution $p(x)$, which is assumed to be governed by a generative process which involves a set of **generative factors** $\mathbf{f} = \{f_i\}_{i \in \mathbb{N}}$. Neural representation learning therefore aims to build an abstract **latent space** $\mathbf{z} = \{z_i\}_{i \in \mathbb{N}}$ through non-linear transformations, capturing as much information as possible about factors. Therefore, a model unsupervisedly learned following a representation learning paradigm is likely to produce an informative latent space, to be leveraged by downstream tasks. Let $\mathbf{y} = \{y_i\}_{i \in \mathbb{N}}$ be the generic notation of the target discrete or continuous label, or the target sample to generate, pertaining to the concerned downstream task.

Furthermore, a sense of **orthogonality** is assumed between generative factors, i.e., observations generated under values taken by a factor f_i should generalize under all settings of other factors $\{f_{i'}\}_{i' \neq i}$. This does not require statistical independence between factors, especially within a finite set of observations. They are expected to influence disjoint properties, potentially correlated but not causally tied. One might rightly not expect to see a pink elephant, but an image representation model properly learned should generalize faced with such an input, by decoupling the abstract concepts of color and type of animal.

Conventional representation learning is demonstrated to be effective for various speech-related downstream tasks. By way of evidence, SUPERB [237] and LeBenchmark [56] are popular benchmarking frameworks proposed to evaluate and compare speech self-supervised representation models in achieving multiple tasks, e.g., *Automatic Speech Recognition* (ASR), *Automatic Speaker Verification* (ASV) or *Automatic Emotion Recognition* (AER), which demonstrate the widespread interest in such methods, e.g., Wav2vec [198, 9], WavLM [27] or HuBERT [94]. Furthermore, some studies are analyzing the internal structure of speech representation models [168, 170, 169, 140], which assessed how acoustic, linguistic, and paralinguistic information are differently captured depending on layer depth in a self-supervised pre-trained model. But the bare structure of the representations still remains mostly uncharted: are the learned concepts explicitly exhibited, or are they

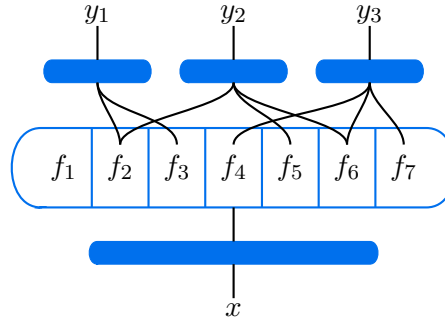


Figure 2.2 – Shared statistical strength between tasks (from Bengio et al. (2014) [13])

diluted within intricate nonlinear relations between latent dimensions?

It is hence argued in Bengio et al.’s initiative work [13] that representation learning may greatly benefit from a more structured knowledge organization: a learning algorithm able to disentangle factors of variations f in its hidden representation z might be efficiently leveraged for multiple downstream tasks, as each task may be fulfilled by subsets of explanatory factors, and each factor might be relevant for several tasks. Coming back to the pink elephant in the room, a disentangled representation is believed to be a convenient support for an image classifier to predict an animal or its color, or for a generative model of images to properly generate any animal in any color, if only one knows how to sample z accordingly.

More broadly, the aim of disentangled representation learning is to advocate the “sharing of statistical strength”, in one hand between the unsupervised phase and underlying supervised tasks by easing the discarding of irrelevant information, and in the other hand across tasks in a multi-task setting. Borrowed from Bengio et al. (2014) [13], Figure 2.2 illustrates such phenomena, with explanatory factors $\{f_i\}_{i \in \{1, \dots, 7\}}$ disentangled in learned representation from observed input x , and used by tasks y_1, y_2 and y_3 . Overlapping subsets of factors used by each task embodies the statistical strength shared among tasks, which is believed to help generalization of learned abstract representation space.

For the aforementioned reasons, representation learning approaches are lacking knowledge organization on purpose in learned latent spaces. Based on this insight, some efforts have been conducted to disentangle speech representations [93, 11, 211, 74], targeting interpretability in learned representations and explicit controllability of speech attributes in synthesis. While not providing a formal definition of a disentangled representation, Bengio et al. more generally provide some hints of expected properties required to reach such task-agnostic representations.

2.1.2 Discovering world’s hidden mechanisms

From the properties hinted by Bengio et al. (2014) [13], what they suggest to be a “good” representation should exhibit the following properties:

Smoothness : close observations should lead to close points in the representation space, i.e., $x_1 \approx x_2 \Rightarrow h(x_1) \approx h(x_2)$ given a learned function h . Smoothness emphasizes robustness of representations, e.g., against adversarial examples [70].

Sparsity : only relevant information is extracted from observations. In other words, leveraging the high modeling capacity of deep neural networks allows the learning of abstract concepts, discarding irrelevant and noisy information, i.e., learned representations are insensitive to small variations of input.

Distributed : multiple features are learned, and can be independently varied, i.e., they are not mutually exclusive. Such representations are sufficiently expressive to model similarities across concepts and generalize to configurations unseen during training.

Disentanglement : as many sources of variations as possible should be disentangled. With the assumption that observed data are generated from complex interactions of hidden source factors, subsidiary tasks are likely to be tied to those explanatory factors, either directly or through simple transformations and combinations. Since target tasks, and thus relevant factors, are still unknown when learning a representation space, the challenge remains in the definition of the prior belief on the nature of useful factors to extract.

Retrospectively, it remains unclear how to define formally what generative factors are. Bengio et al. describe them as the sources of variations whose states and interactions explain a given environment. Goodfellow et al. [68] are portraying factors as separate unobservable sources of influence affecting observable quantities.

Higgins et al. (2018) [85] are following a **symmetry**-based approach to provide a definition of disentangled representation, where underlying factors to disentangle within an environment are properties which might be altered under interactions, while leaving other properties unchanged, hence said to be invariant under symmetric transformations e.g. an object can be characterized by its shape, color, weight, spatial positioning, and can be moved from one place to another without altering its shape, color, or weight. It is worth noting the connection with distributed representation principle, which reflects by essence the real-world symmetry transformations: different concepts can be described through a shared set of (symmetric) abstract properties, which can be independently controlled

without affecting each other. While the symmetrical nature of the universe [148] is out of the scope of this manuscript, one may intuit that apprehensible mechanisms of our world are those ruled by symmetries, which form the basis of animals and humans reasoning about how they perceive and interact with their surroundings, i.e., comprehension of world mechanics are instinctively tied to properties, among which any can be altered while leaving others unchanged. In a sense, disentangle the properties pertaining to a set of observations amounts to find those symmetries.

Goodfellow et al. [68] are also depicting disentanglement through the prism of **causality** [173], a point further explored by Suter et al. (2019) [213]. Learning algorithms are expected to separate on different representation dimensions the various data causes of variations. Factors of variations are hence assumed to be causally independent, i.e., independently affecting observations and not exhibiting causal effects between each other, which follows the distributed representation property. Although it is hard to prevent mutual causations between underlying factors, one may resort to a set of confounders $\mathbf{c} = \{c_i\}_{i \in \mathbb{N}}$ to maintain causal independence while defining an environment. More explicitly, explanatory factors are generally mutually dependent, i.e., $\forall i, j \in \mathbb{N}^2$ such that $i \neq j$, the observation of f_j influences the value of f_i :

$$p(f_i | f_j) \neq p(f_i) \quad \text{or} \quad f_i \not\perp\!\!\!\perp f_j, \quad (2.1)$$

with $f_i \not\perp\!\!\!\perp f_j$ denoting the statistical independence between f_i and f_j . But factors become mutually independent when conditioned on confounders i.e.:

$$p(f_i | f_j, \mathbf{c}) = p(f_i | \mathbf{c}) \quad \text{or} \quad f_i \perp\!\!\!\perp f_j | \mathbf{c}. \quad (2.2)$$

Furthermore, the observation of a data sample $\mathbf{x} = x$ does render factors dependent i.e.:

$$p(f_i | f_j, x) \neq p(f_i | \mathbf{x} = x) \quad \text{or} \quad f_i \not\perp\!\!\!\perp f_j | \mathbf{x} = x. \quad (2.3)$$

More importantly, the intervention on f_j does not impact f_i i.e.:

$$p(f_i | \text{do}(f_j = f)) = p(f_i) \quad \text{or} \quad f_i \perp\!\!\!\perp \text{do}(f_j = f), \quad (2.4)$$

the **do** operator differentiating the **observation** of a realization (i.e., $f = f$) and the **intervention** on the system towards a specific state (i.e., $\text{do}(f = f)$). The action on the system $\text{do}(f = f)$ has the effect of preventing (causal) inference of upstream variables,

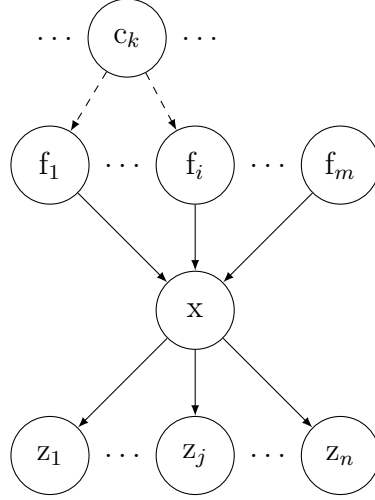


Figure 2.3 – Data generation process

while seeing that $f = f$ lets one infer the probability of the causes of what is observed. To borrow Pearl’s [173] illustration, observing that the sprinkler is on provides a hint on the current season, while manually turning it on prevents the forecast of the ongoing season. This hence leads to the distinction between (2.1) and (2.4). Figure 2.3 illustrates with a Bayesian network the causal generation processes of x by factors $\{f_i\}_{i \in \{1, \dots, m\}}$ under confounders $\{c_i\}_{i \in \mathbb{N}}$, with the predicted latent variables $\{z_i\}_{i \in \{1, \dots, n\}}$. It is believed that learning such causal mechanisms ensures the disentanglement of agnostic explanatory factors, since, according to Goodfellow et al. [68], “the laws of the universe are constant”, which is echoing with the aforementioned symmetry principle.

In practice, learning disentangled representations is usually reduced to the fulfillment of more rudimentary properties. While investigating related studies and efforts towards the definition of disentanglement metrics [84, 109, 55, 188], one may notice that the achievement of disentanglement can be summarized into the assessment of 3 criteria [21], which are henceforth designated as: **modularity**, **completeness** and **informativeness**. Modularity (sometimes misleadingly referred to as disentanglement) refers to how much each latent dimension is informative about only 1 factor, i.e., to what degree each latent corresponds to only 1 factor. Complementarily, completeness (sometimes referred to as compactness) assesses how much each factor is explained by only 1 latent dimension, i.e., to what degree each factor finds only 1 corresponding factor in the representation space. Informativeness (sometimes called explicitness) is finally ensuring that factors are effectively explained by latent space, i.e., factor states can be predicted from latent values. Fig-

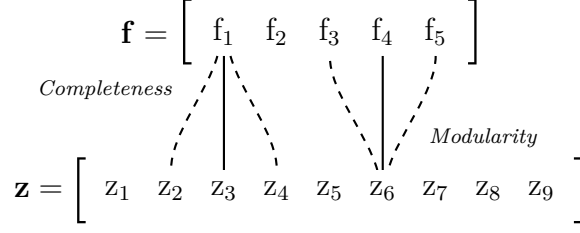


Figure 2.4 – Modularity and completeness criteria
(from Eastwood and Williams (2018) [55])

Figure 2.4 illustrates modularity and completeness criteria for given factors $\mathbf{f} = \{f_i\}_{i \in \{1, \dots, 5\}}$ and a learned latent representation $\mathbf{z} = \{z_i\}_{i \in \{1, \dots, 9\}}$, plain and dashed lines materializing latent informativeness/factor explainability relationships. Latent z_6 is acknowledged to exhibit high modularity if it is only informative about factor f_4 , i.e., dashed relationships with f_3 and f_5 are negligible relative to the plain relationship with f_4 . Similarly, factor f_1 manifests good completeness if dashed relations with z_2 and z_4 are weak compared to the plain relationship with z_3 . Altogether, achieving disentanglement comes down to seeking a bijective mapping between latents \mathbf{z} and factors \mathbf{f} .

It is also worth highlighting the controversy about the completeness criterion. Ridge-way and Mozer (2018) [188] raise the very relevant point that enforcing completeness is likely to be counterproductive in the discovery of explanatory factors. For instance, a factor characterizing any rotation might be explained through its angle $\theta \in [0^\circ, 360^\circ]$, which expresses a better completeness than using its sin and cos, while being just as much descriptive. More broadly, one may find some interest in modeling a complex factor in multiple latent directions.

Lastly, the unsupervised disentanglement of representations is a promising endeavor, that advocates leveraging data itself to learn underlying factors in a general-purpose first stage, before handling downstream supervised tasks in a subsequent stage [13]. Such a semi-supervised scheme is however an elusive undertaking: following the prominent investigations about unsupervised learning of disentangled representations from Locatello et al. (2019) [149], one should not solely rely on great amounts of unlabeled data to extract task-agnostic features. Defining the proper inductive bias is a mandatory phase, in order to infuse prior knowledge about relevant underlying factors to disentangle and incoming downstream tasks. All in all, the choices of neural network architecture, hyperparameters, and regularization strategies are the exploitable inductive biases to adjust the definition of salient sources of variations depending on the context [68]. Learning disentangled repre-

sensation should indeed follow a data-driven procedure, but not without carefully defining prior knowledge to implement through inductive bias.

Despite the absence of a consensual and formal definition of disentanglement at hand, one can find substantial hints in the aforementioned references [13, 68, 85, 149], which are converging to the intent of learning disentangled representations able to discern the explanatory features, symmetrical properties, and causal factors of variation hidden within a set of observations into distinct latent space directions. With these clues in mind, recent years have witnessed a great deal of investigation towards such interests (Figure 2.1). Although, for the sake of clarity, a distinction between the notion of disentanglement in use throughout this manuscript on the one hand, and the one used in some other published articles on the other hand, seems necessary and is advanced hereafter in Subsection 2.1.3.

2.1.3 Distinction from information factorization

Among the publications displaying an interest in disentanglement, a great part of them do not actually share the same concepts followed in this manuscript, being the separation of factors of variations in individual latent dimensions. This distinct paradigm can be referred to as **information factorization**, where learned hidden features are explicitly separated into multiple representations i.e. extracted information is factorized on purpose into distinct and separated latent spaces. For instance, some works factorize (a.k.a., “disentangling”) linguistic and speaker information, by learning two separated representations, to perform Voice Conversion [244]. Information factorization is definitely “disentanglement” in its broad sense, but not as we interpret it in this manuscript. However, information factorization is designated as “disentanglement” in a lot of studies, including many of those considered in Figure 2.1. This is hence reducing the number of publications specifically dealing with generative factors disentanglement, but it remains interesting to note that the notion of disentanglement in its broadest sense (i.e., building more structured representations, in any way, regarding underlying information) is gaining attention over the years. Henceforth, let **disentanglement** refer to the separation of generative factors into distinct latent dimensions, and **factorization** to the separation of information towards distinct representations.

Figure 2.5 illustrates instances of information factorization implementations. Factorization can be performed through a model-driven approach, for instance, with multiple encoders to learn separate representations, i.e., factorized latent spaces, as portrayed in Figure 2.5a: two encoders are separately learning their own representation of input data

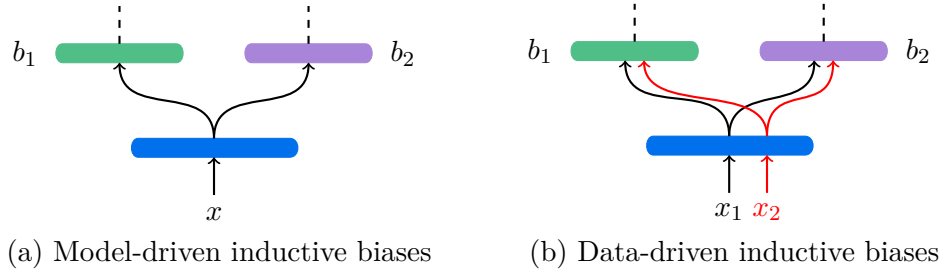


Figure 2.5 – Information factorization illustrations

x , while being jointly optimized. To efficiently discriminate the information captured by each latent space, each side of the bifurcation is characterized by different inductive biases b_1 and b_2 e.g., architecture, regularization, loss function, and so on. For example, in Figure 2.5a, if b_1 is a speaker recognition task loss, and b_2 a transcription task loss, linguistic and extralinguistic information are factorized into their respective representation.

One may also rely on known common or different properties between data samples to force them to be close or far from each other in learned latent spaces. Such a data-driven scheme is depicted in Figure 2.5b, where two different inputs x_1 and x_2 are fed into the model. If properties are known to be different or similar (e.g. gender, speaker identity, emotion), the relative distance or shape of their respective representations can be enforced to meet desired behaviors regarding the known characteristics. Using different properties to compare both representations, i.e., different data-driven inductive biases b_1 and b_2 , can lead to semantically well-separated latent spaces. For instance, knowing that data samples x_1 and x_2 are coming from the same (resp. different) speaker, but with different (resp. same) emotions, one can learn speaker and emotion embeddings by setting b_1 to minimize (resp. maximize) the distance between x_1 and x_2 representations, and letting b_2 maximize (resp. minimize) the distance between x_1 and x_2 representations. Contrastive learning [107] stands among such principled, data-driven, inductive biases. Overall, information factorization leverages prior knowledge about data to influence and explicitly factorize information flow, to help downstream tasks focus on relevant cues.

Amongst the instances of efforts leveraging information factorization for speech processing, Yuan et al. (2021) [240] are following the same principle as in Figure 2.5a by learning a content encoder and a style encoder, to perform style transfer. Polyak et al. (2021) [177] employ three self-supervised encoders to factorize content, F0, and speaker identity in distinct representations, enabling controllable speech synthesis via F0 manipulation and VC. StyleVC [50] also uses 3 unsupervised encoders to separate speaker,

style, and content information, to jointly perform VC and expressive VC. Williams et al. (2021) [233] exploit *gradient reversal* [63] to discard supervisedly speaker information and learn a phoneme representation on the one hand and speaker embedding on the other hand. DRVC [228] uses Figure 2.5b data-driven inductive prior principle, to discriminate content from speaker information. *Cascade Deep Factorization* (CDF) [138] is a fine example of a more advanced model-driven inductive bias principle: learned representations are reinjected further in the model, to enforce the scrapping of already captured features, and guide the modeling of remaining, “residual”, information in another representation space. A hierarchical structure of features is hence built, which can be assumed to accurately model real-world factor relationships, wherein concepts may be tied to more abstract ones [13]. This concept of building abstract knowledge from shallower ones is at the basis of deep learning formulation, with the ready difference that abstract concepts are not explicitly desired interpretable in conventional deep learning hidden layers. Li et al. (2018) [138] are further exploiting CDF by splitting linguistic, speaker and emotional features from speech, with supervision. To close the walk, AutoVC [179] relies on the information injection principle, by conditioning the decoding stage with a speaker embedding, hence enforcing the encoding stage to supply only the remaining speech information, to perform VC.

Although it is not aligned with the principles of interest in this manuscript, information factorization is not incompatible with hints introduced in Subsection 2.1.2. It does make sense and have advantages to rely on inductive biases to explicitly control information flow throughout a model. This can aid disentanglement by performing preliminary filtering to prune the excessive amount of stimuli within data. One can consider factorization as coarse disentanglement of information, and this manuscript is concerned with finer disentanglement of individual underlying factors. To give an example of how both approaches can complement each other and be efficiently combined, FHVAE [92] relies on a pair of self-supervised hierarchical latent spaces, with suited temporal-based inductive biases, to factorize short-term and long-term variabilities of speech, while disentangling speech factors in latent dimensions. FHVAE will be leveraged Chapter 3, to study its ability to disentangle speech attributes from realistic speech data. Hence, the incoming discussions are about disentanglement as it was described previously in Subsection 2.1.2. Subsequently, Section 2.2 covers the neural-based approaches to disentangle factors of variations.

2.2 Neural networks for generative factors discovery

In order to extract the underlying source factors of variations, one needs to model their relationship with observable samples. Autoencoding generative models are good candidates towards this purpose, as they learn to encode data salient variabilities necessary to reconstruct input samples in a decoding stage. With minimal knowledge about data, self-supervised generative models can learn insightful abstract representations, and their disentanglement can be enforced with a suitable regularization scheme.

Among the various existing generative models, the *Variational Autoencoder* (VAE) [115] stands as the most promising framework to achieve disentanglement. Its derivation is succinctly described in Subsection 2.2.1. Then Subsection 2.2.2 presents the subtleties behind the reconstruction accuracy/latent informativeness trade-off when training a VAE. Finally, Subsection 2.2.3 foreshadows why VAE is a suitable framework for disentanglement, and goes through the extensive investigations undertaken to explicitly enforce the disentanglement capacity of VAEs.

2.2.1 Variational Autoencoder (VAE)

A set of observed samples $x \in \mathbb{X}$, of dimensionality n can be considered as the realization of a collection of random variables \mathbf{x} , dawning from a hidden generative process governed by l ground-truth factors of variations $\mathbf{f} = \{f_i\}_{i \in \{1, \dots, l\}}$. The purpose of Bayesian inference is to model this unobservable generative process, by means of m latent features $\mathbf{z} = \{z_i\}_{i \in \{1, \dots, m\}}$. Assuming that latent variables \mathbf{z} belong to a family of prior distributions $p_\theta(z)$ parameterized by θ , an accurate modeling of factors \mathbf{f} through latents \mathbf{z} is achieved if observations are likely to be generated by latents, i.e., if a high likelihood $p_\theta(\mathbf{x}|\mathbf{z})$ is reached. In order to model the generative process, one has to maximize the likelihood probability of the observations marginalized over the latent space:

$$p_\theta(\mathbf{x}) = \int_{\mathbf{z}} p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})d\mathbf{z}. \quad (2.5)$$

However, the marginal likelihood $p_\theta(\mathbf{x})$ is typically intractable. Hence, variational inference principle leverages an approximation of the posterior distribution $q_\phi(\mathbf{z}|\mathbf{x}) \approx p_\theta(\mathbf{z}|\mathbf{x})$, which belongs to a family of distributions parameterized by ϕ . Borrowed from Kingma and Welling (2013) [115], a directed graphical model in Figure 2.6 depicts the overall model, with latent variable \mathbf{z} in a white node, observed variable \mathbf{x} in a shaded node, gen-

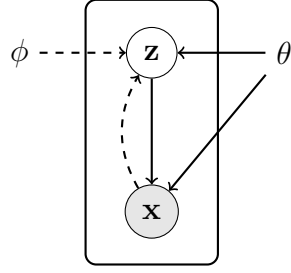


Figure 2.6 – VAE graphical model

erative model parameterized by θ represented with solid lines and inference model with parameter ϕ embodied by dashed lines.

To infer latent features from observations, one has to minimize the dissimilarity between the posterior $p_\theta(\mathbf{z}|\mathbf{x})$ and its variational surrogate $q_\phi(\mathbf{z}|\mathbf{x})$, usually with the Kullback–Leibler divergence:

$$\begin{aligned}
 & D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})) \\
 &= \int_{\mathcal{Z}} q_\phi(\mathbf{z}|\mathbf{x}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})} d\mathbf{z} \\
 &= \int_{\mathcal{Z}} q_\phi(\mathbf{z}|\mathbf{x}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}, \mathbf{x})} d\mathbf{z} + \log p_\theta(\mathbf{x}) \\
 &= \int_{\mathcal{Z}} q_\phi(\mathbf{z}|\mathbf{x}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z})} d\mathbf{z} - \int_{\mathcal{Z}} q_\phi(\mathbf{z}|\mathbf{x}) \log p_\theta(\mathbf{x}|\mathbf{z}) d\mathbf{z} + \log p_\theta(\mathbf{x}) \\
 &= D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) - \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] + \log p_\theta(\mathbf{x}) \\
 &= \log p_\theta(\mathbf{x}) - \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) \\
 &\iff \log p_\theta(\mathbf{x}) \geq \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})). \tag{2.6}
 \end{aligned}$$

Hence, as the marginal log-likelihood $\log p_\theta(\mathbf{x})$ is fixed with respect to $q_\phi(\mathbf{z}|\mathbf{x})$, and based on the Kullback–Leibler divergence non-negativity, it appears that minimizing the distance between the true posterior $p_\theta(\mathbf{z}|\mathbf{x})$ and the variational approximation $q_\phi(\mathbf{z}|\mathbf{x})$ is equivalent to maximizing the *Evidence Lower Bound* (ELBO), which is the corner stone of VAE’s principle:

$$\begin{aligned}
 \mathcal{L}_{\text{ELBO}}(\theta, \phi; \mathbf{x}) &= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\underbrace{\log p_\theta(\mathbf{x}|\mathbf{z})}_{\text{reconstruction error}}] - \underbrace{D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}))}_{\text{posterior deviation from prior}}. \tag{2.7}
 \end{aligned}$$

The prominent works of Kingma and Welling (2013) [115] and Rezende et al. (2014) [187] propose to model the inference model $q_\phi(\mathbf{z}|\mathbf{x})$ and the generative model $p_\theta(\mathbf{z}|\mathbf{x})$ with probabilistic encoder and decoder deep neural networks, respectively. Thus, the variational parameters ϕ and the generative parameters θ are jointly estimated from the data through *Stochastic Gradient Descent* (SGD) optimization. For tractability concerns, one typically lets the latent prior distribution $p_\theta(\mathbf{z})$ be a standard Gaussian, the variational approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ a multivariate Gaussian with diagonal covariance matrix $\Sigma = \sigma^2 * \mathbf{I}_m$ conditioned on \mathbf{x} and ϕ , and the generative model $p_\theta(\mathbf{x}|\mathbf{z})$ a multivariate Gaussian with diagonal covariance conditioned on \mathbf{z} and parameters θ learned by a neural network decoder:

$$p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z}; 0, \mathbf{I}_m), \quad (2.8)$$

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}(\mathbf{x}, \phi), \Sigma(\mathbf{x}, \phi)), \quad (2.9)$$

$$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}(\mathbf{z}, \theta), \Sigma(\mathbf{z}, \theta)). \quad (2.10)$$

Note that under (2.8), the prior $p_\theta(\mathbf{z})$ is free of parameter, i.e., is not conditioned on θ . For convenience, the variational approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ will be henceforth referred to as simply posterior, and $p_\theta(\mathbf{z}|\mathbf{x})$ as the true posterior.

The expectation of the log-likelihood in the ELBO (2.7) can be interpreted as the reconstruction error between input \mathbf{x} and the predicted mean $\boldsymbol{\mu}$, as $(x_i - \mu_i)^2$ appears with the following deviation of the log-likelihood:

$$\begin{aligned} \log p_\theta(\mathbf{x}|\mathbf{z}) &= -\frac{n}{2} \log 2\pi - \frac{1}{2} \log \det \Sigma - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \\ &= -\frac{1}{2} \sum_{i=1}^n \log 2\pi + \log \sigma_i^2 + \frac{(x_i - \mu_i)^2}{\sigma_i^2}. \end{aligned} \quad (2.11)$$

Hence, $\log p_\theta(\mathbf{x}|\mathbf{z})$ can be thought as a reconstruction error cost, as it is optimizing the model to a learned latent space through $q_\phi(\mathbf{z}|\mathbf{x})$ sufficiently informative to be leveraged by the decoder $p_\theta(\mathbf{x}|\mathbf{z})$ to accurately reconstruct the input x , following an autoencoding training scheme. In this respect, many implementations of VAE are considering a deterministic decoder $p_\theta(\mathbf{x}|\mathbf{z})$ using the squared L_2 distance or the *Mean Squared Error* (MSE) between input \mathbf{x} and the predicted reconstruction $\hat{\mathbf{x}} = \boldsymbol{\mu}(\mathbf{z}, \theta)$ in place of the log-likelihood. According to (2.11), the squared L_2 distance or the MSE are equivalent to log-likelihood with a fixed unit likelihood covariance $\Sigma = \mathbf{I}_n$, up to a factor $\frac{n}{2}$ and a

constant c [191]:

$$\begin{aligned}
 \log p_\theta(\mathbf{x}|\mathbf{z}) &= -\frac{1}{2} \sum_{i=1}^n \log 2\pi + (\mathbf{x}_i - \mu_i)^2 \\
 &= -\frac{1}{2} \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + c \\
 &= -\frac{n}{2} \text{MSE}(\mathbf{x}, \hat{\mathbf{x}}) + c.
 \end{aligned} \tag{2.12}$$

The Kullback-Leibler term in the ELBO is to be interpreted as a regularization term in the optimization process. Intuitively, enforcing the posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ to match the standard Gaussian prior distribution $p_\theta(\mathbf{z})$ increases the probabilistic overlap between the individual posterior distributions, by tightening the learned latent space, i.e., reducing the spreading of the posterior means and broadening the posterior variances [20, 156, 2].

An interesting consequence of the D_{KL} term is that it advocates a distributed learned latent space. In the setting framed by (2.8) and (2.9), the Kullback-Leibler divergence between the variational posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and the prior $p_\theta(z)$ in the ELBO (2.7) can be expressed in a very tractable and simple way:

$$\begin{aligned}
 &D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})) \\
 &= D_{KL}(\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}(\mathbf{x}, \phi), \boldsymbol{\Sigma}(\mathbf{x}, \phi))\|\mathcal{N}(\mathbf{z}; 0, \mathbf{I}_m)) \\
 &= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[-\frac{1}{2} \log \det \boldsymbol{\Sigma} + \frac{1}{2} (\mathbf{z}^\top \mathbf{z} - \text{tr}(\boldsymbol{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu})^\top (\mathbf{z} - \boldsymbol{\mu}))) \right] \\
 &= -\frac{1}{2} \log \det \boldsymbol{\Sigma} + \frac{1}{2} \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\mathbf{z}^\top \mathbf{z} - \text{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma})] \\
 &= -\frac{1}{2} \log \det \boldsymbol{\Sigma} - \frac{1}{2} m + \frac{1}{2} \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\mathbf{z}^\top \mathbf{z}] \\
 &= -\frac{1}{2} \log \det \boldsymbol{\Sigma} - \frac{1}{2} m + \frac{1}{2} \sum_{i=1}^m \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [z_i^2] \\
 &= -\frac{1}{2} \log \det \boldsymbol{\Sigma} - \frac{1}{2} m + \frac{1}{2} \sum_{i=1}^m (\sigma_i^2 + \mu_i^2) \\
 &= -\frac{1}{2} \sum_{i=1}^m (\log \sigma_i^2 + 1 - \mu_i^2 - \sigma_i^2).
 \end{aligned} \tag{2.13}$$

It appears that the squared amplitudes of the mean $\boldsymbol{\mu}$ are penalized, i.e., $\sum_{i=1}^m \mu_i^2 = \|\boldsymbol{\mu}\|_2^2$ is to be minimized. It is thence believed that a scenario where multiple dimensions are slightly deviating from the standard Gaussian zero mean is privileged over having a single

dimension strongly straying from $\mathcal{N}(\mathbf{z}; 0, \mathbf{I}_m)$. In other words, high values of μ_i are more penalized than small values, leading to the distribution of the information capacity among latent components¹. It results in a latent space with information driven to be scattered across dimensions rather than concentrated into few ones.

Furthermore, a low Kullback-Leibler divergence in the ELBO induces a latent space less discriminative and informative about data. To concurrently reduce the reconstruction error term, the model has to focus on salient information useful to reproduce the input, to be passed through the information bottleneck [219] embodied by the Kullback-Leibler. Another consequence is that close observations are encouraged to have close latent representations, i.e., reducing the log likelihood is also achieved by smoothing out the latent space. Hence, the optimization of both terms of the ELBO pushes similar data samples to be located in the same vicinity in the latent space. Therefore, the D_{KL} term regulates the structure and the informativeness of the latent space, and ensures that abstract features are learned, hopefully related to the true data generative factors. Put another way, a too high D_{KL} leads to an unstructured, lookup table-like and thus irrelevant latent space. By considering the expectation of the Kullback-Leibler term over observations, one can explicit its regularization function [2, 90, 109]:

$$\begin{aligned}
 & \mathbb{E}_{\mathbf{x} \sim p_\theta(\mathbf{x})} [D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}))] \\
 &= \int_{\mathcal{X}} \int_{\mathcal{Z}} p_\theta(\mathbf{x}) q_\phi(\mathbf{z}|\mathbf{x}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z})} d\mathbf{z} d\mathbf{x} \\
 &= \int_{\mathcal{X}} \int_{\mathcal{Z}} p_\theta(\mathbf{x}) q_\phi(\mathbf{z}|\mathbf{x}) \log \frac{q_\phi(\mathbf{z})}{p_\theta(\mathbf{z})} d\mathbf{z} d\mathbf{x} + \int_{\mathcal{X}} \int_{\mathcal{Z}} q_\phi(\mathbf{x}, \mathbf{z}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z})} d\mathbf{z} d\mathbf{x} \\
 &= \int_{\mathcal{Z}} q_\phi(\mathbf{z}) \log \frac{q_\phi(\mathbf{z})}{p_\theta(\mathbf{z})} d\mathbf{z} + \int_{\mathcal{X}} \int_{\mathcal{Z}} q_\phi(\mathbf{x}, \mathbf{z}) \log \frac{q_\phi(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}) p_\theta(\mathbf{x})} d\mathbf{z} d\mathbf{x} \\
 &= D_{KL}(q_\phi(\mathbf{z}) \| p_\theta(\mathbf{z})) + \mathcal{I}_{q_\phi}(\mathbf{x}; \mathbf{z}), \tag{2.14}
 \end{aligned}$$

with $\mathcal{I}(\cdot; \cdot)$ being the Mutual Information (MI). Hence, based on the Kullback-Leibler non-negativity, over a set of observations, the Mutual Information (MI) between observed data \mathbf{x} and latent variables \mathbf{z} is upper bounded by the Kullback-Leibler term in the ELBO. Given that the MI term $\mathcal{I}_{q_\phi}(\mathbf{x}; \mathbf{z})$ is the informativeness of the latent space regarding the input dataset, its upper bound $\mathbb{E}_{\mathbf{x} \sim p_\theta(\mathbf{x})} [D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}))]$ can be interpreted as the information capacity of the latent space. In simpler terms, the actual amount of

1. Alike Ridge regression [89], the L_2 regularization encourages the distribution of the parameter amplitudes.

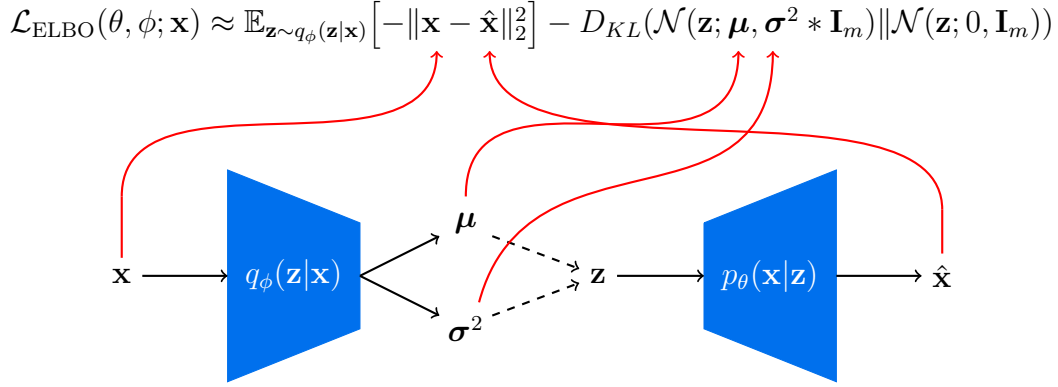


Figure 2.7 – VAE model framework

information filled in the latent space about the data cannot exceed the vessel volume of information allowed by the D_{KL} term:

$$\mathcal{I}_{q_\phi}(\mathbf{x}; \mathbf{z}) \leq \mathbb{E}_{\mathbf{x} \sim p_\theta(\mathbf{x})} \left[D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z})) \right], \quad (2.15)$$

The overall VAE model framework is illustrated in Figure 2.7. During the training stage, the input data \mathbf{x} is fed to the encoder $q_\phi(\mathbf{z}|\mathbf{x})$ to predict a mean vector $\boldsymbol{\mu}$ and a (log)-covariance diagonal $\boldsymbol{\sigma}^2$. A latent vector \mathbf{z} is sampled from $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2 * \mathbf{I}_m)$, as represented by dashed arrows, and is fed to the decoder $p_\theta(\mathbf{x}|\mathbf{z})$ to predict the reconstruction $\hat{\mathbf{x}}$. The sampling procedure is made differentiable by means of the reparameterization trick [115], in order to keep the whole model optimizable through gradient descent. The optimization of such a model is performed by maximizing the ELBO (2.7), which leads to the minimization of the reconstruction error (2.11) and the approximation of the posterior towards the prior distribution (2.13), as embodied by red arrows. During the inference stage, latent samples can be drawn from the prior $p_\theta(\mathbf{z})$ to generate new data samples through $p_\theta(\mathbf{x}|\mathbf{z})$. For any downstream task, one can also feed true data samples to the encoder $q_\phi(\mathbf{z}|\mathbf{x})$ to produce representations, typically in a deterministic way by keeping the most probable point, i.e., $\mathbf{z} = \boldsymbol{\mu}(\mathbf{x}, \phi)$. Further details about the deviations, VAE framework and looser discussions about the ELBO can be found in [115, 187, 114, 46, 68, 52, 15, 2, 20, 156].

VAEs are a powerful framework for estimating arbitrary data distributions $p_\theta(\mathbf{x})$ by approximating the posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ of latent variables \mathbf{z} . However, the investigation of the ELBO (2.7) demonstrates that without a good balance between the two concurrent terms, the model may converge to a suboptimal solution. The trade-off

between reconstruction and information capacity should therefore be carefully addressed, as developed in Subsection 2.2.2.

2.2.2 Information capacity: *A Latent Space Odyssey*

Navigating across the latent space is a perilous journey. As it will be described, one must cautiously balance the trade-off between reconstruction accuracy and information capacity, or they may fall into the abyss of the posterior collapse. At the end of this venture, a disentangled latent space may be discovered as the quest’s reward.

The ELBO trade-off

It has been demonstrated that in its initial formulation, the VAE comprises drawbacks that might lead to undesirable outcomes. Given that the amount of information pertaining to a dataset, i.e., its entropy, varies following the data complexity, modality, sequentially, long- and short-term dependencies and so on, the information bottleneck materialized by the Kullback-Leibler term in the ELBO (2.7) has to be adjusted accordingly. The appropriate amount of information to be transmitted in the latent space is hence to be targeted. In this view, β -VAE [84] introduces a multiplier β to the D_{KL} penalty, controlling the informative bottleneck pressure exercised throughout the encoder. The objective function to maximize becomes:

$$\mathcal{L}_\beta(\theta, \phi; \mathbf{x}, \beta) = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z})). \quad (2.16)$$

The challenge remains in finding the adapted value of β which balances the trade-off between information capacity and reconstruction accuracy. As a matter of fact, poor reconstruction is observed with too great values of β [84]. The visible reason for this trade-off is that in the objective function (2.16), a high value of β leads to an overweighting of the D_{KL} cost and an underrating of the reconstruction error loss during training. In addition to this, (2.15) reveals that minimizing the D_{KL} penalty involves the reduction of the mutual information between data and latent variables, leading to an uninformative latent space.

Based on the decomposition (2.14), InfoVAE [248] proposes to decouple the penalty on the latent space informativeness from the aggregated posterior factorization, with an

objective function not expressed as but equivalent to:

$$\mathcal{L}_{\text{InfoVAE}}(\theta, \phi; \mathbf{x}, \alpha, \lambda) = \mathbb{E}_{\mathbf{x} \sim p_\theta(\mathbf{x}), \mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - \lambda D_{KL}(q_\phi(\mathbf{z}) \| p_\theta(\mathbf{z})) - \alpha \mathcal{I}_{q_\phi}(\mathbf{x}, \mathbf{z}) \quad (2.17)$$

Following this expression of the InfoVAE objective function to maximize, the factorization of the latent space can hence be enforced by penalizing further $D_{KL}(q_\phi(\mathbf{z}) \| p_\theta(\mathbf{z}))$ through a greater λ , without degrading the relevance of the latent space $\mathcal{I}_{q_\phi}(\mathbf{x}, \mathbf{z})$.

Another perceptive to interpret the effect on the encoding capacity of the latent space when manipulating the β can be discerned by examining (2.11). If one posits a fixed value of variance $\sigma^2 = \varsigma^2$ in likelihood Gaussian distribution $p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \mu(\mathbf{z}, \theta), \varsigma^2 * \mathbf{I}_n)$, then the choice of this value plays the same role as tuning β in (2.16) with an MSE or L_2 reconstruction loss in place of log-likelihood, up to a constant and a factor with respect to the MSE both function of ς^2 , as developed by Rybkin et al. (2021) [191]:

$$\mathcal{L}_\sigma(\theta, \phi; \mathbf{x}, \varsigma) = -\frac{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2}{2\varsigma^2} - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z})) + c(\varsigma) \quad (2.18)$$

\equiv

$$\mathcal{L}_{\beta\text{-MSE}}(\theta, \phi; \mathbf{x}, \beta) = -\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 - \beta D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z})) + c. \quad (2.19)$$

Thus, the many implementations using β -VAE with a Gaussian decoder and assuming an MSE penalty are implicitly and equivalently using a raw VAE as defined in (2.7), with an assumed fixed value of decoder variance $\varsigma^2(\beta)$. Built on this insight, Rybkin et al. propose σ -VAE [191], which automatically calibrates the latent information capacity by learning the decoder variance.

Informativeness bounds

Another insightful perspective on the problem can be conferred by rethinking the mutual information between the data and the latent space. By combining the expectation over data of the inequality (2.6) and the decomposition of (2.14), one can deduce that:

$$\begin{aligned} \mathbb{E}_{\mathbf{x} \sim p_\theta(\mathbf{x})} [\log p_\theta(\mathbf{x})] &\geq \mathbb{E}_{\mathbf{x} \sim p_\theta(\mathbf{x}), \mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathcal{I}_{q_\phi}(\mathbf{x}; \mathbf{z}) \\ \Leftrightarrow -\mathbb{E}_{\mathbf{x} \sim p_\theta(\mathbf{x})} [\log p_\theta(\mathbf{x})] &\leq -\mathbb{E}_{\mathbf{x} \sim p_\theta(\mathbf{x}), \mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] + \mathcal{I}_{q_\phi}(\mathbf{x}; \mathbf{z}) \\ \Leftrightarrow \mathcal{H}(x) + \mathbb{E}_{\mathbf{x} \sim p_\theta(\mathbf{x}), \mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] &\leq \mathcal{I}_{q_\phi}(\mathbf{x}; \mathbf{z}). \end{aligned} \quad (2.20)$$

Associated with (2.15), one can meet the findings of Alemi et al. (2018) [2], who demonstrated that the information capacity of the latent space is lower bounded by the entropy of the data discounted by the mean negative reconstruction error, and upper bounded by the mean posterior deviation from the prior:

$$\begin{array}{c}
 \text{data amount of information} \qquad \qquad \qquad \text{latent space informativeness} \\
 \underbrace{\mathcal{H}(\mathbf{x}) + \mathbb{E}_{\mathbf{x} \sim p_\theta(\mathbf{x}), \mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{negative reconstruction error}} \leq \underbrace{\mathcal{I}_{q_\phi}(\mathbf{x}; \mathbf{z})}_{\text{posterior deviation from prior}} \leq \mathbb{E}_{\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{z})} [\underbrace{D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}))}_{\text{posterior deviation from prior}}] . \\
 \hspace{15em} (2.21)
 \end{array}$$

Thus, only the posteriors' shift from the prior can bring some information capacity to the latent space, and the latent space amount of information should at least explain the reconstruction ability of the model. In other words, if the information capacity exhibited by the posteriors' shift from the prior is lower than the data amount of information, one cannot expect to reconstruct without degradation due to the loss of information. This interpretation highlights the importance of finding a proper balance between the lower and upper bounds of the latent information capacity, to get representations both informative and insightful about data.

Posterior collapse

The very extreme case of $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z})) = 0$ means that no matter the input observation x , it will be projected to a standard Gaussian distribution. The latent space hence becomes clueless about the dataset, as it becomes impossible to discriminate latent points all coming from the same distribution $\mathcal{N}(\mathbf{z}; 0, \mathbf{I}_m)$. It is a well-known pitfall referred to as **posterior collapse**, and is illustrated through a simple 2D latent space example in Figure 2.8. Before being trained, the encoder is randomly projecting observations into the latent space, as shown by Figure 2.8a. Resulting posteriors are represented by Gaussians: dots are predicted means $\boldsymbol{\mu}$, and dashed ellipses are predicted variances $\boldsymbol{\sigma}^2$. Thus, each point encompassed by its ellipse is a predicted posterior $q_\phi(\mathbf{z}|x)$ given input x , from which a latent z can be sampled. Two classes are considered, represented in blue and orange. If a good balance is found between reconstruction and disentanglement, a well-formed latent space as in Figure 2.8b can be learned, with posteriors scattered enough to be distinguishable between each other, but suitably aggregating around a centered Gaussian to minimize the Kullback-Leibler cost. The fair degree of overlapping allows a

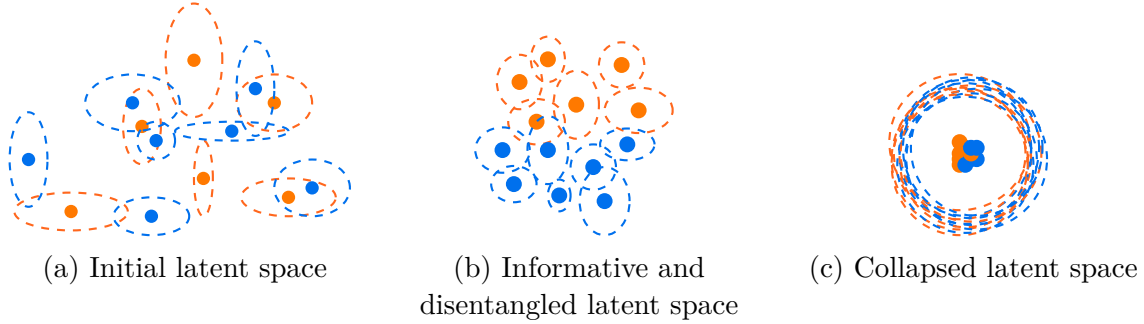


Figure 2.8 – Posterior collapse illustration

smooth latent space, and the classes are disentangled along the vertical axis. Conversely, a latent space fallen into posterior collapse is represented in Figure 2.8c, with all posteriors being predicted to be nearly-standard Gaussians. Such a latent space is uninformative, and arises from a too-strong pressure exerted by the information bottleneck.

Even so, it is worth to note that the overpressure exerted on the information capacity is not the only reason why posterior collapse may happen. As explained by Dai et al. (2020) [35], multiple sources may cause the posterior to collapse. Among them, a too powerful decoder, typically an autoregressive RNN, may be able to perfectly model data distribution $p_\theta(\mathbf{x}|\mathbf{z}) = p_\theta(\mathbf{x})$, thus achieving a high marginal log-likelihood $\log p_\theta(\mathbf{x})$, while ignoring the latent space. The posterior hence falls towards its prior $q_\phi(\mathbf{z}|\mathbf{x}) = p_\theta(\mathbf{z})$, leading to an annealed regularization term $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) = 0$ and clueless representations. The discovery is launched by Bowman et al. (2017) [16], and is the topic of a great deal of efforts to prevent the decoder from locally modeling data information and ensure that the approximated latent distribution encodes meaningful information², through architecture constraints [28], ELBO decomposition [248, 145], D_{KL} annealing [16, 18] (as β -VAE), or optimization scheme adjustments [36, 78, 7].

Towards disentanglement

Altogether, representations learned according to the VAE framework (Subsection 2.2.1) are believed to exhibit properties advised by Bengio et al. [13]. Optimizing the ELBO implies jointly restricting the broadness of the latent space towards the standard Gaussian (through the Kullback-Leibler divergence term) while preserving semantic vicinities (through the reconstruction error penalty). A good trade-off between both penalties,

2. More discussions and insights about posterior collapse in [2, 150, 151, 35].

thanks to the approaches advanced above, leads to a smooth and distributed latent space. In addition, it is believed that VAEs exhibit some prerequisite properties to achieve the automatic discovery of generative factors in their latent space, in a disentangled manner, as it will be elaborated in Subsection 2.2.3.

2.2.3 Enforcing disentanglement: *The Way We Make Contact*

The Variational Autoencoder is a very suitable framework to estimate some data distribution and learn well-structured latent spaces. In addition to the properties reported in Subsection 2.2.1, a great deal of research efforts are studying the disentanglement ability of VAEs, and more specifically with variants built on top of VAE.

Above all, the Kullback-Leibler divergence between the posterior and the prior distributions in the ELBO (2.7) is pressing the posterior to match a standard multivariate Gaussian distribution. One may notice that while the posterior is optimized to approximate a factorized distribution and have a diagonal covariance matrix, latent dimensions tend to be mutually independent. To compile with insights depicted in Subsection 2.2.1, a VAE is learning a **smooth** latent space, restricted to capture **salient** information, **distributed** across **independent** dimensions. It is clear that such a representation space would optimally meet those criteria if its dimensions align with the underlying generative factors of variations \mathbf{f} , as defined in Subsection 2.1.2. True generative factors may not be independent in reality, but it is a fair assumption to start from, akin to Naive Bayes classifier or *Independent Component Analysis* (ICA) [108]. VAE framework hence stands as a good candidate to learn disentangled representations. Rolínek et al. (2019) [189] further demonstrate that the factorized decoder distribution leads to the local orthogonality of the latent space, thus dimensions are implicitly optimized to convey independent variations, i.e., data principal components.

Based on the previous discussions in Subsection 2.2.1 and Subsection 2.2.2, it appears that the β -VAE [84] can provide a convenient way to control the disentanglement ability of a model, by controlling the enforcement of the independence between latent dimensions with β , but not without intricacies described in Subsection 2.2.2. Thus, a higher value of β enforces statistical independence between latent directions, but degrades reconstruction accuracy. In other words, imposing a small D_{KL} brings disentanglement through the first term of (2.14), but also reduces the informativeness of the latent space about the observation through the second term.

Following this intuition, (CCI-VAE) [20] explicitly relaxes the constraint on the infor-

mation capacity by a value κ , which is gradually increased during training. This procedure enforces the latent space to absorb the most salient factors of variations in initial steps, and grants extra capacity in further training steps to capture other factors that may also contribute to the reduction of the reconstruction error. An hyperparameter γ analogous to β in β -VAE (2.16) lets one control the information bottleneck pressure:

$$\mathcal{L}_{\text{CCI-VAE}}(\theta, \phi; \mathbf{x}, \gamma, \kappa) = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - \gamma |D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) - \kappa|. \quad (2.22)$$

Moreover, by further decomposing the information capacity upper bound in (2.15), one can reveal that it is equal to the total correlation over latent dimensions, added by the latent dimension-wise deviation from the unidimensional prior distribution:

$$\begin{aligned} D_{KL}(q_\phi(\mathbf{z})||p_\theta(\mathbf{z})) &= \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z})} \left[\log \frac{q_\phi(\mathbf{z})}{\prod_{i=1}^m p_\theta(z_i)} + \sum_{i=1}^m \log q_\phi(z_i) - \log \prod_{i=1}^m q_\phi(z_i) \right] \\ &= D_{KL}(q_\phi(\mathbf{z})||\prod_{i=1}^m q_\phi(z_i)) + \sum_{i=1}^m D_{KL}(q_\phi(z_i)||p_\theta(z_i)) \\ &= \underbrace{\mathcal{TC}(\mathbf{z})}_{\text{total correlation}} + \sum_{i=1}^m \underbrace{D_{KL}(q_\phi(z_i)||p_\theta(z_i))}_{\text{dimension-wise } D_{KL}}. \end{aligned} \quad (2.23)$$

It hence appears that disentanglement is actually carried out by the total correlation term, as it imposes independence between latent dimensions. FactorVAE [109] and β -TCVAE [26] built on this result, by penalizing further the total correlation with a parameter γ , enforcing disentanglement without altering further the reconstruction ability through $\mathcal{I}_{q_\phi}(\mathbf{x}; \mathbf{z})$ as β -VAE:

$$\begin{aligned} \mathcal{L}_{\beta\text{-TCVAE}}(\theta, \phi; \mathbf{x}, \beta) &= \mathbb{E}_{\mathbf{x} \sim p_\theta(\mathbf{x}), \mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathcal{I}_{q_\phi}(\mathbf{x}; \mathbf{z}) \\ &\quad - \sum_{i=1}^m D_{KL}(q_\phi(z_i)||p_\theta(z_i)) \\ &\quad - \beta \mathcal{TC}(\mathbf{z}). \end{aligned} \quad (2.24)$$

Kumar et al. (2018) [130] emphasize disentanglement with another approach, by introducing in the objective function a penalty over the covariance of the marginal posterior distribution $q_\phi(\mathbf{z})$, to be regulated towards the identity matrix. The aggregated posterior is optimized to match a factorized prior, to accentuate the statistical independence between latent dimensions and hopefully learn disentangled factors. To do so, Kumar et al.

start with the law of total covariance:

$$\begin{aligned}\text{Cov}_{q_\phi(\mathbf{z})}(\mathbf{z}) &= \mathbb{E}_{p_\theta(\mathbf{x})}[\text{Cov}_{q_\phi(\mathbf{z}|\mathbf{x})}(\mathbf{z})] + \text{Cov}_{p_\theta(\mathbf{x})}(\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\mathbf{z}]) \\ &= \mathbb{E}_{p_\theta(\mathbf{x})}[\boldsymbol{\Sigma}(\mathbf{x}, \phi)] + \text{Cov}_{p_\theta(\mathbf{x})}(\boldsymbol{\mu}(\mathbf{x}, \phi)),\end{aligned}\quad (2.25)$$

to define DIP-VAE-I objective function by regularizing $\text{Cov}_{p_\theta(\mathbf{x})}(\boldsymbol{\mu}(\mathbf{x}, \phi))$:

$$\begin{aligned}\mathcal{L}_{\text{DIP-VAE-I}}(\theta, \phi; \mathbf{x}, \lambda_{od}, \lambda_d) &= \mathcal{L}_{\text{ELBO}}(\theta, \phi; \mathbf{x}) - \lambda_{od} \sum_{i \neq j}^m [\text{Cov}_{p_\theta(\mathbf{x})}(\boldsymbol{\mu}(\mathbf{x}, \phi))]_{ij}^2 \\ &\quad - \lambda_d \sum_{i=1}^m \left([\text{Cov}_{p_\theta(\mathbf{x})}(\boldsymbol{\mu}(\mathbf{x}, \phi))]_{ii} - 1 \right)^2.\end{aligned}\quad (2.26)$$

In addition, DIP-VAE-II is defined by penalizing directly $\text{Cov}_{q_\phi(\mathbf{z})}(\mathbf{z})$:

$$\begin{aligned}\mathcal{L}_{\text{DIP-VAE-II}}(\theta, \phi; \mathbf{x}, \lambda_{od}, \lambda_d) &= \mathcal{L}_{\text{ELBO}}(\theta, \phi; \mathbf{x}) - \lambda_{od} \sum_{i \neq j}^m [\text{Cov}_{q_\phi(\mathbf{z})}(\mathbf{z})]_{ij}^2 \\ &\quad - \lambda_d \sum_{i=1}^m \left([\text{Cov}_{q_\phi(\mathbf{z})}(\mathbf{z})]_{ii} - 1 \right)^2.\end{aligned}\quad (2.27)$$

The presented models: β -VAE, CCI-VAE, β -TCVAE, FactorVAE and DIP-VAE-I/II all provide convenient hyperparameters to calibrate the latent dimension independence, and let one connect with the disentanglement capacity of the learned latent space. They effectively demonstrate disentanglement behaviors and improvements over the basic formulation of VAE, in an unsupervised principle, i.e., fully automatic discovery of underlying factors. However, as discussed at the end of Subsection 2.1.2, it is argued by Locatello et al. (2019) [149] that one should not expect to efficiently disentangle some data without resorting to inductive biases to guide the disentanglement. They actually demonstrate the limitations of the models introduced earlier, and conclude on the importance of incorporating prior knowledge about data, by means of specific neural architectures, or (semi-)supervision.

According to this idea, Guided-VAE [44] proposes in its supervised version to explicitly drive latent space dimensions to learn some given tasks, with classifiers optimized through excitation and inhibition losses. Each task is assigned a latent dimension, and the excitation loss maximizes the prediction accuracy of the selected latent, while the inhibition penalty prevents other latents from predicting the task. A semi-supervised extension of VAE is advanced by Kingma et al. (2014) [118], wherein a categorical latent variable is

inserted to represent the partially annotated data, sampled from posterior inference when not observed. Another semi-supervised VAE is proposed by Paige et al. [167], which allows one to consider latent variables from which some labels are partially observed and available, hence providing a semi-supervised training framework. To achieve speech synthesis, FHVAE [92] leverages the prior knowledge of multi-scaled temporal information of speech, to factorize the latent space into two distinct representations, each exhibiting disentanglement behaviors. Capacitron [11] leverages the knowledge of transcription and speaker identity to condition the latent space posterior distribution, which is hierarchically factorized to embody local and global variations. CCI-VAE principle is also exploited to explicitly control the information capacity of both latent spaces. GMVAE-Tacotron [93] capitalizes on observed annotations (e.g., speaker identity) to learn two latent spaces, one for the unobserved factors, optimized as in the regular VAE, and a second one conditioned on the observed factor to learn a Gaussian mixture prior distribution, each component representing a class.

To summarize, the VAE framework, augmented with substantial proposed modifications and refinements, can learn insightful representations with mutually independent latent dimensions. However, it remains unclear if those latent directions can efficiently align with true generative factors. For instance, FHVAE [92], Capacitron [11], and GMVAE-Tacotron [93] are well-designed speech disentanglement models, and some speech attributes seem to be disentangled when going through individual latent dimensions and listening to the synthesized speech. But apart from this manual latent manipulation and subjective listening, no quantitative method is leveraged to objectively assess the disentanglement of speech attributes. To this end, some metrics are developed to quantitatively assess the degree of disentanglement of the proposed approaches. But as the true underlying factors of variations are generally unknown in real data, one has to rely on synthetic data to actually use those metrics. Some widely used synthetic corpora are depicted in Section 2.3.

2.3 Synthetic corpora

In order to determine the disentanglement potential of developed models, a proper framework is necessary, where the true explanatory factors are known. This is why the great majority of the concerned studies use synthetic corpora, to inspect the behavior of their approach within playgrounds, which is simple but totally explained by a small set

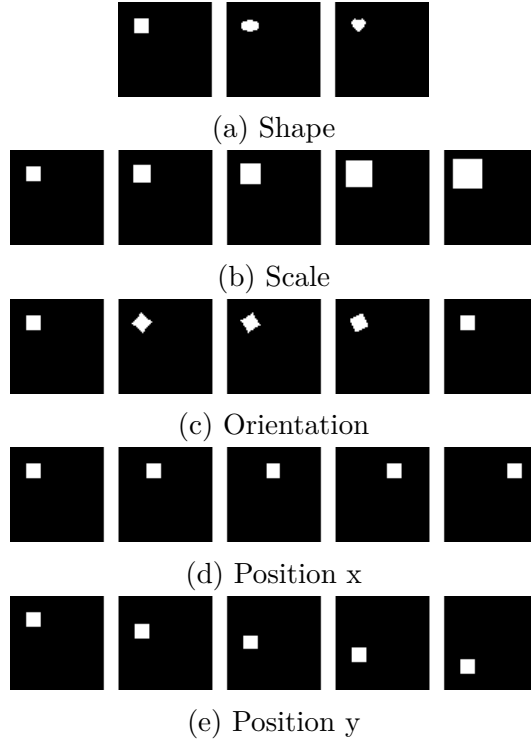


Figure 2.9 – dSprites factors of variation

of variables.

The simple and extensively used synthetic corpus dSprites [157] is considered as a baseline dataset to investigate disentanglement model performances. It comprises 5 factors, shown in Figure 2.9 through samples drawn from the corpus. It is a corpus of black and white images, with different shapes (Figure 2.9a), at different scales (Figure 2.9b), with multiple orientations (Figure 2.9c), and at different x (Figure 2.9d) and y (Figure 2.9e) coordinates.

Cars3D [184] is a collection of generated 3D models of cars, as it can be noticed in Figure 2.10. The 3 generative factors are: the view elevation (Figure 2.10a), azimuth (Figure 2.10b) and the car type (Figure 2.10c).

Furthermore, SmallNORB [137] is a corpus of 3D modeling of toys comprising 4 generative factors, as one can observe in Figure 2.11. There are 5 types of toy shapes (Figure 2.11a), with different views of elevation (Figure 2.11b), azimuth (Figure 2.11b) and lightning (Figure 2.11d).

Some other corpora frequently used are worth to mention, as Shapes3D [19] and MPI3D [67]. In such synthetic corpora, generative factors are completely independent:

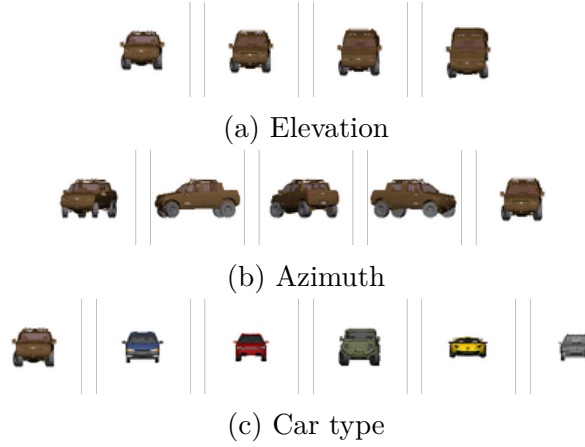


Figure 2.10 – Cars3D factors of variation

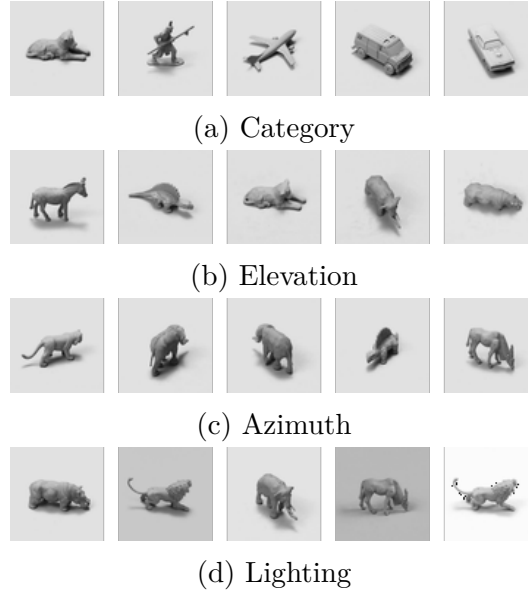


Figure 2.11 – SmallNORB factors of variation

all combinations of factor values are generated, and one has no influence on the other, i.e., the complete grid of possible factor combinations is browsed. Hence, one has to keep in mind that this non-natural exhaustiveness might bias the conclusions of the resulting trained models. Other more realistic datasets are therefore often used to mitigate this phenomenon, such as CelebA [147], 3D chairs [8], or 3D faces [171]. They are more complex, as the true underlying factors are unknown, and all the combinations of the generative factors cannot be covered. It is consequently more difficult to disentangle them, but they also allow experiments closer to real-world applications. They are mostly used for manual

and subjective assessments through reconstruction visualizations.

Nonetheless, it remains useful to rely on synthetic data, especially to employ the various disentanglement metrics proposed to objectively rate the alignment of latent dimensions with the known generative factors. Some instances of such available metrics are described in Section 2.4.

2.4 Measuring disentanglement

The many ways believed to learn disentangled representations have been covered in Section 2.2. Although those models arise from well principled formulations, their disentanglement capacity is implicitly ensued by the optimization towards a factorized marginal posterior i.e. independence of latent dimensions. Consequently, no natural and obvious procedure arises to objectively measure the degree of disentanglement of a given model.

One can rely on the generative ability of the VAE framework to go through the learned latent space and visualize the impact of each dimension on the reconstruction. It is a procedure called latent space *traversal*, broadly used to visually demonstrate the disentanglement power of a proposed model. Some instances of traversals can be found in the next chapter, Table 3.12 and Table 3.13. But traversals do not scale to large investigations, especially with speech synthesis, as it is tedious to listen to every single reconstruction while smoothly traversing each latent direction. In such listening assessments of speech disentanglement, one is looking for continuous transformations of speech attributes. But it might be troublesome to acknowledge which attribute is being altered, and put a precise word on a perceived modification. This echoes with the troubles encountered in perceptual assessment of speech, discussed in Section 1.1.2. Listeners are also prone to semantic satiation [100] when listening over and again to nearly identical synthesis: some latent dimensions might not exhibit any modification on synthesized speech when traversed, hence verbal transformation effect [231] i.e. auditory illusions of phonetic distortions due to mental tiredness, are likely to be experienced and annihilate the relevance of the evaluation.

Thence, the evaluation of speech attributes disentanglement has to be performed using methodical and objective approaches. In this sense, a plethora of metrics have been proposed, each defining and measuring its own properties. While some metrics are consistent with each other, some contradictions arise. This issue stems from the lack of a consensual definition, as discussed in Subsection 2.1.2. Authors fill in the missing points of the

informal shared concept of disentanglement with their own interpretations and requirements, following their context, leading to disparate metrics and potential inconsistencies. As pointed out by Carbonneau et al. (2021) [21], metrics mostly do not correlate when compared for model selection, and do not exhibit the same behaviors in totally simulated scenarios. Adbi et al. (2019) [1] noticed that metrics were not consistent with manual traversal assessments. Understanding such discrepancies is of great interest in developing more reliable metrics, which is part of the concerns of Chapter 4.

Metrics, however, remain the best means to acquire insight when investigating the alignment of latent dimensions with some generative factors. They are worth of interest to discern the boundaries of our comprehension of how information is structured within deep latent representations, and what is expected from them. Following the taxonomy introduced by Carbonneau et al. (2021) [21], one can distinguish 3 types of supervised metrics: predictor-based ones described in Subsection 2.4.1, information theory-based in Subsection 2.4.2, and intervention-based in Subsection 2.4.3. Still, according to Carbonneau et al., and as described in Subsection 2.1.2, metrics usually assess one property among: modularity, completeness and informativeness. However, the common limitation of those metrics is the mandatory knowledge of the true generative factors to disentangle: one still needs to know what to disentangle to appraise if a model disentangles well. A very limited number of unsupervised metrics have been proposed, which are described in Subsection 2.4.4. In the incoming definitions, a set of m latents $\mathbf{z} = \{z_i\}_{i \in \{1, \dots, m\}}$ and a set of l factors $\mathbf{f} = \{f_j\}_{j \in \{1, \dots, l\}}$ are considered.

2.4.1 Predictor-based

An instinctive way to measure the alignment of each latent variable with underlying factors is to try to predict factor values from the learned latent space, with regressors for continuous factors or classifiers for discrete ones. Hence, completeness is achieved if each factor can be predicted from a single latent (or a closed subset of latents following the definition of completeness, as discussed in Section 2.1). In the other way round, a high modularity is reached if each latent is relevant to predict only one factor. Finally, the informativeness naturally arises from the factor prediction accuracy.

Composed of 3 metrics, DCI [55] stands for Disentanglement, Completeness and Informativeness. On a first stage, an $m \times l$ importance matrix R is built, with R_{ij} being the relative weight assigned to latent z_i in the prediction of factor f_j . To compute the disentanglement score, a.k.a., modularity, one has to follow the following process. For

each latent z_i , let P_i be the probability distribution of latent z_i being important in the prediction of each factor. Then, let P_{ij} be the probability of latent z_i being important in the prediction of factor f_j :

$$P_{ij} = \frac{R_{ij}}{\sum_{k=1}^l R_{ik}} \quad (2.28)$$

Let $\mathcal{H}_l(P_i)$ the entropy of distribution P_i , with base l logarithm (denoted as \log_l) for normalization purpose, to ensure that $\mathcal{H}_l(P_i) = 1$ if P_i is uniformly distributed:

$$H_l(P_i) = - \sum_{j=1}^l P_{ij} \log_l P_{ij} \quad (2.29)$$

The disentanglement score of each latent z_i has the following definition, which ranges between 0 and 1, the higher the better:

$$D_i = 1 - H_l(P_i) \quad (2.30)$$

To each latent i is assigned a weight ρ_i , defined as the latent's total importance against the total importance conveyed in R , to discard uninformative latents:

$$\rho_i = \frac{\sum_{j=1}^l R_{ij}}{\sum_{k=1, j=1}^{m, l} R_{kj}} \quad (2.31)$$

The overall disentanglement D is the weighted average of latents disentanglement:

$$D = \sum_{i=1}^m \rho_i D_i \quad (2.32)$$

Completeness is computed using the importance matrix in the other way round, throughout the following steps. For each factor f_j , let Q_j the probability distribution of factor f_j being explained by each latent. Then, let Q_{ij} the probability of factor f_j being explained by latent z_i :

$$Q_{ij} = \frac{R_{ij}}{\sum_{k=1}^m R_{kj}} \quad (2.33)$$

Let $\mathcal{H}_m(Q_j)$ the entropy of distribution Q_j , with base m logarithm:

$$\mathcal{H}_m(Q_j) = - \sum_{i=1}^m Q_{ij} \log_m Q_{ij} \quad (2.34)$$

The completeness score of each factor f_j is defined as:

$$C_j = 1 - \mathcal{H}_m(Q_j) \quad (2.35)$$

Overall completeness score C is the average of factors completeness:

$$C = \frac{1}{l} \sum_{j=1}^l C_j \quad (2.36)$$

Finally, informativeness of each factor I_j is deduced from the prediction accuracy of each predictor. It is advised by Carbonneau et al. (2021) [21] to use a random forest model as the factor regressor or classifier from latents. The overall informativeness is the mean informativeness over factors.

The Separated Attribute Predictability (SAP) score [130] similarly to DCI builds an importance matrix R , where R_{ij} is the R^2 score of a linear regression or the classification accuracy of the prediction of factor f_j from latent z_i . Then, for each factor, the difference between the top 2 values is retained and averaged to get the overall SAP score. This metric, therefore, gauges the completeness of a latent space.

Furthermore, Explicitness [188] measures the informativeness of a representation space, and is defined as the mean Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) curves, obtained from logistic regressions performed on each factor class (one-versus-rest).

Considering the prediction accuracy of factors from latents seems to be a logical approach. But one has to be aware that those procedures might be highly dependent on hyperparameters, and subject to (regressor or classifier) predictor-related biases. Such methods, therefore, have to be carefully manipulated when used as metrics for disentangling model selection.

2.4.2 Information theory-based

Another way to measure the dimension/factor alignment is to rely on information theory-based quantities, as mutual information, to estimate the amount of information conveyed by each latent direction.

The Mutual Information Gap (MIG) [26] is a measure of completeness, and also relies on an importance matrix R , in which R_{ij} is the mutual information between latent z_i and factor f_j . It is then built on the same concept as SAP score [130], by considering the

differences between the two greatest mutual information quantities, for each factor. Gaps are normalized by the respective factor entropy, and averaged to get the global MIG score:

$$MIG = \frac{1}{l} \sum_{j=1}^l \frac{\mathcal{I}(f_j; z_a) - \mathcal{I}(f_j; z_b)}{\mathcal{H}(f_j)} \quad (2.37)$$

with z_a and z_b being the most and second most informative latents for each factor f_j .

Modularity score as introduced by Ridgeway et al. (2018) [188] relies on the same mutual information-based importance matrix as MIG, but is a measure of modularity, by computing for each latent z_i the normalized MSE between the greatest importance, obtained from factor k , and the remaining mutual information values:

$$\delta_i = \frac{1}{R_{ik}^2} \text{MSE}(R_{ik}, R_{ij, j \neq k}) \quad (2.38)$$

The modularity of z_i is then $1 - \delta_i$, and the overall modularity is the averaged value across latents.

UniBound [221] is a more advanced information theory-based metric, in that it employs Partial Information Decomposition (PID) [234] to propose a metric similar to MIG but with a finer insight. PID provides a principled framework to decompose the information pertaining to more than 2 variables, into comprehensible and interpretable *partial information* pieces. More precisely, the amount of information conveyed by a set of source variables \mathbf{s} about a target variable t can be decomposed into *unique*, *redundant* and *synergistic* pieces of information. Figure 2.12 illustrates how those partial informations partition the total mutual information between 2 sources s_1 and s_2 and a target t . The two circles represent the individual mutual information $\mathcal{I}(t; s_1)$ and $\mathcal{I}(t; s_2)$, and the overall ellipse embodies the overall information conveyed by both sources about the target, $\mathcal{I}(t; s_1, s_2)$. The information shared by sources, i.e., *redundantly* conveyed by both sources, is the quantity $\mathcal{R}(t; s_1 : s_2)$ illustrated by the overlap of the circles. For each source, there is a remaining amount of information, not overlapping with each respective counterpart, i.e., the *unique* information of each source. Those quantities are noted $\mathcal{U}_1 = \mathcal{U}(t; s_1)$ and $\mathcal{U}_2 = \mathcal{U}(t; s_2)$. Finally, the information provided by the combination of the sources i.e. only accessible by observing both variables, is the *synergistic* information $\mathcal{S}(t; s_1 : s_2)$.

With this insight in mind, it becomes clear that MIG is overestimating the amount of information disentangled by the most informative latent. All in all, the intent of MIG is to obtain the information uniquely carried by the most informative latent. But MIG is only

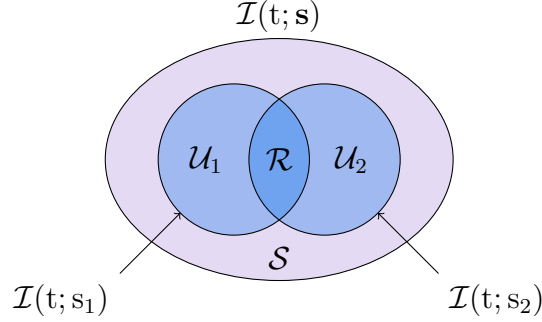


Figure 2.12 – PID illustration

discounting the information coming from the second most important latent. Therefrom, some redundant information overlapping with other latents and information uniquely captured by other latents are not discounted by the gap. To address this issue, Tokui and Sato [221] propose to consider the gap between the most important latent and all other latents:

$$\text{UNIBOUND} = \frac{1}{l} \sum_{j=1}^l \frac{\max_i [\mathcal{I}(f_j; z_i) - \mathcal{I}(f_j; \mathbf{z}_{\setminus i})]}{\mathcal{H}(f_j)} \quad (2.39)$$

Information theory-based metrics have the advantage of not directly relying on hyperparameters to be computed, though they are dependent on the quantization scheme applied to enable discrete mutual information usage. The choice of mutual information estimation might also influence the resulting measure.

2.4.3 Intervention-based

By fixing one factor and randomly drawing others, intervention-based metrics assess whether a dimension of the encoded latent space remains stable throughout samples. In other words, such a metric ensures that while fixing a factor, a latent variable stays invariant with respect to interventions on factors.

Z_{diff} [84], a.k.a., β -VAE score, constitutes a batch of pairs of latent representations, each pair being encoded from data samples sharing a common generative factor, while others are drawn randomly. This is used to build a classification problem, where the mean of the absolute difference of the generated pairs forms one input value, while the associated fixed factor is the label. A set of such batches is created, to train a simple linear classifier to predict which factor was fixed to generate each batch based on mean absolute differences. This metric hence grants a good score when, for each factor, at least

one latent direction remains consistent to factor's fluctuations, and robust to variations of other factors. Z_{diff} hence measures the modularity of a latent space. However, failure modes have been identified with this metric [109, 204], reconsidering its relevance.

$Z_{\text{min-var}}$ [109], a.k.a., FactorVAE score slightly modifies the Z_{diff} procedure, in an attempt to address its drawbacks. Given a fixed factor, a batch of latent representations is generated, normalized by the latent empirical standard deviation across the dataset. Then, the dimension with smaller variance within the batch is retained, and forms a training point with the fixed factor as label to be predicted by a majority-vote classifier. Dead latent dimensions are pruned when computing the empirical standard deviation across the dataset, in order to ignore them when selecting the dimension with smaller variance.

Finally, Interventional Robustness Score (IRS) [213] specifies a causal framework to define the modularity of latent components. Therefore, the proposed metric relies on the causal effects of *nuisance* factors' perturbations on individual latent dimensions, knowing that a drawn *relevant* factor is kept fixed. More formally, let z_k a latent component, f_i a realization of factor f_i and \mathbf{f}_J a realization of the remaining factors \mathbf{f}_J ($J = \{1, \dots, l\} \setminus \{i\}$). IRS is based on the Post Interventional Disagreement (PIDA), being the expected absolute distance from z_k under realization f_i when experiencing a perturbation \mathbf{f}_J :

$$\text{PIDA}(k|f_i, \mathbf{f}_J) = \left| \mathbb{E}[z_k|do(f_i)] - \mathbb{E}[z_k|do(f_i, \mathbf{f}_J)] \right|, \quad (2.40)$$

the do operator being described in Subsection 2.1.2. The perturbation maximizing this distance is considered:

$$\text{MPIDA}(k|f_i, J) = \sup_{\mathbf{f}_J} \text{PIDA}(k|f_i, \mathbf{f}_J), \quad (2.41)$$

and weighted by each f_i realizations:

$$\text{EMPIDA}(k|i, J) = \mathbb{E}_{f_i}[\text{MPIDA}(k|f_i, J)]. \quad (2.42)$$

The IRS is then the normalization by the maximal deviation from expected z_k under any variations of factors \mathbf{f} , with a slight modification to get a score ranging between 0 and 1, the higher the better:

$$\text{IRS}(k|i, J) = 1 - \frac{\text{EMPIDA}(k|i, J)}{\text{EMPIDA}(k|\emptyset, \{1, \dots, l\})}. \quad (2.43)$$

Then, the factor f_i minimizing the normalized expected maximum perturbation on z_k under all other factors $\mathbf{f}_{\{1,\dots,l\}\setminus\{i\}}$ variations is kept to get the modularity D_k of z_k :

$$D_k = \max_{i \in \{1,\dots,l\}} \text{IRS}(\{k\}|\{i\}, \{1, \dots, l\} \setminus \{i\}). \quad (2.44)$$

While having the advantage of not making prior assumptions on the relationships between latents and factors, intervention-based metric behaviors are strongly grounded in the sampling schemes, i.e., the number and size of batches, to estimate expected distances and variances.

2.4.4 Unsupervised metrics

The numerous proposed approaches testify the lack of a formal and consensual definition of disentanglement. They are bounded by the knowledge of the true generative factors, which is a significant obstacle towards the experimentation of disentanglement models in real and complex scenarios. Few attempts have yielded unsupervised metrics for disentanglement assessment, and they are mainly developed in and for image processing contexts.

Unsupervised Disentanglement Ranking (UDR) [51] is based on Rolínek et al.’s (2019) [189] founding that if a given model converges until disentanglement, it will always do so through similar latent spaces, up to signed permutations. Based on this principle, UDR performs pair-wise comparisons over a large set of trained latent spaces, and assigns a high disentanglement for models that find many other models with a similar latent space.

Variation Predictability (VP) [249] metric proposes to compare pairs of decoded data, generated from latent representations with only one differing dimension. A classifier is trained to predict the differing latent direction, based on the difference between each pair. The accuracy of the predictor is considered as the disentanglement score. Hence, if variations of each latent component are easily recognizable, the latent space is assumed to be disentangled.

Pertaining to computer vision concerns, Traversal Perceptual Length (TPL) [250] leverages a VGG16 [206] model to compute the accumulated perceptual distance between generated images while traversing a model. It is hence assumed that when traversed, disentangled representations should lead to smooth variations on generated images, i.e., small perceptual differences. This metric has the advantage of not relying on a swarm of models to train, but is more suited for image evaluation. Likewise, Perceptual Path

Length (PPL) [174] also employs deep visual representations from computer vision neural networks to estimate the perceptual similarity between generated images. Nevertheless, both metrics are believed to be more related to the smoothness of the latent space than its disentanglement.

While UDR relies on a bunch of trained models, other metrics assume that perceptual distances can be estimated by way of output differences or deep features. The latter perceptual-based metrics are promising research directions for future unsupervised disentanglement metrics, however, no analogous evaluation scheme has yet been proposed for speech disentanglement assessment.

2.5 *Opening the pod bay doors*

Ultimately, extensive research on disentanglement learning and assessment has led to a wealth of insights, with a diverse range of principles and approaches depending on the desiderata. With the expansion of unsupervised learning of general-purpose speech representations [9, 94, 27], it becomes clear that artificial intelligence is reaching a strong capacity to extract meaningful knowledge from utterances. Disentanglement strives to go one step further, by supplying explicit and interpretable representations, and is a promising paradigm towards the understanding of the very intricate speech attributes.

The properties advocated by Bengio et al. (2013) [13] of latent representations, namely smoothness, distributness, and disentanglement, have been argued in Section 2.2 to be achievable by disentanglement learning through VAEs. Although the fine-grained disentanglement of speech characteristics is still limited, some studies exhibit encouraging results, especially for controllable synthesis [92, 93, 11]. Nevertheless, finding the appropriate inductive biases and conditioning schemes with available knowledge to acquire full control over speech synthesis is a challenging problem that is the topic of many ongoing academic and industrial investigations.

As described in Section 2.4, quantifying the disentanglement of a model is all but trivial. Additional complexities pertaining to speech attributes and quality assessment make the evaluation of speech disentanglement an open problem, i.e., traversing latent dimensions requires the listening of every generated utterance, leading to a tedious manual evaluation, prone to semantic satiation and auditory illusions. Hence, there is currently no reliable and scalable appraisal scheme for speech disentanglement.

In short, learning to disentangle is a difficult undertaking per se, and the challenge of

adapting these methods to the particularities of speech is all the greater. Hence, a gap is still to be bridged between disentanglement learning and speech processing, and this manuscript is only gently opening the pod bay doors of the great odyssey towards latent spaces aligned with speech intricacies. Upcoming parts, Chapter 3 and Chapter 4, will depict the contributions made to this enterprise provided within the scope of the present thesis.

PART II

Speech Attributes Disentanglement

TOWARDS SPEECH ATTRIBUTES DISENTANGLEMENT

Along Chapter 1 were detailed the various variations pertaining to speech, broadly in terms of linguistic, paralinguistic, and extralinguistic facets according to Section 1.1. Neural methods leveraged to address those aspects are depicted in Section 1.2, but their ability to efficiently and independently control each attribute is still limited. Hence, Chapter 2 introduced promising hints to disentangle individual characteristics and control them in synthesis.

Procedures to learn an abstract representation space able to align salient factors of variations with distinct dimensions have gained a lot of interest over the past years. Disentanglement learning is now a full-fledged research area, aiming to automatically distinguish and separate independent properties of observed data. It discards irrelevant information for underlying tasks, and overall provides interpretable representations, leverageable in generative models for controlled data synthesis.

Learning disentangled representations, however, remains a challenge. No formal definition is yet accepted, leading to the lack of a well-defined metric, and the knowledge of ground-truth factors to disentangle is still necessary, limiting studies to synthetic datasets. In order to unveil the involved complications, Section 3.1 describes preliminary experiments conducted on synthetic image corpora.

Together with the temporal intricacies of speech perceptual properties, learning disentangled speech representations is even more challenging. To investigate solutions towards this purpose, one may benefit from a well-defined frame for speech, such as synthetic image corpora. To this end, diSpeech is introduced in Section 3.2, with the results of experiments conducted using it.

Despite the growing number of materials focused on disentanglement principles, applications to realistic data struggle to recover the same results as on synthetic data. Authentic utterances inherently convey more variabilities, which are still hard to formally

Model	Hyperparameter		nb latents
VAE			$\{8, 16, 32\}$
β -VAE	β	$\{2, 4, 8, 16, 32, 64, 128\}$	$\{8, 16\}$
β -TCVAE	β	$\{2, 4, 8, 16, 32, 64, 128\}$	
FactorVAE	γ	$\{2, 4, 8, 16, 32, 64, 128\}$	
CCI-VAE	γ	$\{2, 8, 32, 128\}$	
	κ	$\{5, 10, 25, 50, 100\}$	

Table 3.1 – Models hyperparameter settings

define and hierarchize (Section 1.1). In addition, if such an aspect could be precisely defined, it cannot be completely covered based solely on observations, making its disentanglement further complex. In an attempt towards disentanglement of real speech data, Section 3.3 reports the results and the behaviors observed from more advanced speech disentanglement models trained on real utterances.

3.1 Synthetic image datasets disentanglement

The most convenient medium to investigate the disentanglement abilities of a generative model is images. As described in Section 2.3, synthetic image datasets are typically used to assess an algorithm’s efficiency, as one can literally observe the captured information by traversing each representation dimension and examining the generated variations. Such corpora are suited playgrounds to prospect the behavior of disentanglement models and metrics.

Hence, Subsection 3.1.1 details the experimental setup, methodology and expectations, whereas Subsection 3.1.2 presents the results and Subsection 3.1.3 draws the conclusions.

3.1.1 Setup and expectations

For those preliminary experiments, synthetic image corpora described in Section 2.3 are leveraged, namely dSprites [157], Cars3D [184] and SmallNORB [137]. The very first experiments are conducted with the basic VAE framework, as described in Subsection 2.2.1. Using synthetic corpora and the materials provided by `disentanglement_lib` [149], several numbers of latent dimensions ($\text{nb latents} \in \{8, 16, 32\}$) are tested to appraise the

	Parameter	Value
Training	Batch size	256
	Training steps	300000
Adam optimizer	β_1	0.9
	β_2	0.999
	ϵ	1e-08
	α	1e-04

Table 3.2 – Training parameters

effect on disentanglement metrics: DCI [55], MIG [26], IRS [213], Z_{diff} [84] and $Z_{\text{min-var}}$ [109]. The degree of accordance between metrics can also be assessed in this preliminary set of experiments. In addition, some traversals are inferred to visualize the actual information captured by each latent direction, and testify the coherence with metrics.

To go further towards the disentanglement of the synthetic corpora, some of the more advanced models described in Subsection 2.2.3 are tested, with multiple settings of hyperparameters and number of latent dimensions. The trained models are β -VAE [84], β -TCVAE [26], FactorVAE [109] and CCI-VAE [20], with the hyperparameters presented in Table 3.1. The hyperparameters β and γ regulate the information bottleneck pressure. Values among $\{2, 4, 8, 16, 32, 64, 128\}$ are tested, to assess the effect of the bottleneck on the latent space structure. For CCI-VAE, κ controls the information capacity, explicitly let to the latent space (see Subsection 2.2.3 and equation (2.22)). Multiple values of κ , within $\{5, 10, 25, 50, 100\}$ are tested, also to explore the effect on the learned representations. In addition, two sizes of latent space, 8 and 16 components are used. Following `disentanglement_lib`, models are implemented with a convolutional encoder and decoder. Hence, only the optimization scheme, i.e., the ELBO, differentiates the various models. All models follow the same training scheme and an Adam optimizer [116] setting, as displayed in Table 3.2.

While the basic VAE model is likely to reach a good reconstruction loss, better disentanglement abilities are expected from its deviations: β -VAE, β -TCVAE, FactorVAE, and CCI-VAE, especially for salient factors of variation. This assumption is assessed through disentanglement metrics and some traversals. One can expect to determine the best-performing model and setting thanks to those disentanglement metrics. Consistency between metrics and with observed traversals is also to be verified, as metrics reliability is discussed by Carbonneau et al. (2021) [21] and debated by Locatello et al. (2019) [149].









































































nb latents	dSprites						Cars3D					
												
8												
16												
32												
	SmallNORB											
												
8												
16												
32												

Table 3.3 – Reconstructions of VAE with 8, 16 and 32 latent dimensions on dSprites, Cars3D and SmallNORB. Real image samples are displayed below their respective corpus name, and the corresponding reconstructions are listed below.

3.1.2 Results

Reconstructions of VAEs trained on dSprites, Cars3D, and SmallNORB can be found in Table 3.3. Pictures right below corpus names are original samples drawn from the corpora, and the images listed below are the corresponding reconstructions, each line representing a model with different number of latent dimensions (nb latents). For the models trained on dSprites and Cars3D, reconstructions are fairly good, regardless of the number of latent. Reconstructions are, however, fuzzier for SmallNORB, which can be explained by its greater complexity. Nevertheless, the lighting, rotation, and global shape of the objects are still recognizable. Overall, all models have converged towards states where their latent spaces are fairly informative about the data.

In order to acquire a deeper understanding of the information conveyed by the learned latent spaces, traversals are displayed in Table 3.4, from models trained with 8 latent components. Each line is a distinct latent direction, being traversed between -2 and 2, with 6 steps. When a dimension is traversed, other ones are set to 0. Traversals clearly uncover different behaviors among corpora. Factors of dSprites do not seem to be prop-

erly disentangled, i.e., variations along individual dimensions are irregular and somewhat random. The good reconstruction over dSprites does not translate into an insightful latent space. Concerning Cars3D, factors seem to be smoothly varying along some axis, but still in an entangled way. With SmallNORB, the dimension 0 clearly and smoothly controls the lighting. Despite the fuzzy reconstruction quality, dimension 2 seems to capture the object category. Other dimensions appear to capture some variations, but reconstructions are too noisy to come to further conclusions. Hence, despite the absence of any mechanism to enforce disentanglement, VAE exhibits some disentanglement behaviors, yet they are limited.

To settle VAE’s disentanglement capability, metrics resulting from the same VAE models are presented in Table 3.5. The first point arising from the bar plots is the discrepancy between metrics. Z_{diff} always forecasts a very optimistic disentanglement, while MIG is noticeably pessimistic. It also appears that the number of latent dimensions does not have a significant influence on the disentanglement scores. Overall, it remains difficult to conclude on the disentanglement ability of the raw VAE models through metrics.

Lets now examine the disentanglement of other models, designed to favor disentangled representations. As depicted in Subsection 3.1.1, 4 models are tested: β -VAE, β -TCVAE, FactorVAE, and CCI-VAE; and a set of hyperparameters are tested for each of them, leading to a substantial amount of models to investigate. To get a clearer picture, Table 3.6, Table 3.7 and Table 3.8 show the disentanglement metrics and the reconstruction loss over dSprites, Cars3D, and SmallNORB, respectively. Models trained with 8 latent dimensions are depicted in the left column, and those with 16 dimensions in the right one. Then, each line is a different model, and in each figure, metrics are plotted against information bottleneck hyperparameters, β or γ , following the model. CCI-VAE has additional lines and figures, as it comprises a second hyperparameter κ , the information capacity. The black dashed line is the reconstruction loss, which represents the mean distance between original inputs and reconstructions, reached at the end of the training.

The disentanglement of dSprites, measured in Table 3.6, remains hard to assess through metrics, as in Table 3.5 i.e. some metrics reach high scores while others are quite low, in almost all situations. However, common variation patterns can be discerned for some hyperparameter settings. For instance, β -VAE seems to reach better scores with 8 latents and $\beta = 32$, or FactorVAE with 8 latents and $\gamma = 32$. However, reconstruction error highly increases when the information bottleneck pressure controlled by β or γ are too important, which reveals the trade-off between information bottleneck and reconstruction

Latent	Traversal frames											
	dSprites						Cars3D					
	-2					2	-2					2
0												
1												
2												
3												
4												
5												
6												
7												
	-2					2	SmallNORB					
0												
1												
2												
3												
4												
5												
6												
7												

Table 3.4 – Traversals of VAE with 8 latent dimensions trained on dSprites, Cars3D and SmallNORB

capacity.

Regarding disentanglement metrics on Cars3D, in Table 3.7, intervention-based ones clearly appear to be very optimistic compared to others, more reserved. Z_{diff} and $Z_{\text{min-var}}$ reach high scores, while IRS stands around 0.5, and others rarely rise above 0.4. In DCI,

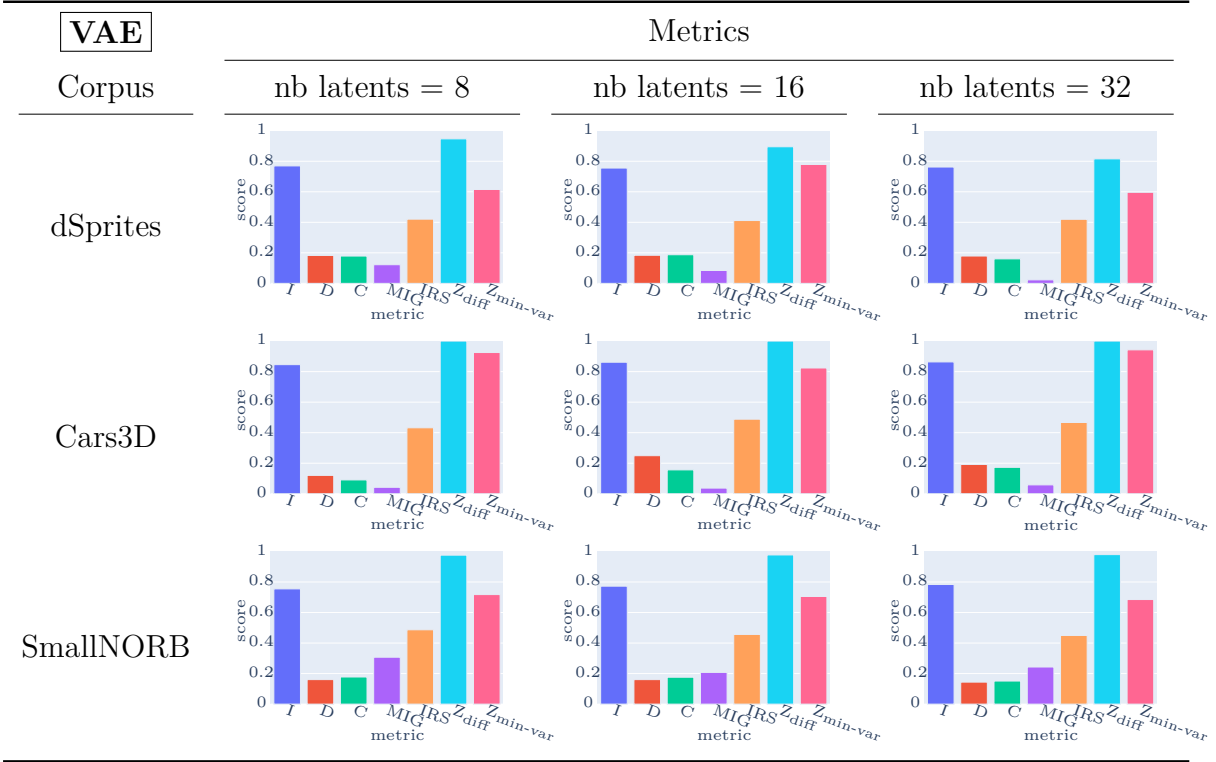
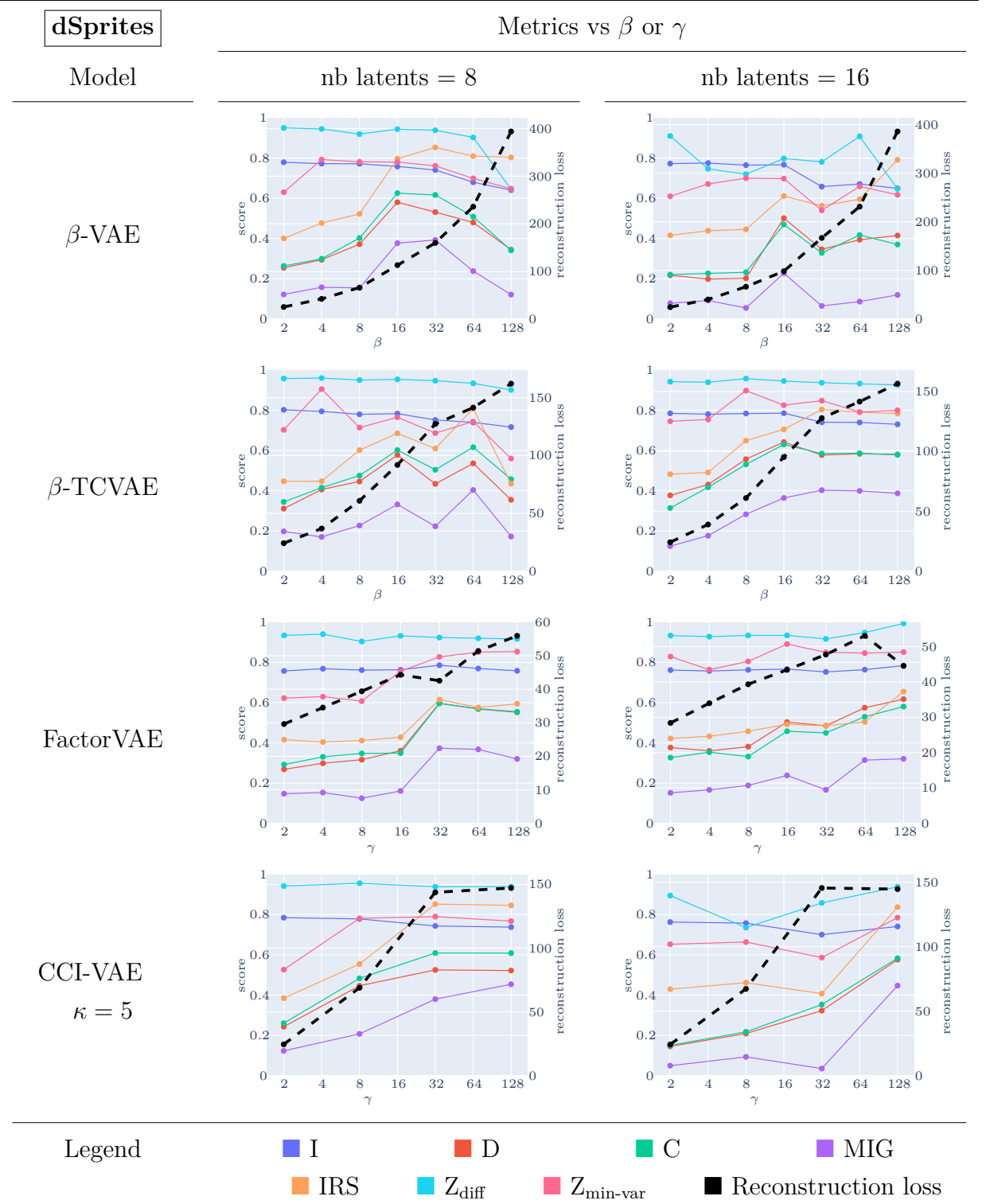


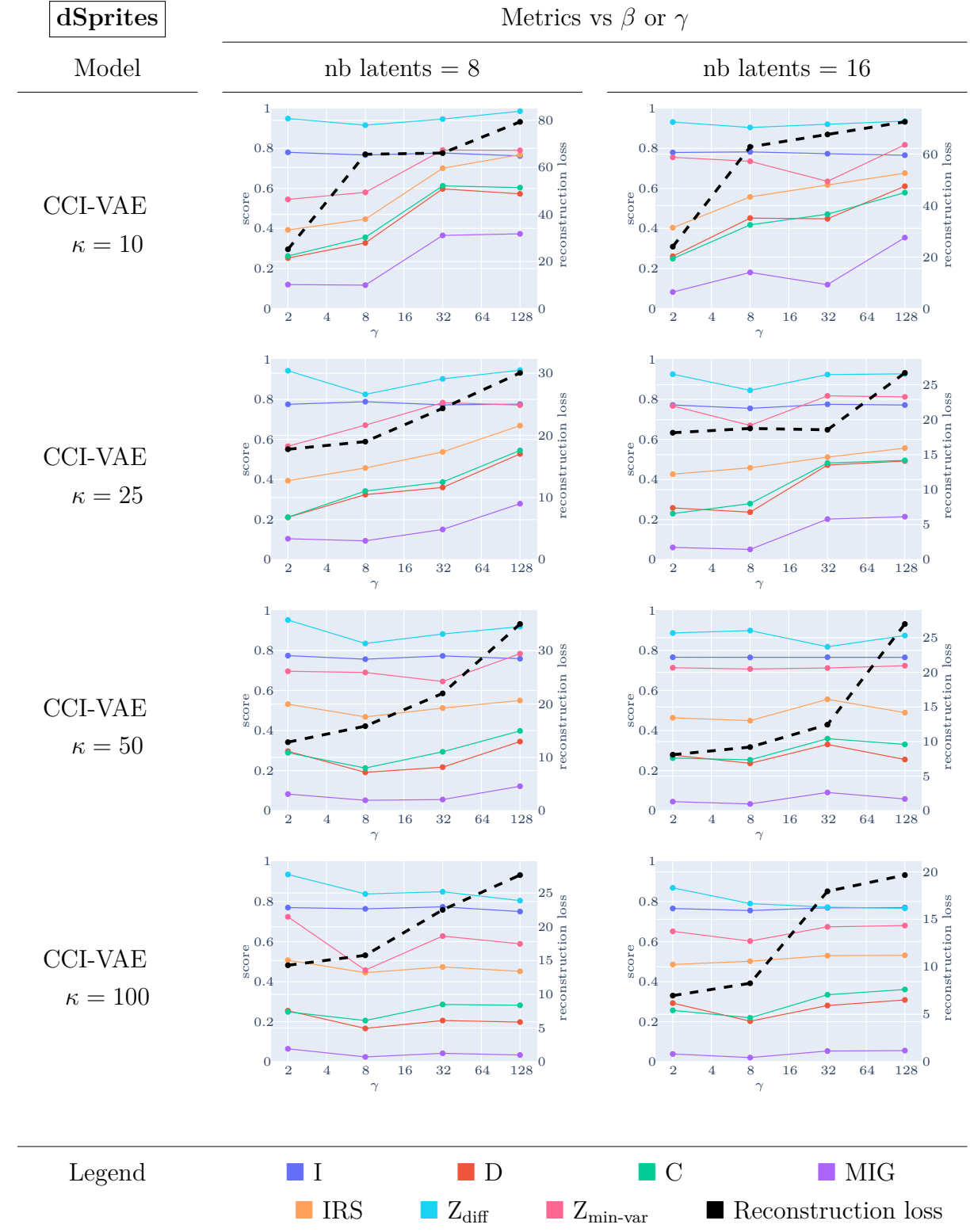
Table 3.5 – Metrics of VAE trained on dSprites, Cars3D and SmallNORB

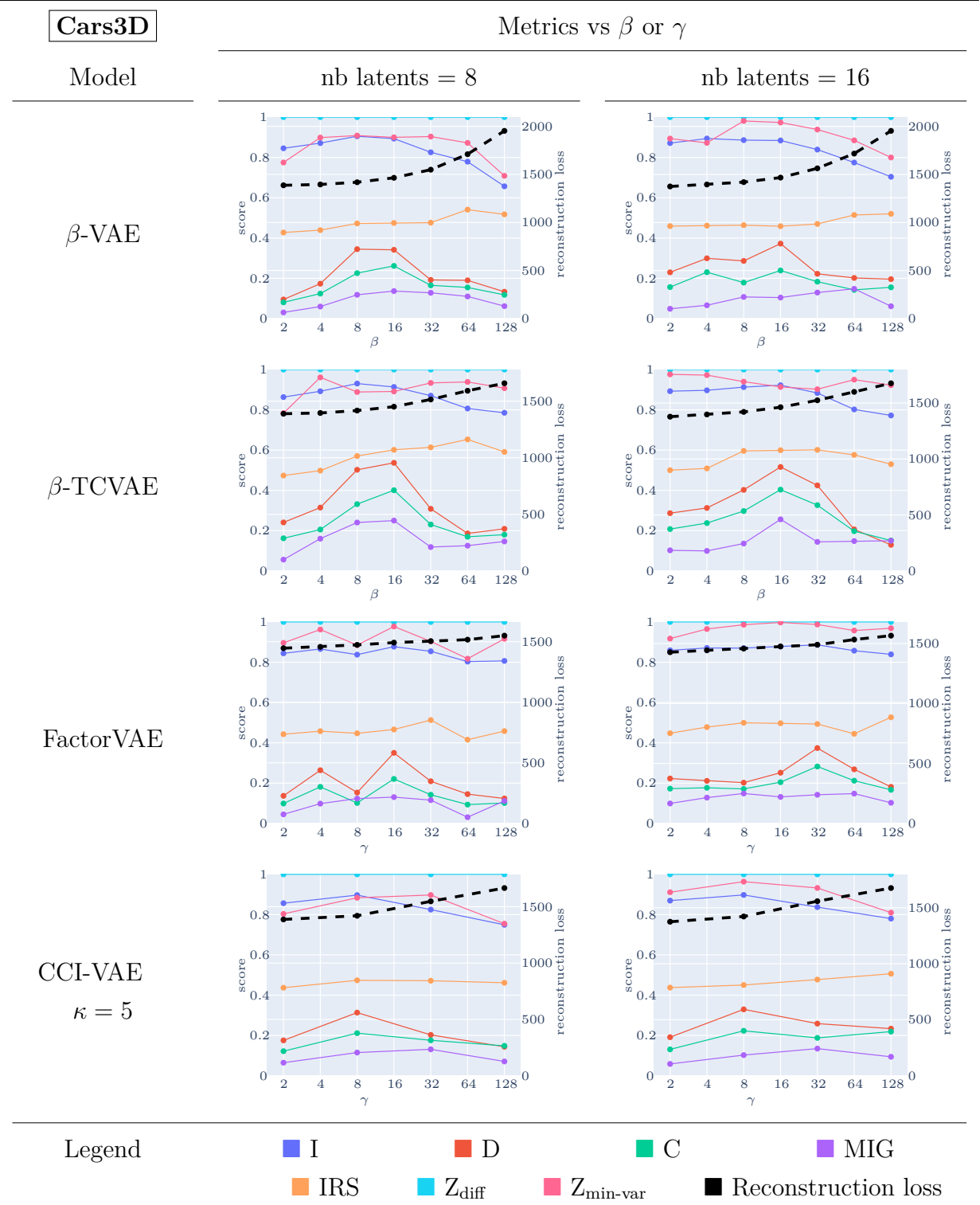
Informativeness (I) reaches high scores, but actually measures a global quantity of information, and not some degree of disentanglement (see Subsection 2.4.1 for more details about DCI). As with dSprites, it is hard to decide which model performs the best disentanglement, but some peaks can be identified when varying the information bottleneck pressure, with β or γ . As with dSprites, too high values of β or γ degrade the reconstruction ability of the models.

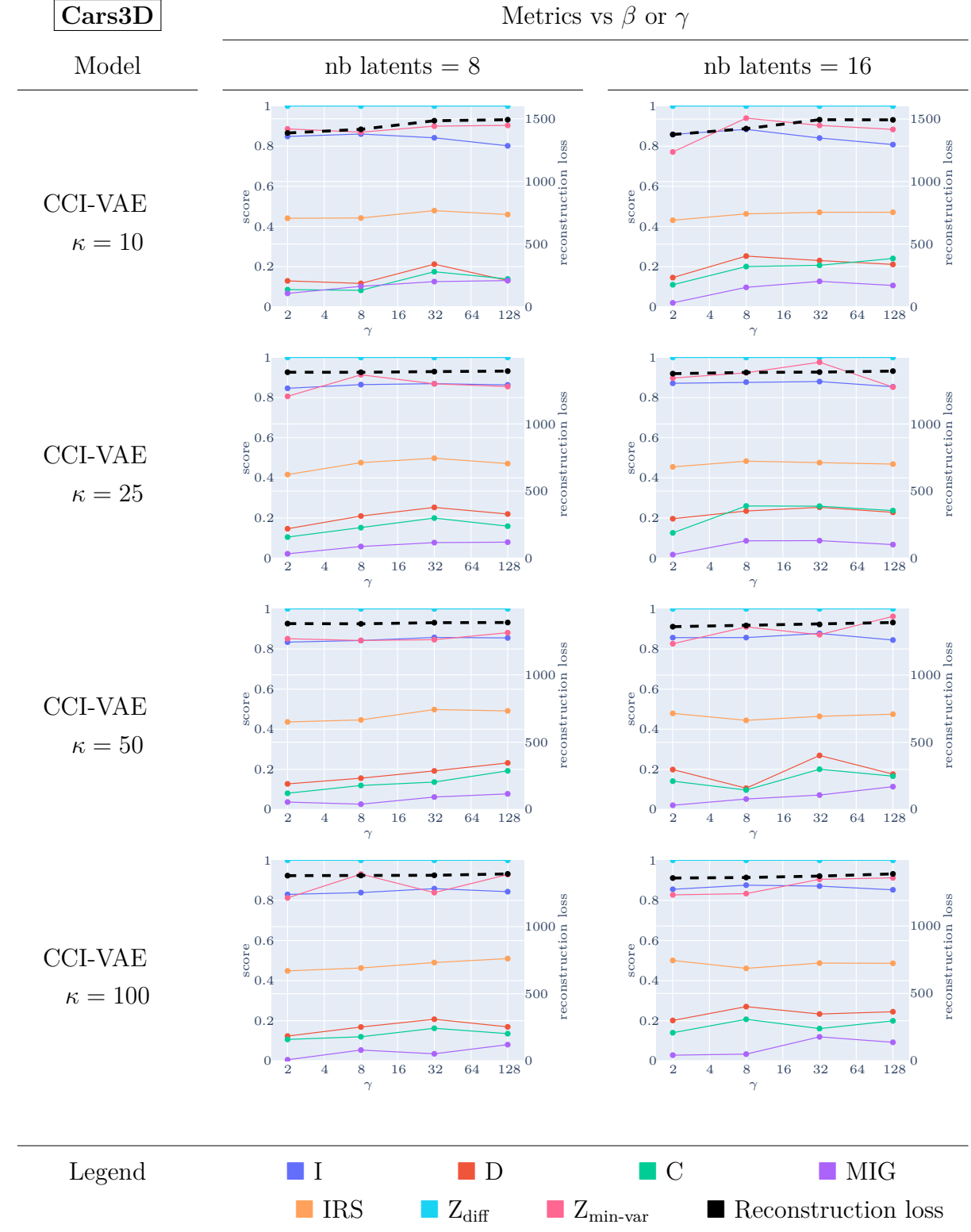
With SmallNORB, in Table 3.8, Disentanglement and Completeness from DCI, and MIG scores express poor disentanglement performances. Here again, no model distinctly stands out from the crowd regarding disentanglement metrics. It is also worth noting that the reconstruction error is constantly high. Basic VAE was already struggling to properly reconstruct images from SmallNORB, hence, it is not surprising that enforcing further disentanglement can only worsen the reconstruction error.

In general, higher disentanglement scores are reached on dSprites and Cars3D. But it remains hard to determine the best-performing model regarding the metrics. Good scores are reached in some settings, but often at the cost of a very degraded reconstruction accuracy.

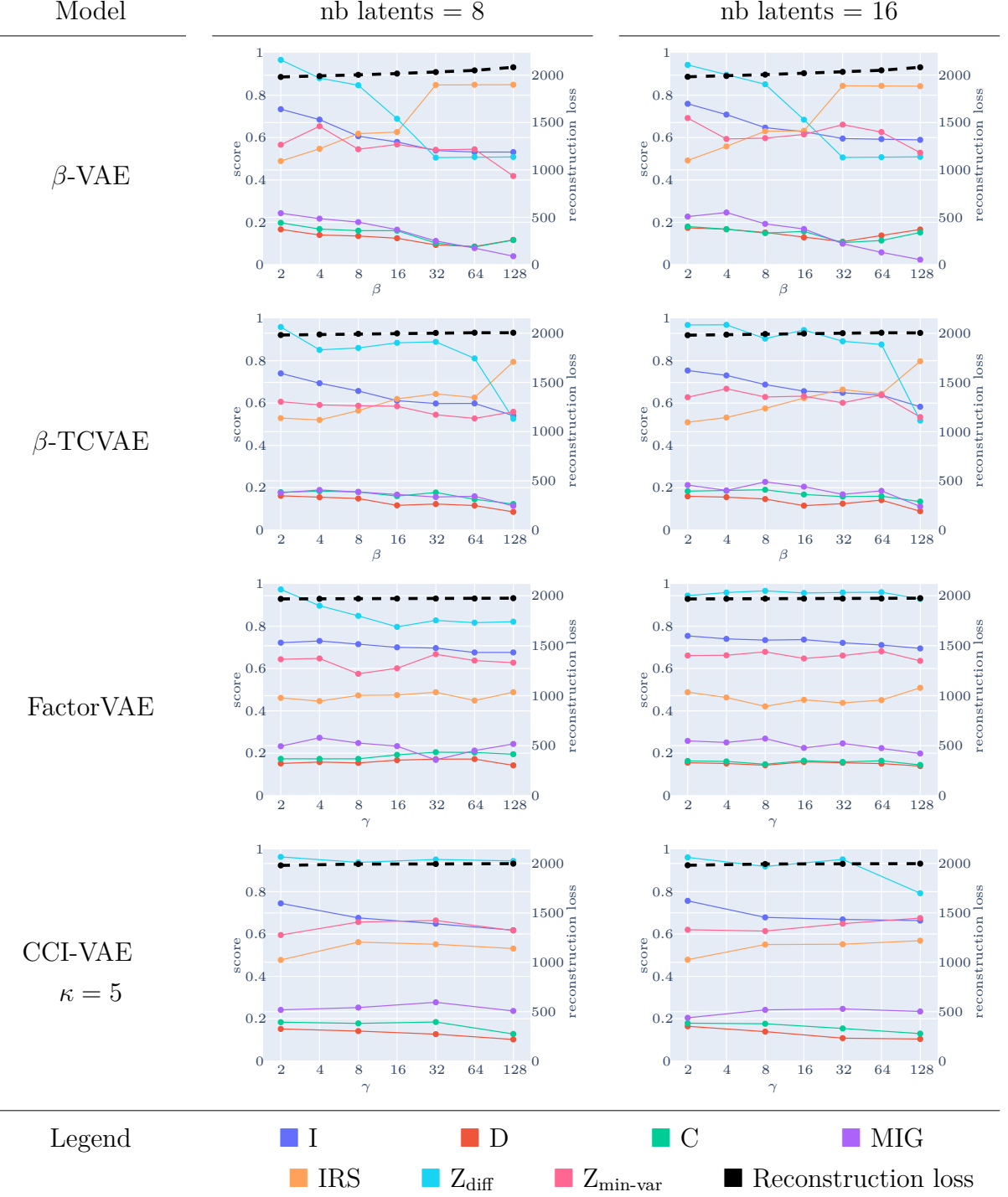



 Table 3.6 – Metrics of β -VAE, β -TCVAE, FactorVAE and CCI-VAE on dSprites

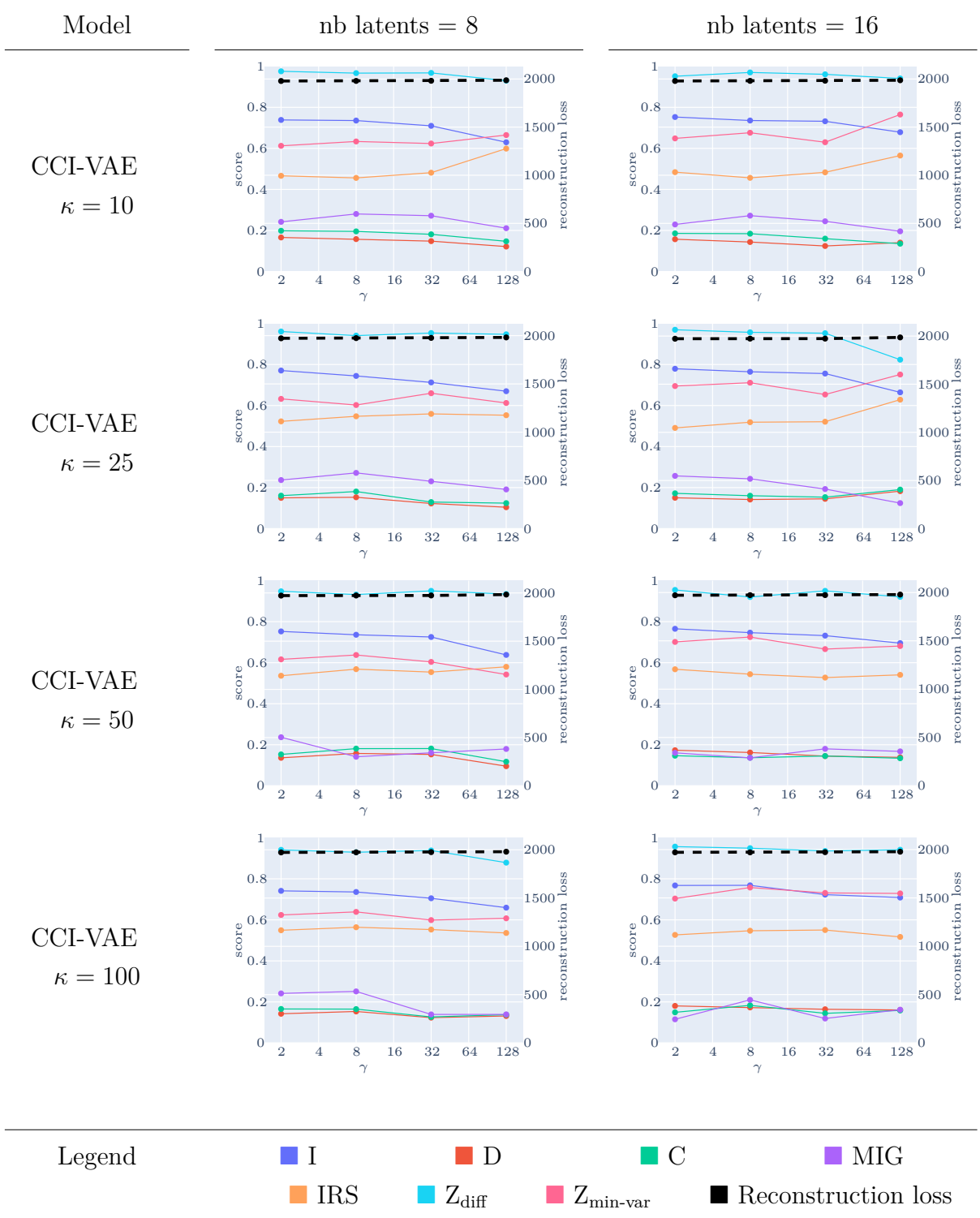



 Table 3.7 – Metrics of β -VAE, β -TCVAE, FactorVAE and CCI-VAE on Cars3D

SmallNORB

Metrics vs β or γ 

SmallNORB

 Metrics vs β or γ

 Table 3.8 – Metrics of β -VAE, β -TCVAE, FactorVAE and CCI-VAE on SmallNORB

At this point, it is interesting to investigate some reconstructions of the learned models, especially to ensure the effect of the information bottleneck pressure controlled by β or γ , and the information capacity controlled by κ in CCI-VAE, as described in Subsection 2.2.3. To illustrate those effects, such reconstructions on Cars3D are displayed in Table 3.9, with increasing β in β -VAE and increasing κ in CCI-VAE with fixed $\gamma = 128$, both with 8 latent dimensions. As expected, increasing β in β -VAE reduces the reconstruction quality. Around $\beta = 32$, the bottleneck pressure becomes too important, and progressively prevents the car color information from being transmitted, and the generated images become noisy. One can also observe that increasing κ in CCI-VAE improves the reconstruction quality, which is consistent with the idea that κ explicitly represents the amount of information authorized to be conveyed by the latent space.

Overall, higher scores than raw VAE are reached on dSprites and Cars3D with the considered variants, which confirms the intuitions of disentanglement enforcement developed in Subsection 2.2.3. No real improvements are found on SmallNORB, which might reveal inappropriate model architecture or tested hyperparameter values. To confirm the improvement of disentanglement announced by the metrics, Table 3.10 shows for each corpus traversals of β -TCVAE with $\beta = 8$ and 8 latents, empirically selected regarding its fair reconstruction/disentanglement trade-off in Table 3.6, Table 3.7 and Table 3.8, to be compared from Table 3.4. First of all, dSprites seems well disentangled: latent 0 controls the scale of the sprite, 2 and 4 the vertical and horizontal positions, respectively. The type of shape and the orientations seem to be conveyed by latents 3, 5, and 6, but they are more fuzzy to distinguish. Subsequently, Cars3D is harder to analyze, but it is clear that some factors are conveyed along dimensions, even not figuring among Cars3D’s ground truth generative factors (illustrated in Figure 2.10). The azimuth is captured by latent 0, while elevation seems to be controlled by latent 1, which also conveys the car orientation (left or right). The color appears to be separated from the concept of car type in latent 7, while also being captured by latents 3 and 5, which also control the car type factor. Other less obvious factors seem to be varying along remaining latents, as the car roundness in latent 2, and the windows darkness in latent 6. This is an interesting behavior that demonstrates the limitation of supervised disentanglement metrics, which may not reflect such factors being captured, not standing among pre-defined generative factors, yet is relevant and insightful. Finally, visualizations are less convincing for SmallNORB. Generated images are noisy, but one can recognize the lighting changing along latent 0. It is quite fuzzy, but the object shape, azimuth, and orientation factors seem to be somehow















β	β -VAE	κ	CCI-VAE
			
2		5	
4		10	
8		25	
16		50	
32		100	
64			
128			

Table 3.9 – Reconstructions on Cars3D of β -VAE , both with 8 latent dimensions. Real image samples are displayed below model names, and the corresponding reconstructions are listed below.

entangled in latents 3, 6 and 7. More generally, smooth variations along latent dimensions can be observed, in some cases disentangling ground-truth generative factors, entangling some of them, or even capturing other relevant variations.

3.1.3 Conclusions

Thanks to the described experiments, one can appraise the disentanglement capacity of VAE-based models, the influence of models’ hyperparameters, and the behavior of disentanglement metrics. Experiments also pointed out the difficulty to draw clear conclusions: which model to choose? in which hyperparameter settings? how to deal with the reconstruction/disentanglement trade-off? how to efficiently ensure that relevant factors are captured, even unsupervised ones? Visualization and traversals are of good help, but one definitely cannot rely on such manual assessments when handling multiple models and complex data, such as speech.

Furthermore, results displayed in Table 3.6, Table 3.7 and Table 3.8 demonstrate disagreements between metrics. Some are far more optimistic (Z_{diff} , $Z_{\text{min-var}}$) than others (MIG, Disentanglement, Completeness), and in a few cases, they exhibit inverse variations. This comforts the discussions of Carbonneau et al. (2021) [21], who also witnessed

Latent	Traversal frames											
	dSprites						Cars3D					
	-2					2	-2					2
0												
1												
2												
3												
4												
5												
6												
7												
	-2					2	SmallNORB					
0												
1												
2												
3												
4												
5												
6												
7												

Table 3.10 – Traversals of β -TCVAE with 8 latent dimensions trained on dSprites, Cars3D and SmallNORB

contradictory behaviors between metrics.

All in all, disentanglement learning is definitely not an easy task to investigate. Models are implicitly enforcing disentanglement, simple data with known factors is still required, and metrics are not trivial to interpret. Thence, to pave the way towards the disentan-

gment of real speech data, a synthetic speech corpus, analogous to the ones used in this section, would be a great help. This is rightly the concern of Section 3.2, which introduces such a synthetic corpus for speech purposes, and some experiments conducted on it.

3.2 diSpeech : a synthetic toy dataset for speech disentanglement

In this section is described *diSpeech* [247], the first synthetic speech dataset intended for speech disentanglement experiments. Therefore, in diSpeech, “di” stands for “disentanglement”. As a first step, this dataset is constrained by synthesizing only vowel waveforms lasting one second.

The purpose of this corpus is to provide a suitable playground for machine learning speech disentanglement models, with known and well-distributed acoustic features. As described in Section 1.1, speech intricacies prevent one to formally define a closed set of distinct perceptual attributes, e.g., quality, prosody, emotion, identity. Acoustic cues are more alienated from perception, but are better descriptors of speech. Consequently, as a first stone to pave the way to the disentanglement of perceptual facets of speech, the introduced corpus uses acoustic features as generative factors, and is based on open source implementations, as described in Subsection 3.2.1. Subsection 3.2.2 depicts disentanglement experiments performed on diSpeech, along with visualizations and analysis results. Towards realistic speech, Subsection 3.2.3 portrays experiments conducted on TIMIT vowel disentanglement with a model trained on TIMIT and inferred on diSpeech. Discussions are provided in Subsection 3.2.4, perspectives in Subsection 3.2.5, and the overall conclusions are presented in Subsection 3.2.6.

3.2.1 Corpus description

The proposed corpus of synthetic vowels diSpeech is composed of five generative factors: the first three formants F1, F2, F3; the fundamental frequency F0; and F0’s fade rate, which is merely refer to as fade. Mel-spectrograms illustrating those attributes are shown in Figure 3.1. The minimum and maximum values used for each factor are displayed in Table 3.11.

To synthesize vowels, Klatt Synthesizer [119] is utilized, which is a formant-based synthesizer that provides a complete set of parameters potentially able to generate quite

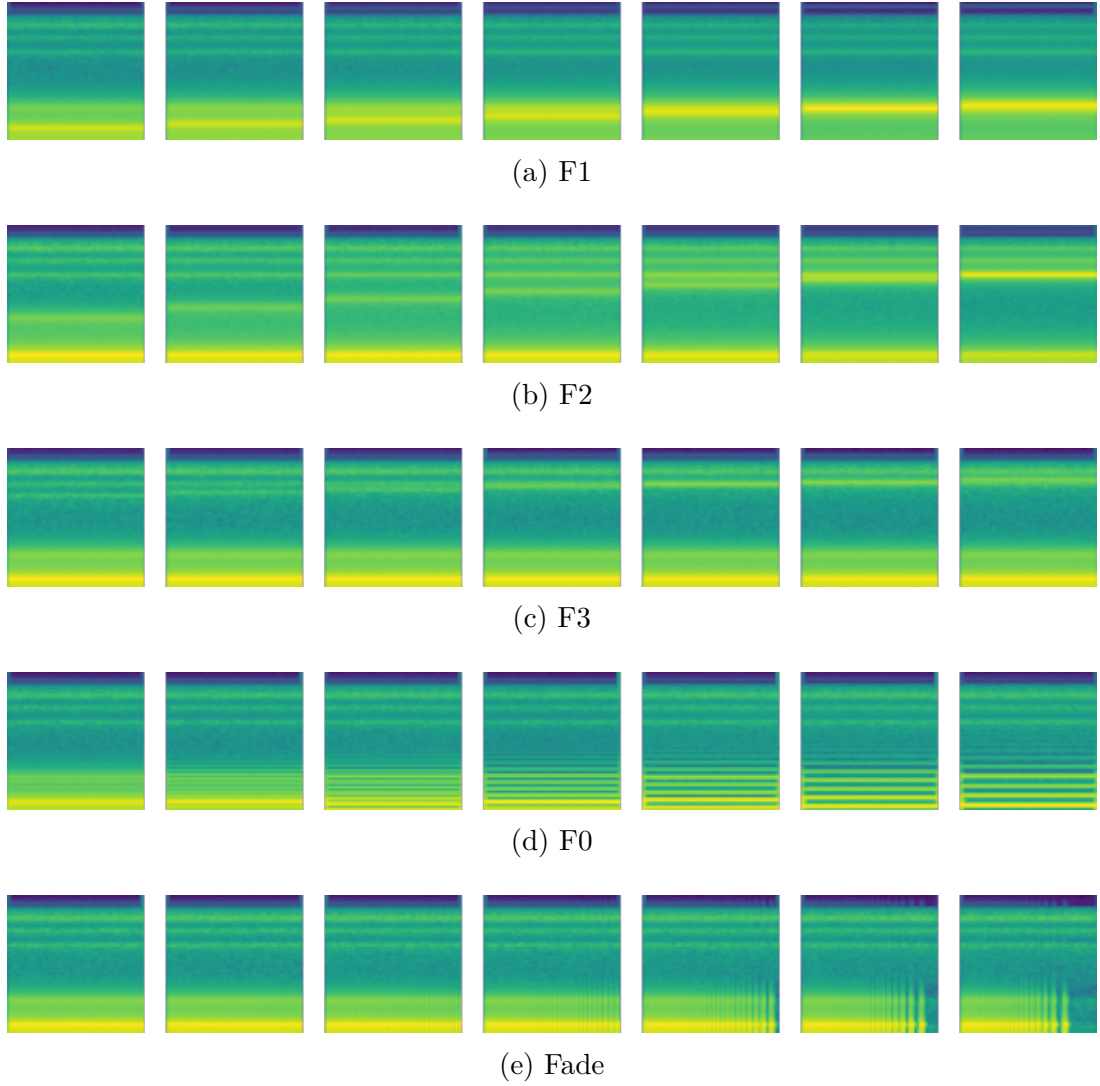


Figure 3.1 – diSpeech factors of variation

	F1	F2	F3	F0	fade
min	275	779	2579	50	0
max	830	2585	3815	200	99

Table 3.11 – Range of values for diSpeech generative factors

realistic speech. More precisely, *tdklatt* is used, an open source Python implementation of Klatt Synthesizer¹.

For the experiments to follow in this manuscript, 15 equally spaced values are dis-

1. <https://github.com/guestdaniel/tdklatt>

cretized for each factor within the ranges defined in Table 3.11. All combinations of the 15 possible values of the 5 factors are used, hence, the total number of samples is $15^5 = 759375$.

The first three formants are employed to properly cover the vowel space. The minimum and maximum values reached by each formant are determined based on [66], as reported in Table 3.11. Since all the combinations of the 15 equally spaced values for each formant are used, and not only the reference vowel formant values from Georgetown et al. (2012) [66], “intermediate” vowels, or “vowel-like” allophones are also generated. For the sake of conciseness, they are simply referred to as vowels in this manuscript.

The fade factor represents the difference between F0’s initial and final values. It also provides temporal variations, leading to more “natural-sounding” vowels. More precisely, the fade factor is a percentage ($\in [0, 99]$), defining the proportion of the initial F0 value the vowel will reach at its end. For instance, F0 can be constant (fade = 0) or linearly decreasing to 50% of its initial value (fade = 50).

Furthermore, the generative factors to be used and their values are explicitly defined here for reproducibility matters. But Klatt synthesizer actually supports a large set of parameters, that can be tuned to generate other vowel variations and, more generally, phonemes or entire sentences. Hence, in the corpus generation code, publicly available on GitHub², there is no constraint on the parameters to tune, enabling any parameter to be considered as a generative factor. It means that diSpeech can (and is intended to) be extended to consonants, words, or sentences and is not limited to vowels.

In order to extract meaningful features from vowel audios, and also to seamlessly run experiments as in Section 3.1 with `disentanglement_lib` [149], the synthesized vowels are processed to obtain inputs homologous to dSprites [157] or Cars3D [184] i.e. 64×64 images. From vowels of 1 second with a sample rate of 16kHz, log mel-spectrograms are computed with 64 mel filters (between 80Hz and 7600Hz), and a FFT of length 1024 and hop length 252. Mel-spectrograms highlighting the variations of each factor are displayed in Figure 3.1, where the value of each factor is increasing from left to right, within their respective range (Table 3.11). Formant variations are shown in Figure 3.1a, Figure 3.1b and Figure 3.1c, F0 in Figure 3.1d and fade in Figure 3.1e.

2. <https://github.com/Orange-OpenSource/diSpeech>

3.2.2 Synthetic vowels disentanglement

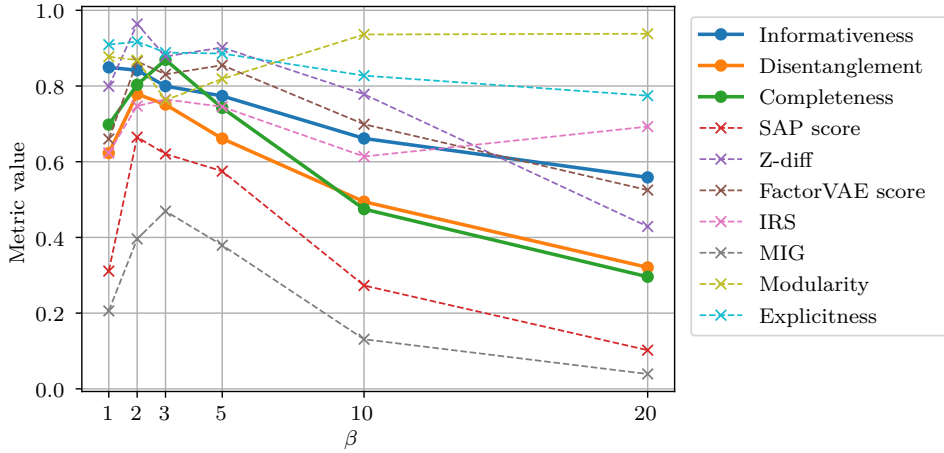
With a new dataset of synthetic vowels in hands, the disentanglement ability of β -VAE [84] can be experimented, the best value of the hyperparameter β estimated, and factors successfully disentangled identified. A β -VAE is trained on diSpeech with several values of $\beta \in \{1, 2, 3, 5, 10, 20\}$, and a latent space of 10 dimensions.

The disentanglement of each model can be qualified with visualization assessments. Table 3.12 on page 114 stores the generated mel-spectrograms through traversals (see Section 2.4) of each dimension for each trained model. Overall, models with $\beta \in \{1, 2, 3, 5, 10\}$ seem to exhibit latent dimensions capturing factor-related variations. F1 can be noticed to be moving up over latent 5 for $\beta = 1$, latent 8 for $\beta = 2$ and 3, latent 0 for $\beta = 10$, and moving down along latent 6 for $\beta = 5$. Similar observations in other latent components can be made for F2 and F3. Fade is observed with decreasing/increasing energy at the end of reconstructions over models with $\beta \in \{1, 2, 3, 5\}$. Gaps between harmonics appear/disappear when traversing some latent dimensions in models with $\beta \in \{1, 2, 3\}$, hence disentangling F0.

Visualizations through traversals show that a too great value of β leads to a less informative learned latent space, which can be interpreted as the posterior collapse pitfall. F0 and fade are the most complex attributes to capture, as they are left out when β reaches 5 and 10, respectively. Otherwise, the generative factors defined in diSpeech seem to be disentangled, unless the proper values of β are used. But it remains hard to deduce the optimal value of β that better addresses the disentanglement/posterior collapse trade-off (see Subsection 2.2.2).

Therefore, in order to objectively assess the best value of β , evaluations are launched with metrics provided by `disentanglement_lib`, namely DCI [55], SAP score [130], Z-diff [84], FactorVAE score [109], IRS [213], MIG [26], and Modularity and Explicitness [188]. Those scores, depending on the value of β , are plotted in Figure 3.2. In particular, the three metrics constituting the DCI (Disentanglement, Completeness and Informativeness) are emphasized in solid lines, as their relevance has been assessed by the literature and previous experiments. At first sight, metrics seem not to agree on the disentanglement quality. SAP score and especially MIG badly rate disentanglement, whereas Explicitness and Modularity are giving good scores. Nevertheless, the overall variations of the metrics indicate that for most of them, higher values are reached for $\beta = 2$.

With a focus on DCI importance matrix in Figure 3.3, a factor-wise analysis can be performed to directly visualize the relative importance of each latent to predict each

Figure 3.2 – β -VAE disentanglement metrics on diSpeech depending on β

factor. The factor-wise analysis can be observed for each value of $\beta \in \{1, 2, 3, 5, 10, 20\}$, and one may notice that accordingly to Figure 3.2, values of β greater than 5 lead to degraded overall disentanglement. As what can be observed in Table 3.12, β -VAE fails to capture F0 and fade with β greater than 5. Overall, fade and F0 seem always entangled together.

With a focus on $\beta = 2$ in Figure 3.3b, it clearly appears that latents 1, 6, and 7 are important to predict F3, F2, and F1, respectively, confirming the insights inferred visually from Table 3.12. Figure 3.3b also shows that F0 and fade are learned but entangled together in latents 2 and 3, which does not clearly appear in traversals. This is, however, not surprising, as both factors are strongly correlated. The fade rate is also introducing temporal variations, hence making it more difficult to model.

This experiment shows that β -VAE successfully achieves disentanglement of formants on diSpeech. They are correctly learned and aligned each on distinct latent axes, as indicated by single peaks in Figure 3.3, while F0 and fade are entangled together. Hence, DCI allows one to objectively identify disentangled/entangled factors and corresponding disentangling/entangling latents. The disentanglement of synthetic vowels being assessed, one may wonder if similar behaviors can be observed with real vowels, which is in the interest of Subsection 3.2.3.

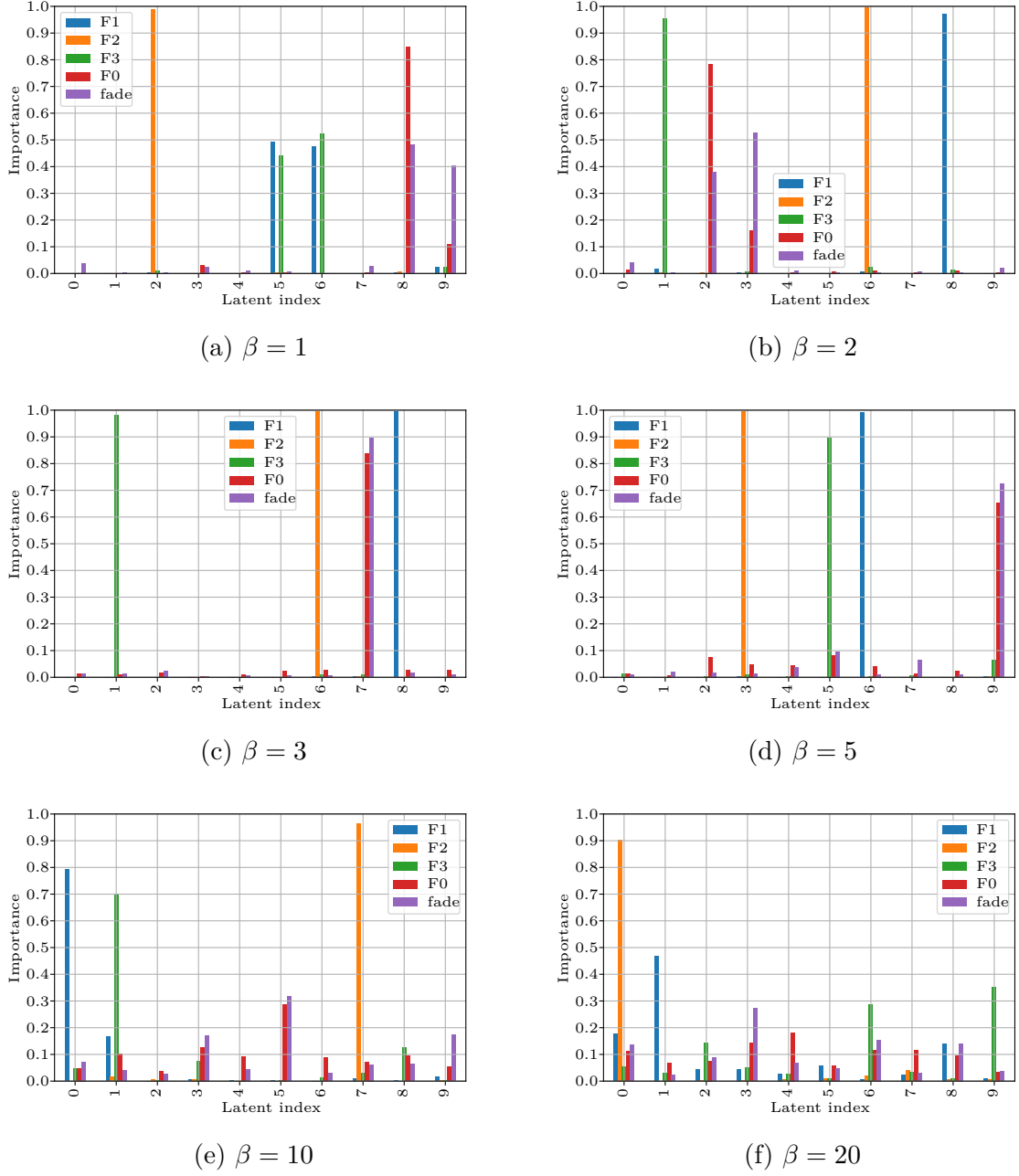


Figure 3.3 – Factor-wise DCI importance matrix analysis of β -VAE trained on diSpeech

3.2.3 Real vowels disentanglement

One main obstacle of real speech disentanglement is the absence of knowledge of generative factors. It is thus proposed to use diSpeech to compute DCI of models trained on real vowels. Hence, β -VAE models are trained on TIMIT’s isolated vowels (always

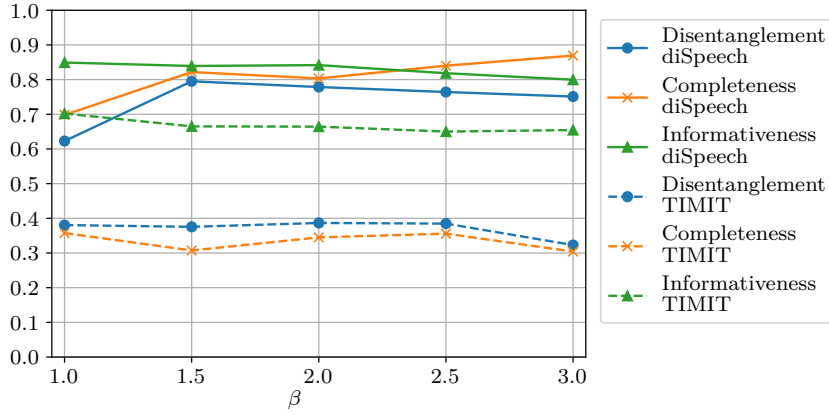


Figure 3.4 – DCI depending on β for β -VAE trained on diSpeech (solid) and on TIMIT’s vowels then inferred on diSpeech (dashed)

with 10 latent dimensions), symmetric-padded to reach 1 second, and preprocessed as in diSpeech to have equivalent inputs. Multiple values of β are tested, focused around 2 ($\beta \in \{1, 1.5, 2, 2.5, 3\}$).

Latent space traversals of each model can be found in Table 3.13. While for all models, some latent dimensions seem to capture relevant information (F0, formants, amplitude), it is mostly hard to clearly discern which variations are captured. Variations also appear to be entangled along some dimensions.

Hence, a better insight can be disclosed by taking a learned β -VAE and encoding diSpeech. One is then able to measure disentanglement learned on TIMIT’s vowels, relative to diSpeech’s factors. Figure 3.4 shows that β -VAEs trained on diSpeech’s reach better DCI values than when trained on TIMIT, which is expected: there is no guarantee that a model trained on TIMIT will be unsupervisedly following diSpeech’s factors, and observing latent traversals on TIMIT’s models in Table 3.13 confirms that they are not well disentangled. But the theoretical Disentanglement score for a totally entangled latent space tends to be 0, as experienced by Carbonneau et al. (2021) [21] in subsection 5.2. Thus, the non-zero DCI values reached by TIMIT’s models suggest a partial disentanglement, as can be observed in Table 3.13: some latents somehow capture variations from single formants, while still being a bit entangled with other factors. TIMIT dataset also includes, by its very nature, far more variations than the five factors defined in diSpeech: prosody, timbre, emotion, and so on. diSpeech’s factors are hence less salient in TIMIT, which explains why they display attenuated scores.

As with diSpeech, a more detailed analysis can be made with DCI importance matrix.

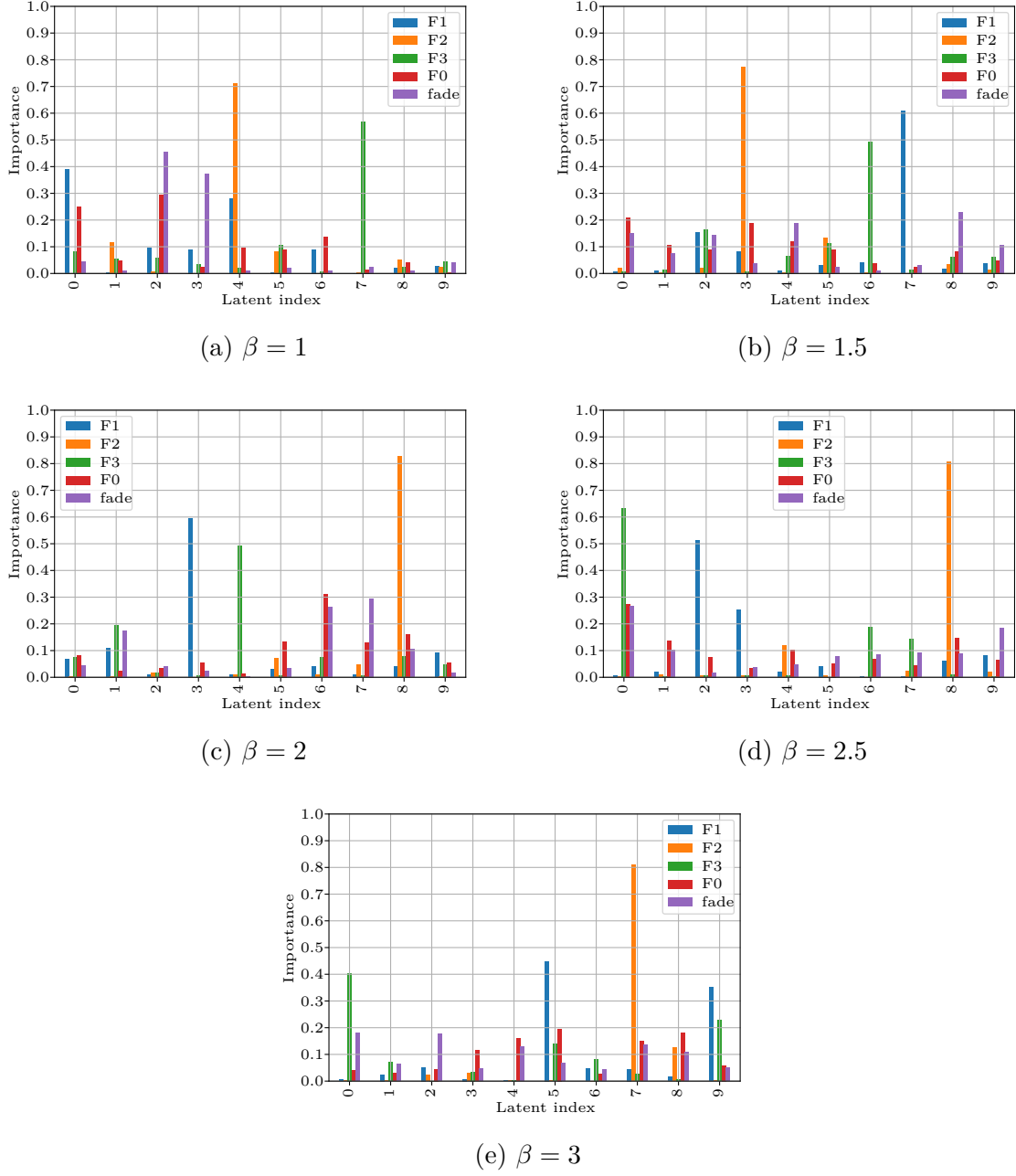


Figure 3.5 – Factor-wise DCI of β -VAE trained on TIMIT

Figure 3.5 shows the resulting factor-wise analysis of the trained β -VAE models. It is confirmed that overall, disentanglement performances are below what was observed with diSpeech. Factors are less well captured, especially F0 and fade with values of $\beta > 1$, while factors are still partially disentangled for $\beta \leq 2$. Overall, F2 seems to be the easiest factor

to disentangle, while other latents tend to be entangled in multiple latent dimensions when β is increasing.

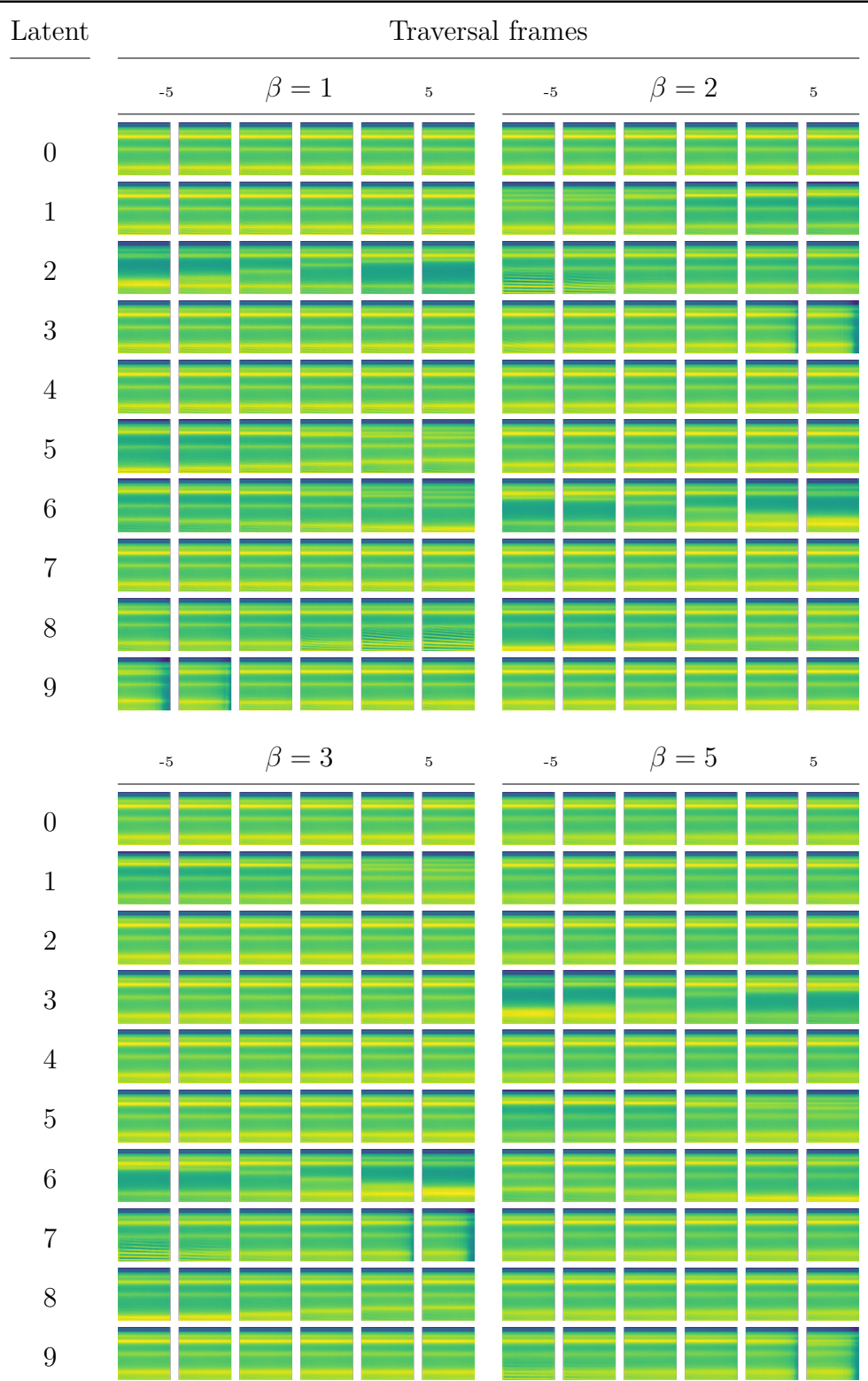
Even if β -VAE does not disentangle perfectly TIMIT’s vowels, relative to the defined factors, diSpeech is shown to allow the computation of disentanglement scores for models trained on real speech, which is unprecedented. One may note that all vowel phonemes of TIMIT’s phonetic transcription were used as extracted samples. Hence, diphthong phonemes are also included during models’ training, increasing training data complexity and divergence with respect to evaluation data. Regarding Table 3.13, diphthong variations seem however to be successfully captured in latents 2 and 3 with $\beta = 1$, disclosing the limitations of using a corpus with pre-defined factors for disentanglement studies.

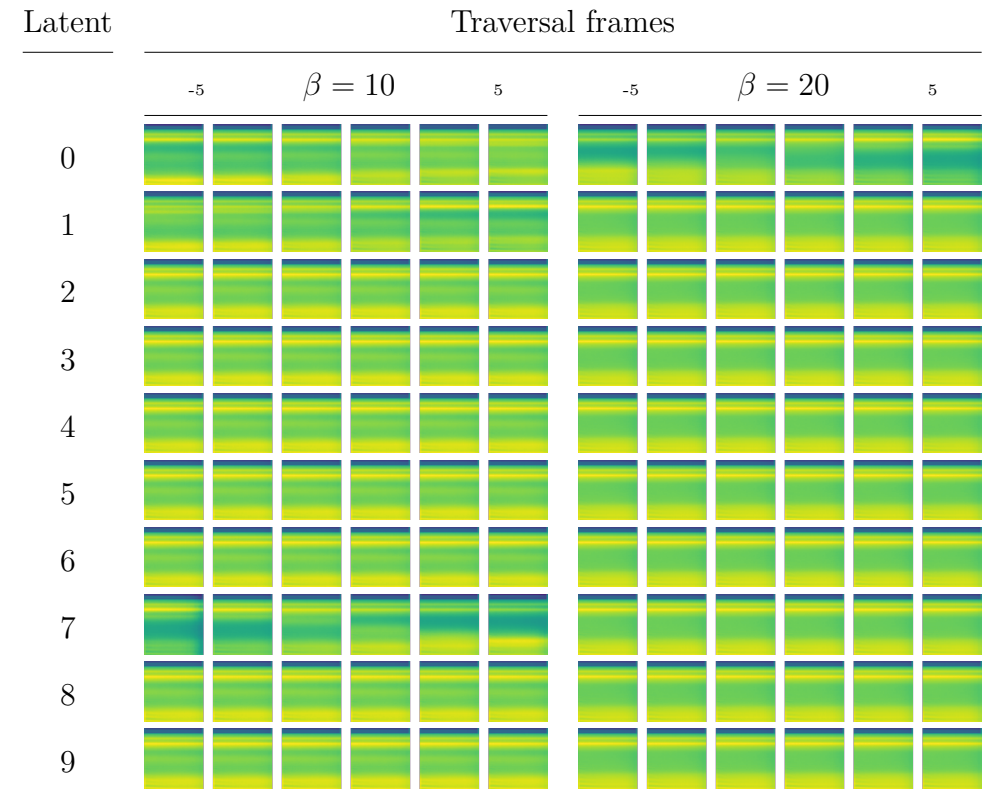
3.2.4 Discussions

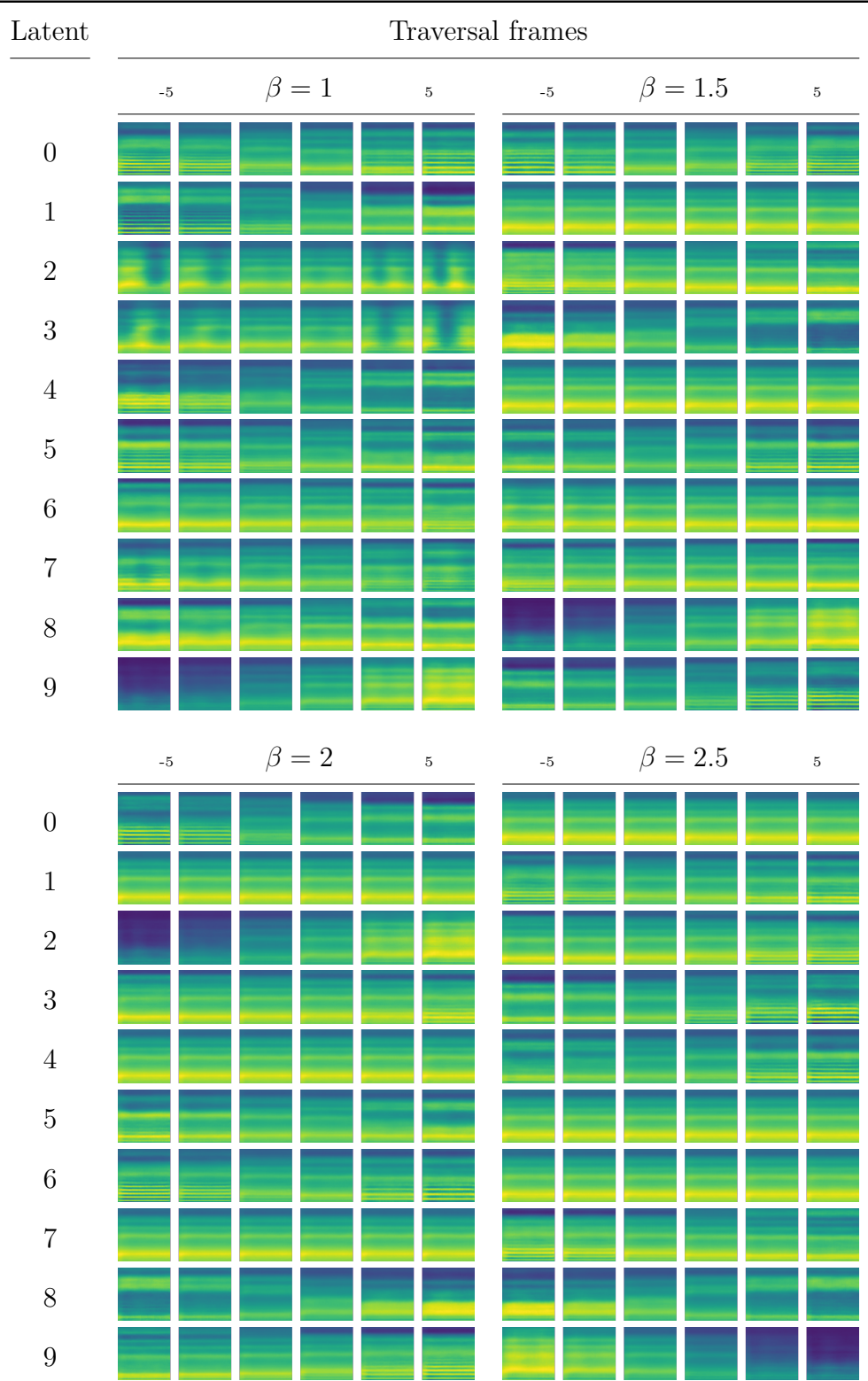
As pointed out in experiment descriptions, β -VAE is not able to achieve a truly efficient disentanglement on TIMIT’s vowels, compared to performances on diSpeech or more generally on other synthetic image corpora. As mentioned in Subsection 3.1.3, speech disentanglement has specific obstacles, due to time dependencies and complex relations between generative factors. Defining a set of perfectly independent factors is already not trivial. On the other hand, assuming independence may lead to exploitable results, as does naive Bayes classifier.

As speech attributes are hard to annotate and subjective, unsupervised disentanglement is a promising approach to automatically extracting relevant and interpretable features for tasks with few annotations. But there is no guarantee that models will align with expected factors, or if one does not set expectations, identifying disentanglement factors is not simple. On top of that, nothing ensures that a latent will learn useful features. It is also noteworthy that nothing prevents factors from being captured by more than one latent (e.g., rotation as angle or sin and cos components). Completeness is hence not an absolute score to blindly follow, but more an indication to carefully interpret. This emphasizes the importance of also monitoring the modularity (Disentanglement in DCI) property of a model, i.e., ensuring that only one factor is captured by each latent. All those elements to be considered reflect the complexity of the analysis of disentanglement models.

Furthermore, experiments were performed with convolution-based β -VAE, which may not have the capacity to handle time dependencies and speech-related complexities previ-




Table 3.12 – Traversals of β -VAE trained on diSpeech



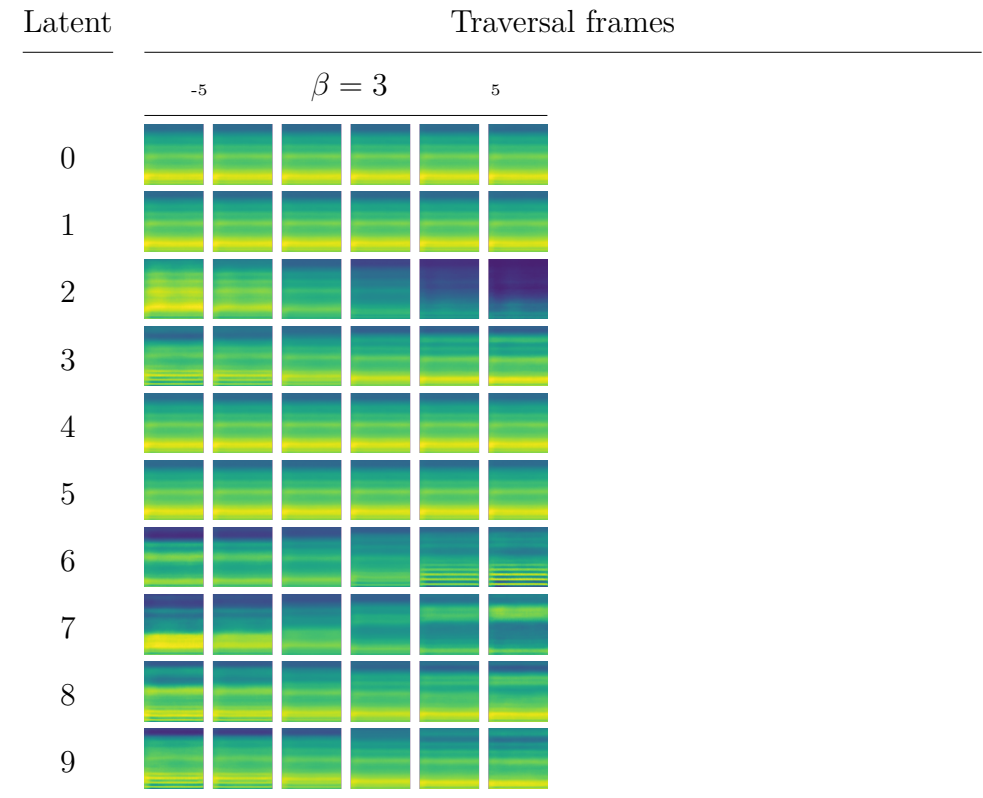


Table 3.13 – Traversals of β -VAE trained on TIMIT

ously mentioned. Leveraging more advanced models, such as those described in Chapter 2 and autoregressive models, would lead to better results.

3.2.5 Perspectives

Conducted experiments show that generative factors of synthetic vowels can be partly disentangled with a β -VAE. But it also appears that transfer to evaluate TIMIT real vowel disentanglement is complicated.

Extending diSpeech towards more realistic content, with other phonemes (e.g., consonants), variable durations, combinations of phonemes, and so on, would hopefully lead to a more reliable disentanglement evaluation on real data, a wider coverage of speech factors, and a fairer approximation of speech complexities.

As the disentanglement process is unsupervised, and hence data-driven, the generative factors can be well disentangled or not, depending on the way they appear in the input audio features. Choosing MFCC instead of mel-spectrograms may have an influence on

disentanglement performances. Similarly, the evaluation of the reconstruction error could be modified to better reflect the similarity between spectrograms (e.g., MCD [120]).

Hierarchical dependencies issues inherent to speech could be addressed by variants of β -VAE, such as VQ-VAE, AnnealedVAE, NVAE or novel strategies parallelizing “multi- β s”. Note also that a well-defined latent space, for instance, inspired by Poincaré embeddings [160], could be helpful.

3.2.6 Conclusion

In this section, a new corpus of synthetic phonemes called diSpeech has been presented. It has been designed to study disentanglement of voice attributes. Its first declination relies on synthetic vowels, parameterized by the fundamental frequency, its fade rate, and the first three formants. It has been used in disentanglement experiments based on β -VAE model.

It results in a clear disentanglement of formants, whereas the remaining two factors stay partly entangled, emphasizing the influence of the nature of a generative factor on its disentanglement.

diSpeech paves the way towards disentanglement evaluation on real speech, as shown in experiments on TIMIT’s vowels in Subsection 3.2.3. Forthcoming studies and improvements of the corpus and methodology have finally been proposed.

3.3 Real speech disentanglement

The disentanglement ability of VAE framework and its variations have been demonstrated on corpora of synthetic images in Section 3.1, and on a corpus of synthetic vowels in Section 3.2. This section steps further towards real speech disentanglement, by leveraging a more advanced VAE-based model: FHVAE [92], trained on TIMIT [64] and Bref120 [133] corpora.

While encouraging results have been found with synthetic images and vowels, it is uncertain if similar behaviors can be obtained from real speech data. As one has no access to true generative factors of variations, one can only rely on available annotations (e.g., speaker identity, gender, accent) to employ supervised metrics. One can also leverage traversals, but repetitive listening tests along each latent dimension are a tedious task, and are prone to semantic satiation, i.e., mental tiredness, negating the reliability of

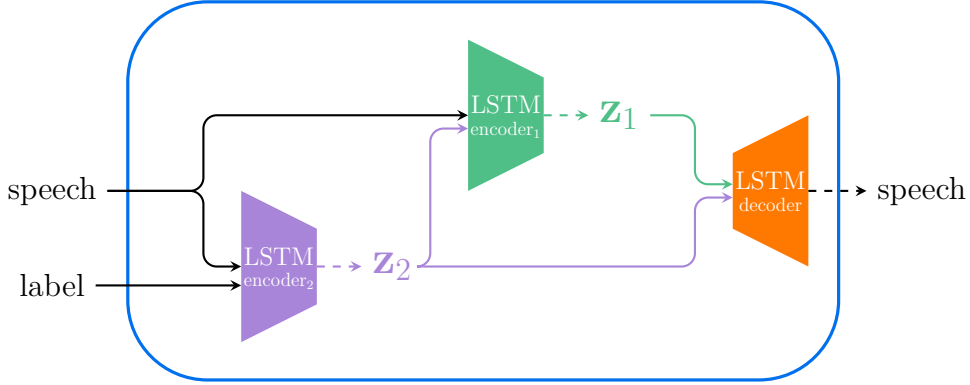


Figure 3.6 – FHVAE architecture

traversal evaluations. Hence, Subsection 3.3.1 describes the leveraged model framework FHVAE. Subsection 3.3.2 succinctly details the training corpora TIMIT and Bref120, Subsection 3.3.3 provides the training setup, and Subsection 3.3.4 the experiments performed with the supervised metric DCI [55] over available annotations.

3.3.1 Factorized Hierarchical VAE (FHVAE)

Among the few attempts to disentangle speech attributes in an self-supervised fashion, Factorized Hierarchical VAE (FHVAE) is a promising yet simple framework, as illustrated in Figure 3.6. The information flow is factorized into 2 latent spaces, \mathbf{z}_1 and \mathbf{z}_2 , in a hierarchical manner, as \mathbf{z}_2 models utterance level variations, and the prediction of \mathbf{z}_1 is conditioned on \mathbf{z}_2 to let \mathbf{z}_1 ignore global features and capture finer-grained information.

In the FHVAE framework, utterances are considered as **sequences** of 200ms **segments**. Let a dataset \mathbb{X} of d utterances X , i.e., $\mathbb{X} = \{X^{(i)}\}_{i \in \{1, \dots, d\}}$, each utterance being split into segments x , i.e., $X^{(i)} = \{x^{(i,j)}\}_{j \in \{1, \dots, N_i\}}$, N_i being the total number of segments in sequence $X^{(i)}$. In other words, let $x^{(i,j)}$ the j -th segment among N_i of sequence $X^{(i)}$, the latter being the i -th utterance among the d of dataset \mathbb{X} . The set of training data is composed of shuffling segments x , regardless of the originated sequence. Segments are preprocessed to n -dimensional mel-spectrograms. The model is hence optimized with batches of segments, which provides a very scalable framework for substantial datasets and long utterances. As in VAE formulation, segments x are supposed to follow a distribution $p_\theta(\mathbf{x})$, and it is assumed that their generative process involves m -dimensional latent variables \mathbf{z}_1 and \mathbf{z}_2 , such that the likelihood $p_\theta(\mathbf{x}|\mathbf{z}_1, \mathbf{z}_2)$ is a multivariate Gaussian distribution (3.1), with mean $\boldsymbol{\mu}_{\mathbf{x}}$ and covariance diagonal $\boldsymbol{\sigma}_{\mathbf{x}}^2$ conditioned on latent vari-

ables \mathbf{z}_1 and \mathbf{z}_2 i.e. predicted from learned decoder neural networks $f_{\mu_{\mathbf{x}}}$ and $f_{\sigma_{\mathbf{x}}^2}$. Latent \mathbf{z}_1 is assumed to model segmental variations, and follows a centered Gaussian distribution (3.2), parameterized by a covariance diagonal $\sigma_{\mathbf{z}_1}^2$, being a hyperparameter. Concerning \mathbf{z}_2 , it is also assumed to follow a Gaussian prior distribution $p_{\theta}(\mathbf{z}_2|\mu_2)$ (3.3), parameterized by a hyperparameter $\sigma_{\mathbf{z}_2}^2$ and conditioned on μ_2 , which itself follows a centered Gaussian prior (3.4) parameterized by $\sigma_{\mu_2}^2$. μ_2 is intended to be conditioned on a sequence index i , which identifies the sequence $X^{(i)}$ pertaining to the considered segment x , as it will be explained further. In the forthcoming experiments, covariance diagonal hyperparameters are defined as $\sigma_{\mathbf{z}_1}^2 = 1$, $\sigma_{\mathbf{z}_2}^2 = 0.25$ and $\sigma_{\mu_2}^2 = 1$, similarly to Hsu et al.'s [92] experiments.

$$p_{\theta}(\mathbf{x}|\mathbf{z}_1, \mathbf{z}_2) = \mathcal{N}(\mathbf{x}; f_{\mu_{\mathbf{x}}}(\mathbf{z}_1, \mathbf{z}_2), f_{\sigma_{\mathbf{x}}^2}(\mathbf{z}_1, \mathbf{z}_2) * \mathbf{I}_m) \quad (3.1)$$

$$p_{\theta}(\mathbf{z}_1) = \mathcal{N}(\mathbf{z}_1; \mathbf{0}, \sigma_{\mathbf{z}_1}^2 * \mathbf{I}_m) \quad (3.2)$$

$$p_{\theta}(\mathbf{z}_2|\mu_2) = \mathcal{N}(\mathbf{z}_2; \mu_2, \sigma_{\mathbf{z}_2}^2 * \mathbf{I}_m) \quad (3.3)$$

$$p_{\theta}(\mu_2) = \mathcal{N}(\mu_2; \mathbf{0}, \sigma_{\mu_2}^2 * \mathbf{I}_m) \quad (3.4)$$

Following variational inference methodology (see Subsection 2.2.1), true posterior distributions are approximated with parameterized multivariate Gaussian distributions. \mathbf{z}_1 is conditioned on input segment \mathbf{x} and latent variable \mathbf{z}_2 (3.5). This dependence between \mathbf{z}_1 and \mathbf{z}_2 induces a hierarchical structure of the latent spaces, intended to model the multi-scale nature of speech attributes, i.e., explicitly separate utterance level and frame level variations. Posterior parameters $\mu_{\mathbf{z}_1}$ and $\sigma_{\mathbf{z}_1}^2$ are predicted from encoder neural networks $g_{\mu_{\mathbf{z}_1}}$ and $g_{\sigma_{\mathbf{z}_1}^2}$. The second latent \mathbf{z}_2 is similarly approximated by a Gaussian distribution (3.6), which is parameterized by encoder neural networks $g_{\mu_{\mathbf{z}_2}}$ and $g_{\sigma_{\mathbf{z}_2}^2}$ trained to predict mean $\mu_{\mathbf{z}_2}$ and variance $\sigma_{\mathbf{z}_2}^2$ given an input segment \mathbf{x} . μ_2 's surrogate posterior distribution is a Gaussian distribution (3.7), parameterized by a fixed covariance diagonal $\sigma_{\mu_2}^2$ and a learned deterministic function $g_{\mu_{\mu_2}}$ which maps each utterance index i to a mean vector $\tilde{\mu}_2$.

$$q_{\phi}(\mathbf{z}_1|\mathbf{x}, \mathbf{z}_2) = \mathcal{N}(\mathbf{z}_1; g_{\mu_{\mathbf{z}_1}}(\mathbf{x}, \mathbf{z}_2), g_{\sigma_{\mathbf{z}_1}^2}(\mathbf{x}, \mathbf{z}_2) * \mathbf{I}_m) \quad (3.5)$$

$$q_{\phi}(\mathbf{z}_2|\mathbf{x}) = \mathcal{N}(\mathbf{z}_2; g_{\mu_{\mathbf{z}_2}}(\mathbf{x}), g_{\sigma_{\mathbf{z}_2}^2}(\mathbf{x}) * \mathbf{I}_m) \quad (3.6)$$

$$q_{\phi}(\mu_2|i) = \mathcal{N}(\mu_2; g_{\mu_{\mu_2}}(i), \sigma_{\mu_2}^2 * \mathbf{I}_m) \quad (3.7)$$

A crucial point is the role of μ_2 , that is, to introduce the knowledge of which utterance $X^{(i)}$ the segment \mathbf{x} is coming from. Hence, \mathbf{z}_2 is encouraged to learn utterance-related

variations. Furthermore, Hsu et al.’s implementation directly considers the posterior mean $\tilde{\boldsymbol{\mu}}_2$ as the prior mean $\boldsymbol{\mu}_2$ without sampling from the posterior $q_\phi(\boldsymbol{\mu}_2|i)$, i.e., $\boldsymbol{\mu}_2^{(i)} = g_{\boldsymbol{\mu}_{\mu_2}}(i) = \tilde{\boldsymbol{\mu}}_2^{(i)}$, which is equivalent to setting a null covariance $\boldsymbol{\sigma}_{\tilde{\boldsymbol{\mu}}_2}^2 = \mathbf{0}$. The posterior approximation $q_\phi(\boldsymbol{\mu}_2|i)$ becomes a shifted Dirac-delta function $\delta(\boldsymbol{\mu}_2 - \tilde{\boldsymbol{\mu}}_2)$. In other words, $q_\phi(\boldsymbol{\mu}_2|i)$ is a trainable lookup table of prior means $\tilde{\boldsymbol{\mu}}_2$, one for each sequence.

All in all, encoder functions $g_{\boldsymbol{\mu}_{\mathbf{z}_1}}$, $g_{\boldsymbol{\sigma}_{\mathbf{z}_1}^2}$, $g_{\boldsymbol{\mu}_{\mathbf{z}_2}}$ and $g_{\boldsymbol{\sigma}_{\mathbf{z}_2}^2}$; and decoder functions $f_{\boldsymbol{\mu}_{\mathbf{x}}}$ and $f_{\boldsymbol{\sigma}_{\mathbf{x}}^2}$; are LSTM-based neural networks, as specified in Figure 3.6. The utterance index i can be seen as a kind of unsupervised label, but one may design a very similar system where the lookup table $g_{\boldsymbol{\mu}_{\mu_2}}$ is indexed following a given annotation (e.g., speaker identity, gender), leading to a latent \mathbf{z}_2 conditioned on a supervised label, hence trained to capture its variations. The ELBO to maximize deriving from this architecture has the following form [92]:

$$\begin{aligned} \mathcal{L}_{\text{FHVAE}}(\theta, \phi; \mathbf{x}, i) = & \mathbb{E}_{q_\phi(\mathbf{z}_1, \mathbf{z}_2|\mathbf{x})} \left[\log p_\theta(\mathbf{x}|\mathbf{z}_1, \mathbf{z}_2) \right] \\ & - \mathbb{E}_{q_\phi(\mathbf{z}_2|\mathbf{x})} \left[D_{KL}(q_\phi(\mathbf{z}_1|\mathbf{x}, \mathbf{z}_2) \| p_\theta(\mathbf{z}_1)) \right] \\ & - D_{KL}(q_\phi(\mathbf{z}_2|\mathbf{x}) \| p_\theta(\mathbf{z}_2|\boldsymbol{\mu}_2^{(i)})) \\ & + \frac{1}{N_i} \log p_\theta(\boldsymbol{\mu}_2^{(i)}). \end{aligned} \quad (3.8)$$

With this framework, Hsu et al. successfully performed speaker verification, voice conversion and denoising by manipulating the sequence level latent \mathbf{z}_2 , and observed smooth transformations when traversing dimensions of \mathbf{z}_1 (formants, phonetic cues) and \mathbf{z}_2 (pitch, formant ranges). FHVAE thus exhibits the potential to disentangle real speech factors, especially in \mathbf{z}_2 for utterance level annotations, which is the hypothesis stressed in Subsection 3.3.4 with the training data described in Subsection 3.3.2.

3.3.2 Training data

Experiments with FHVAE have been conducted on Bref120 [133] and TIMIT [64] datasets. Bref120 is a 236-hour corpus of French reading speech with over 120 speakers (55 males and 65 females), with around 50 to 60 sentences per speaker. Two recording channels are available, with one clearly noisier than the other. TIMIT is a very widely used speech dataset consisting of 630 speakers of American English, which includes both read and spontaneous speech.

Corpus	Labels	Classes
Bref120	channel	clean, noisy
	speaker identity	120
	gender	male, female
	height	[1.50, 1.92]
	weight	[38, 90]
	education	junior high school, trade school, high school degree, 2 year university, 4 year university or more, unknown
	smoker	yes, no
	age	[17, 68]
	native language	German-Russian, Arabic, Spanish, French, Luxembourger, Portuguese
TIMIT	gender	male, female
	speaker identity	630
	region	New England, Northen, North Midland, South Midland, Southern, New York City, Western, army brat ³
	age	[21,75]
	height	[1.45, 1.98]
	ethnicity	white, black, American Indian, Spanish-American, oriental, unknown
	education	high school, associate degree, bachelor's degree, master's degree, PhD, unknown

Table 3.14 – Bref120 and TIMIT leveraged annotations

Both are supplying annotations, detailed in Table 3.14. Bref120 comprises nine labels: channel type, speaker identity, gender, height, weight, education level, smoker, age, and native language. In the described experiments, the native language label has been ignored, as it appears to be highly imbalanced ($\approx 90\%$ of French natives). TIMIT contains seven labels: gender, speaker identity, region, age, height, ethnicity, and education level.

A great diversity of annotations is available, among which many are categorical ones (e.g., speaker identity, channel, education), while others are treated as continuous variables (e.g., height, weight, age). They are all considered constant across utterances and speakers, hence one can expect FHVAE’s latent \mathbf{z}_2 to be more susceptible to capture those variations. Experiments described in Subsection 3.3.4 assess if some of them are disentangled by FHVAE, thanks to DCI [55] metric. However, for some labels, it is quite uncertain that they can be efficiently conveyed and retrieved from speech, e.g., education, height, and weight. It is hence unlikely for FHVAE to disentangle such outsider factors, but they are nonetheless leveraged in experiments. Thus, salient factors are expected to be disentangled: channel type, gender, age, and regional accent in TIMIT.

3.3.3 Training setup

The same training protocol described by Hsu et al. (2017) [92] is followed in the experiments depicted in this manuscript. Encoder and decoder LSTM layers are composed of 256 hidden units, and are followed by fully connected layers to predict means and variances of variational distributions. Latents \mathbf{z}_1 and \mathbf{z}_2 are both 32-dimensional. Adam optimizer is leveraged, with $\beta_1 = 0.95$, $\beta_2 = 0.999$, $\epsilon = 1e - 8$, and learning rate set to $3e - 4$. Speech utterances are sampled at 16kHz then preprocessed into mel-spectrograms with 400 FFT frequencies, a hop length of 160, and 80 mel-coefficients. Utterances are segmented into sequences of 200ms segments. As each mel-spectrogram frame temporally conveys 10ms, each segment is an 80×20 matrix. Batches of such segments of size 256 are constituted for training, for 100 epochs.

With the discussions provided in Subsection 2.2.3 about the trade-off between reconstruction and disentanglement, controlling the information bottleneck was considered during the experiments. To this end, the very same idea advanced by Rybking et al. (2021) [191] with σ -VAE, as illustrated in (2.19), is adopted. Hence, multiple values of decoder covariance diagonal $\sigma_{\mathbf{x}}^2$ are tested: 1, 0.1, and 0.01. The basic setting where it is predicted from a fully connected layer is also retained, i.e., $\sigma_{\mathbf{x}}^2 = f_{\sigma_{\mathbf{x}}^2}(\mathbf{z}_1, \mathbf{z}_2)$. To this extent, it is expected to observe better disentanglement performances with a higher $\sigma_{\mathbf{x}}^2$.

3. Army brat are children of military personnel, known to frequently move from a region to another.

3.3.4 Evaluations

As extensively discussed in Section 2.4, measuring disentanglement is to this day still only feasible when the generative factors are known. Some unsupervised metrics have been proposed, but pertain to computer vision concerns, or require legions of trained models to be reliable (UDR). Hence, to evaluate FHVAE disentanglement on real speech data, the supervised metric DCI has been used to measure the disentanglement of available annotations of Bref120 and TIMIT. To get an insight into label disentanglement along models training, Table 3.15 and Table 3.16 contain the Completeness of annotations, depending on the training epoch, for both latents \mathbf{z}_1 and \mathbf{z}_2 , and for the multiple decoder covariance $\sigma_{\mathbf{x}}^2$ settings, as described in Subsection 3.3.3. The Completeness is previously argued in Subsection 3.2.4 to be hard to interpret, as several latent dimensions might capture a same factor, but it is believed that Completeness remains a useful indicator and efficient criteria to oversee the behavior of a model regarding multiple factors to disentangle. The log likelihood $p_{\theta}(\mathbf{x}|\mathbf{z}_1, \mathbf{z}_2)$ on the validation set, along epochs, is also plotted, which corresponds to the negative reconstruction error. Hence, the higher the log likelihood, the better the reconstruction quality.

In Bref120, the channel type is clearly well disentangled regarding Table 3.15, by \mathbf{z}_1 in early stages, and by \mathbf{z}_2 in further steps. Subsequently, gender, height, and weight labels reach Completeness around 0.5, which uncovers the correlations between those annotations: height and weight are likely to be correlated with one’s gender. With this bias in mind, it is unlikely that FHVAE actually learned variations truly related to one’s height or weight. Consequently, channel type and gender let aside, poor disentanglement is achieved with the remaining labels. The disentanglement/reconstruction trade-off seems to be well balanced for $\sigma_{\mathbf{x}}^2 = 0.1$: Completeness scores remain stable at the end of the training, while achieving similar log likelihood (around -30) than predicted covariance $f_{\sigma_{\mathbf{x}}^2}(\mathbf{z}_1, \mathbf{z}_2)$. Other values of $\sigma_{\mathbf{x}}^2$ exhibit degraded reconstruction performances.

Concerning TIMIT, very poor results can be observed in Table 3.16. Latent \mathbf{z}_2 manages to capture gender and height annotations, but as with Bref120, this might result from correlation between gender and height. Here again, a good disentanglement/reconstruction (likelihood around -40) trade-off is attained with $\sigma_{\mathbf{x}}^2 = 0.1$, although the Completeness score remains low.

Overall, it stands out that annotation information is captured by \mathbf{z}_1 in early training stages, and tends to be conveyed by \mathbf{z}_2 in further training steps. As \mathbf{z}_2 is conditioned to learn sequence level variations, this is the expected behavior. However, Completeness

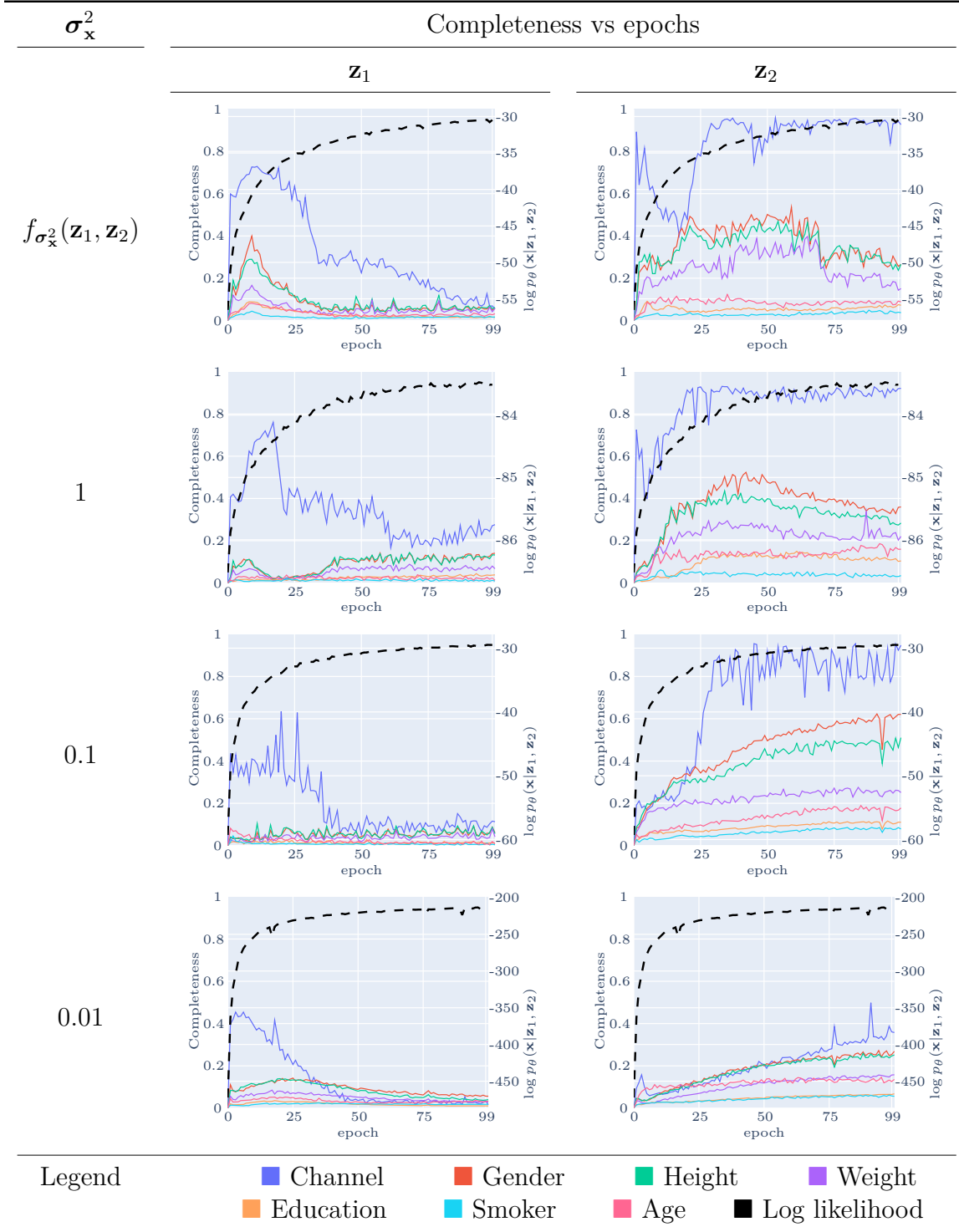


Table 3.15 – Factor-wise Completeness depending on the training epoch of FHVAE, on Bref120

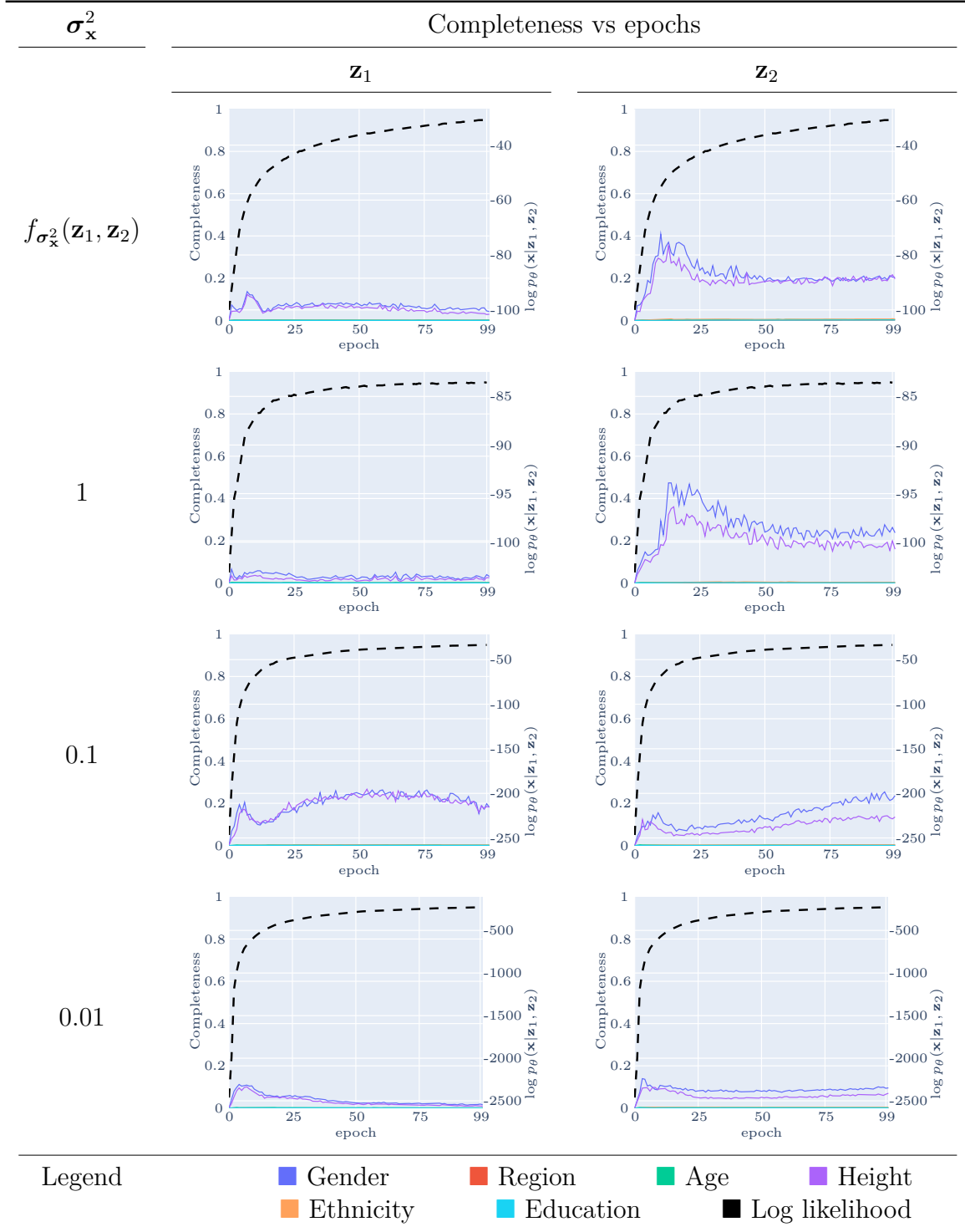


Table 3.16 – Factor-wise Completeness depending on the training epoch of FHVAE, on TIMIT

scores remain poor for most factors, especially on TIMIT. When some label Completeness scores reach values above 0.2, it appears that it may be related to a bias with the gender attribute, i.e., height and weight Completeness seem to be closely correlated with gender. As DCI does not consider relations between factors, and assumes them to be independent, Completeness can be misleading, as observed in those experiments. Furthermore, higher Completeness scores are reached with higher $\sigma_{\mathbf{x}}^2$, which comforts the intuitions developed in Subsection 2.2.3.

3.4 Conclusions and discussions

Along this chapter, disentanglement learning has been studied, from preliminary works on synthetic images in Section 3.1, through similar experiments on synthetic vowels in Section 3.2, to end with investigations towards real speech disentanglement in Section 3.3.

Leveraging `disentanglement_lib`, experiments on synthetic images with the models described in Section 2.2 have been performed, and analyzed with the metrics detailed in Section 2.4. Intuitions about model performances, disentanglement/reconstruction trade-off and hyperparameter aftermath have been provided through reconstructions, traversals, and graphs of metrics and reconstruction loss.

With those insights, similar experiments have been described on an introduced corpus of synthetic vowels, diSpeech. It has been demonstrated that a β -VAE model can successfully disentangle formants, but struggles to deal with temporal variations of pitch. An attempt to disentangle TIMIT’s vowels has been performed, employing a β -VAE model trained on TIMIT and inferred on diSpeech.

The proposed synthetic playground, diSpeech, is a first step to bridge the gap between disentanglement of synthetic images and real speech data. With a more advanced model, FHVAE, experiments have been conducted to assess the disentanglement of annotated factors on real speech corpora, Bref120 and TIMIT, though with reserved results. This highlights the exploratory state of this research direction, and the complexity of transferring theoretical advancements in disentangling models to real-life cases. It also discloses the lack of a well-acknowledged taxonomy of speech attributes, as discussed in Section 1.1, which prevents one from relying on clear reference points: what constitutes speech is still unclear, and interactions between voice characteristics are intricate and complex to deal with, especially for disentanglement purposes. More generally, assessing the disentanglement of speech data appears to be a complex and tedious task, in its current state.

Experiments also revealed that, in addition to requiring knowledge of true generative factors, disentanglement metrics can be convoluted to read and interpret. With this concern in mind, Chapter 4 digs further into inconsistent behaviors experienced with DCI metric, and proposes to address them with a Mutual Information (MI)-based importance matrix. A Partial Information Decomposition (PID)-based framework is also advanced, which is believed to better coincide with the disentanglement desiderata stated in Section 2.1.

MEASURING DISENTANGLEMENT

Assessing disentanglement is still an open problem, as discussed in Section 2.4. Depending on the adopted definition of “what is a disentangled representation”, a wide spectrum of metrics have been proposed and challenged [21, 149]. Furthermore, most related studies handle image disentanglement, and conveniently use synthetic image datasets (Section 2.3) as true factors of variations are known, which is required to measure disentanglement. Concerning speech, the toy dataset diSpeech [247] introduced in Section 3.2 is for now the only available analogous dataset. The described study is basically tied to such a synthetic corpus, as using a realistic voice dataset implies relying on annotated attributes (and not true independent factors of variations) which may not be exhaustive or well-balanced enough to assess metric reliability. The experiments described in Section 3.3 try to disentangle such annotations on Bref120 and TIMIT, with mixed results and conclusions.

It has been demonstrated in Chapter 3 that metrics can uncover interesting behaviors within disentangling models. Nevertheless, considering a single disentanglement score for a model remains too high-level to truly disclose hidden disentanglement-related behaviors. As concluded by Carbonneau et al. (2021) [21], metrics should be considered for each factor separately, to gain a better perception of the performances of a given model. Therefore, a deeper analysis is described in Section 4.1, based on DCI [55], to extend one’s interpretation of latent / factor relations in a β -TCVAE trained on diSpeech. A *latent decimation* process is then proposed, for disentanglement analysis. Applied to diSpeech, it reveals misleading outcomes of the existing metrics in some situations. These observations finally lead in Section 4.2 to an alternative way to compute DCI, inspired by MIG [26], ending up to Mutual Information-based DCI (MIDCI). Experiments are only presented on diSpeech, but it is worth noting that same results were obtained and assessed on the various synthetic datasets and with the wide range of models implemented in `disentanglement_lib`.

Furthermore, Section 4.3 steps further in the definition of disentanglement metrics,

by proposing a PID-based measure of completeness, which has the benefit of taking into account the intricate inter-latent and inter-factor relationships.

4.1 Latent decimation for metric consistency

The review of disentanglement metrics proposed by Carbonneau et al. lists a range of metrics based on different approaches and assumptions. Even for synthetic datasets with a limited number of generative factors and latents, the disentanglement measure may vary significantly from one metric to another, as shown by Locatello et al. [149, Fig 2], Carbonneau et al. [21, Fig 3] and in Subsection 3.1.2. Obviously, it makes it difficult to choose an appropriate metric, and it thus appears useful to compare their assumptions and the approximations they rely on, so as to emphasize their advantages and drawbacks.

This is developed in Subsection 4.1.1, with a focus on the DCI, MIG and Z_{diff} [84] metrics. It is then described in Subsection 4.1.2 disentanglement evaluations on diSpeech augmented with in-depth analysis of metrics.

4.1.1 Metrics comparison

One major advantage of the DCI metrics is that they provide three indicators, that measure three different aspects of the disentanglement (Section 2.4). This is in line with Carbonneau et al.’s advice that disentanglement properties should be considered distinctly.

Also, the DCI [55] metrics are computed thanks to an importance matrix in which each component represents the relationship between the latents and the generative factors. This is useful as it allows a per-factor analysis instead of a global score. Indeed, as a general rule, some factors can be well disentangled while others are not, due to the structure of the data, the nature of the factor, or its impact on the data generation.

On the other hand, the components of the importance matrix (the importance weights) are deduced from the parameters of a regressor (or classifier when categorical factors are concerned) trained to predict the factors knowing the latents. Although they are clearly influenced by the information about each factor contained in each latent, which is relevant for the metrics, these amounts can be altered by the kind of regressor used, the implementation, the assumed relationship between latents and factors (is it linear or not?), and so on.

Information-based approaches such as the MIG [26] score do not suffer these draw-

backs, as they rely on the computation of the MI between factors and latents, which is often used as a generalized correlation coefficient [209]. But still, there are algorithmic parameters to be chosen. In addition, correlations between latents and between factors are ignored. Also, the metric relies on gaps between the most and second most important MI for each factor, favoring information to be located in a single latent for each factor, and disadvantaging cases where a factor might need two latents to be perfectly captured. This point has been further discussed in Subsection 2.1.2. In addition, it totally misses out the Disentanglement part of DCI as the latents capturing multiple factors are not penalized [21].

The Z-diff metric [84] and its variants also use a prediction algorithm to provide their outcome, but through a low-complexity linear classifier, by design. Thus, the score is less dependent on tunable parameters. Nevertheless, its principle consists in finding the most correlated latent to a given factor, ignoring possible correlations to other latents, which often makes its disentanglement evaluation too optimistic, as observed in Subsection 3.1.2.

Following [21, Tab 2], DCI is the metric that covers the most characteristics. Pragmatically, it is indeed convenient to have a precise idea of latent / factor relationships, factor-wise Completeness and latent-wise Disentanglement. MIG has the advantage of not being influenced by predictor intricacies, but has a too restrictive assumption of disentanglement by using MI gaps.

4.1.2 A closer look to metrics

Thanks to `disentanglement_lib` [149], a broad range of experiments of disentangling models on diSpeech corpus have been conducted. Here are the retained results of β -TCVAE trained with $\beta = 10$ and 8 latent dimensions, as it reached good performances, but similar observations were made with other models and datasets. Z_{diff} , MIG, and DCI¹ analyses are presented in Figure 4.1. Metrics values are reported in Figure 4.1a. As in Subsection 3.1.2, Z-diff suggests a really good disentanglement, while other metrics are more mitigated, especially MIG.

But these global measures keep the disentanglement of each factor hidden. Hence, Figure 4.1b reports MIG, Completeness and Informativeness for each factor, showing that performances highly depend on the considered factor. Formants (F1, F2, and F3) seem well disentangled, while pitch (F0) and fade have poor MIG, Completeness and

1. implemented with XGBoost library, for faster computation

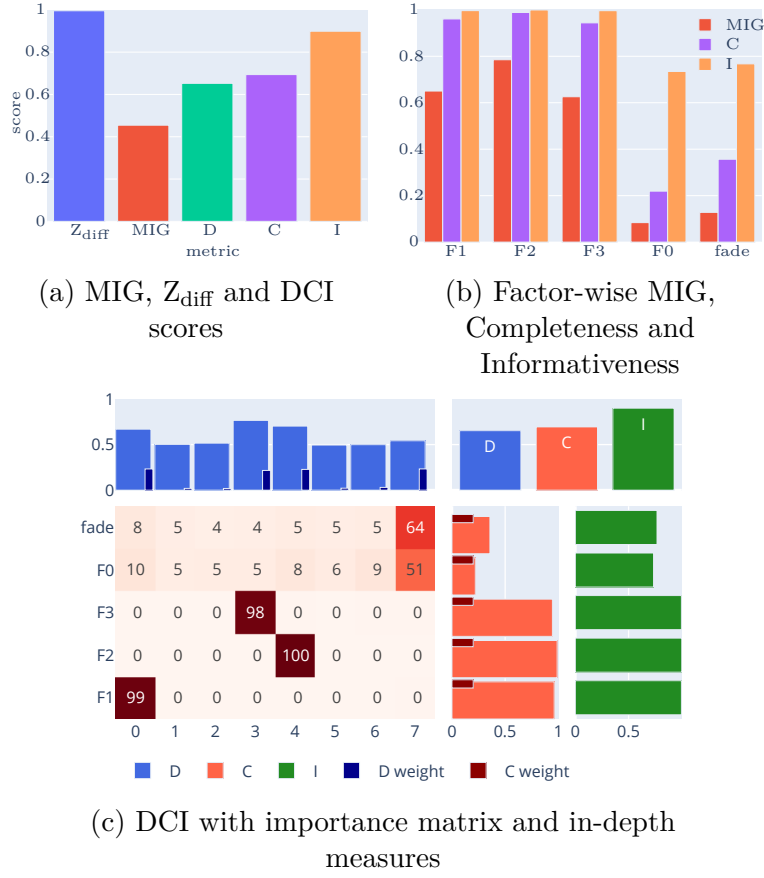
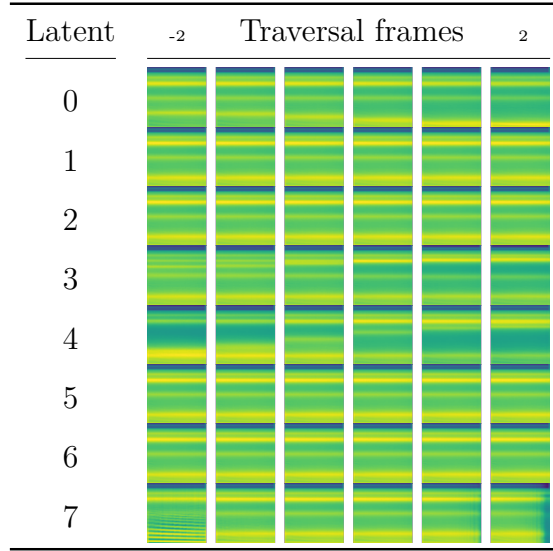


Figure 4.1 – diSpeech disentanglement evaluation

Informativeness.

A closer look at the DCI importance matrix in Figure 4.1c indicates which latent disentangles each formant : F1: latent 0, F2: latent 4, and F3: latent 3. Cell values are the percentage of importance (feature importance \times 100). Figure 4.1c also aligns the importance matrix with entropy-based factor-wise Completeness (right part) and latent-wise Disentanglement (up part), and factor-wise Informativeness. Latent and factor variable importance weights (ρ in [55]) are also reported next to their respective values (thin dark bars). Figure 4.1c is thus an informative yet condensed view of factor / latent relations. It is also suggested by traversals in Table 4.1, where the corresponding formants (F1, F2, and F3) are clearly moving in dimensions 0, 4, and 3, respectively, and only in them.

Table 4.1 – β -TCVAE traversals on diSpeech

4.1.3 Procedure description

In order to figure out if metric outcomes correctly reflect the disentanglement properties of a latent representation, experiments based on a procedure coined *latent decimation* are conducted. The idea is to remove the most informative latents with respect to a given factor, and measure how much of its information has been lost. This loss is evaluated thanks to a predictor (same as DCI), trained to predict the factor from the remaining latents, and the accuracy drop is used to measure the information loss. Thus, if a factor is well disentangled, removing its most important latent should result in a drastic drop in accuracy.

The consistency of the predicted importance can be further challenged by considering the new importance order without the most important latent, and removing the most important once again among the remaining latents. This process is repeated until only one is left, and the importance order is reported at each step, to ensure that latent importances stay consistent throughout the process. A reliable importance matrix is hence expected to exhibit a latent importance ordering stable across decimation steps.

4.1.4 Results on diSpeech

The latent decimation performed with the model described in Subsection 4.1.2 is depicted in Figure 4.2. For each factor, the most important latent (with respect to DCI

importance matrix) is removed to rerun prediction. Then, the latents importance is deduced again, and the new most important latent is removed. This process is repeated until one latent is left. The R^2 scores of each iteration and factor are plotted in Figure 4.2a. Contra-intuitively, from the results in Subsection 4.1.2, factors are still predictable with decent accuracy, meaning that factors' information is not only contained in the most important latents, and not that well disentangled as suggested by DCI.

At each decimation step, one can keep track of latent importances order to assess consistency throughout iterations. The ordered latents at each decimation for F1 are logged in Figure 4.2b: in each column, latent index are stacked in a importance ascending order, and the color scale reflects the importance value. It appears that the importance order is not consistent: latent 7 is the second most important latent at the beginning, but is reported to be the most important only 5 steps further. Similar inconsistent behavior can be observed, with latent 4 and 2, for instance. Similar behaviors are observed with other factors: for F2 in Figure 4.2c, latent 3 starts as the second most important latent, but is decimated only at step 4. With F3 in Figure 4.2d, latent 6 is the fourth most important latent in the initial step, but is the most important in the next step. It also still achieves a prediction accuracy above 0.8, according to Figure 4.2a, while being assigned a very low relative importance in the initial state in Figure 4.2d. For F0 in Figure 4.2e and fade in Figure 4.2f, latent orders are quite stable, although some inconsistencies can still be noticed, for instance: latent 5 for F0 or latent 0 for fade.

These changes underline that information about factors can be spread in other latents, while being announced disentangled by metrics (Figure 4.1a). The traversals in Table 4.1 are also misleading. Hence, it is hypothesized that factor information can be conveyed by multiple latents, but is neither used by predictors (for DCI computation) nor decoders (for transversal generation).

As pointed out, the good disentanglement of formants deduced by DCI is compromised by the latent decimation sanity check. This is following Locatello et al.'s [149] conclusion on the importance of the assessment of the practical benefit of disentanglement. Biases induced by predictors lead to misleading DCI scores. It can be overcome by using an importance matrix based on MI.

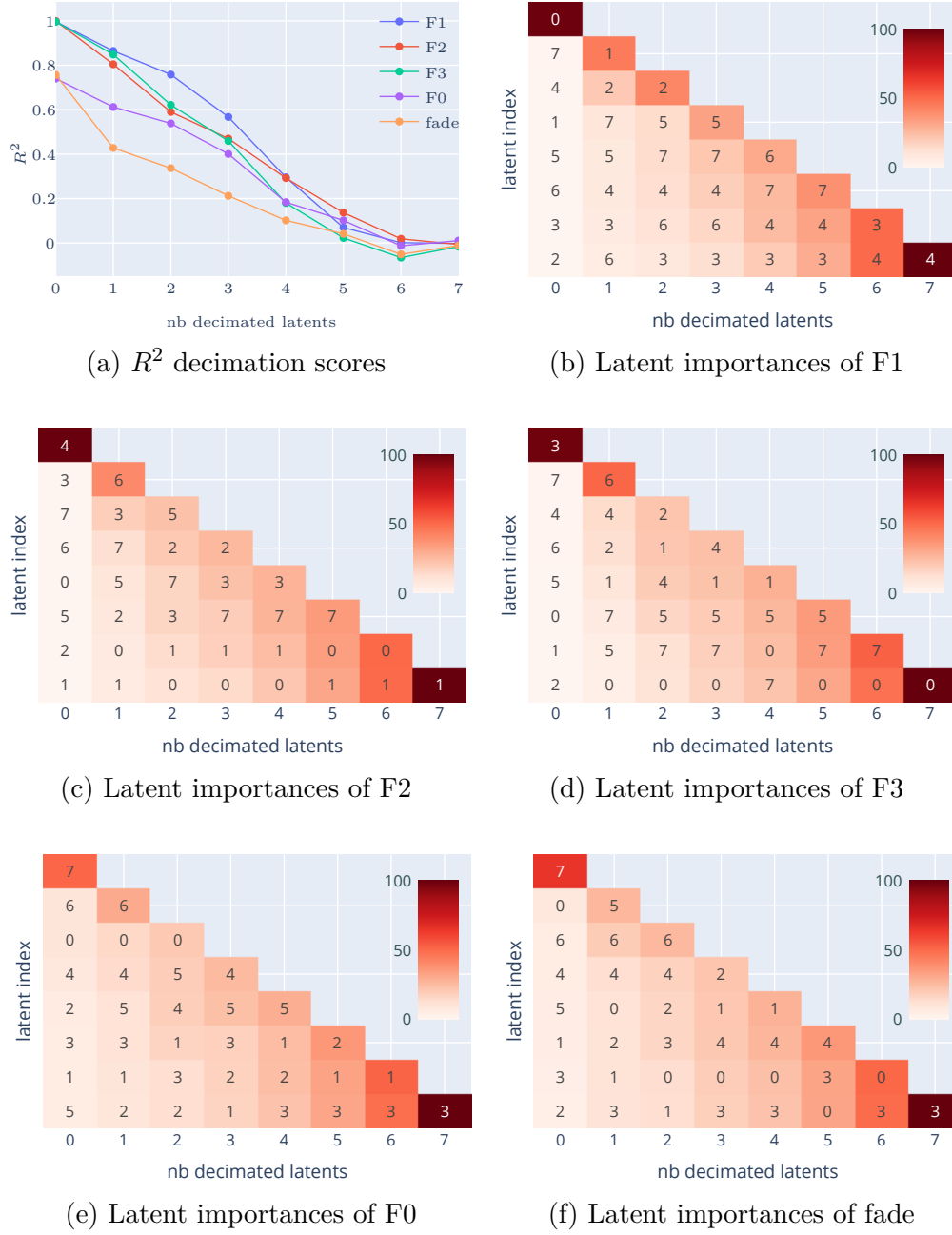


Figure 4.2 – diSpeech latent decimation

4.2 MIDCI

DCI appears in the literature and the experiments in Subsection 4.1.2 as a useful metric to disclose factor / latent relations. But it has also been shown that DCI assessments can be contradicted by the latent decimation procedure. Hence, the Mutual Information-based

DCI (MIDCI) metric is proposed, which is detailed hereafter. Its accordance with latent decimation is then demonstrated.

4.2.1 Definition

In order to overcome predictor biases in DCI, it is proposed to compute an importance matrix based on MI as done in MIG and deduce Disentanglement and Completeness as in DCI. Let $i \in \{1, \dots, l\}$ and $j \in \{1, \dots, m\}$, with l the number of factors and m the number of latents. The MI matrix is defined as:

$$R_{i,j} = \frac{\mathcal{I}(f_i; z_j)}{\mathcal{H}(f_i)}, \quad (4.1)$$

with $\mathcal{I}(f; z)$ the MI between factor f and latent z . MI is divided by $\mathcal{H}(f)$, f 's entropy, so that $R_{i,j} \in [0, 1]$. Straightforwardly, Disentanglement and Completeness are defined as Eastwood et al. [55], by using entropy along latents and factors, respectively.

Note that $S = \sum_{j=1}^l R_{i,j}$ does not necessarily equal 1, as $R_{i,j}$ embodies f_i 's rate of information captured by z_j which can be incomplete ($S \leq 1$) or redundant ($S \geq 1$), due to “cross-information” shared with other latents.

One can also define an information-theory-based formulation of the factor-wise Informativeness Info_i , as:

$$\text{Info}_i = \frac{\mathcal{I}(f_i; z_1, \dots, z_m)}{\mathcal{H}(f_i)}. \quad (4.2)$$

Extended to a global measure of Informativeness Info along all factors, the formulation becomes:

$$\text{Info} = \frac{\mathcal{I}(f_1, \dots, f_l; z_1, \dots, z_m)}{\mathcal{H}(f_1, \dots, f_l)}. \quad (4.3)$$

Those definitions are illustrated through Venn diagrams in Figure 4.3. In Figure 4.3a are represented the interactions between information (i.e., entropy) conveyed by two latents z_1 and z_2 and a factor f_i . The Informativeness defined in (4.2) is represented in the orange area, being the part of f_i overlapped by either factor. This orange part is then normalized by the factor's entropy, $\mathcal{H}(f_i)$, to get a score between 0 and 1, the higher the better. Note that it is different from summing f_i 's Mutual Information with each latent, as it would count multiple times the interaction information, being the area at the intersection of the three variables, in the center of Figure 4.3a. To be generalized with more than one factor, Figure 4.3b illustrates how intricate the interactions can become when two

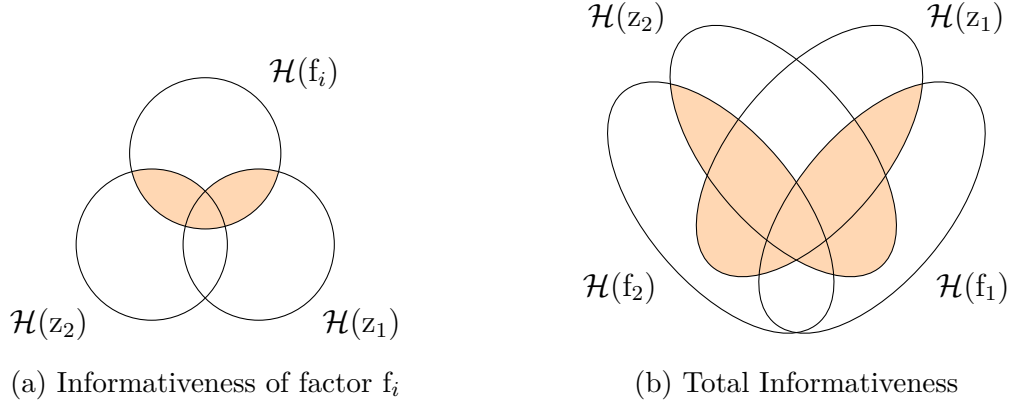


Figure 4.3 – Informativeness Venn diagrams

latents and two factors, f_1 and f_2 , are considered. Here again, the orange area is the amount of factor information captured by latents, being normalized by the total amount of information conveyed by factors in (4.3). The total amount of information is expressed as the joint entropy of all factors, taking into account the correlations between factors. This is believed to be a more representative definition of informativeness than the mean prediction accuracy of multiple predictors.

Nevertheless, in practice, the great number of data points and a possibly important number of latents and factors result in a multivariate distribution, which makes the computation of Info_i and Info a complex challenge, as efficiently estimating MI is still an open challenge. This definition of MIDCI takes benefits from both DCI and MIG: MI based importance matrix overcomes predictor biases, and latent-wise Disentanglement / factor-wise Completeness provides in-depth insights into latent/factor relationships.

4.2.2 Consistency assertion

Coming back to diSpeech disentanglement, applying MIDCI is equivalent to replacing the importance matrix in Figure 4.1c with the MI matrix, resulting in Figure 4.4a. In conformity with *latent decimation*, Completeness appears less optimistic.

In order to assess if MIDCI is closer than DCI to *latent decimation* latents ordering, Normalized Kendall τ distance (K_n) [105] is employed. Intuitively, the normalized Kendall τ distance $K_n(\tau_1, \tau_2)$ between two orderings τ_1 and τ_2 is the number of steps a bubble-sort algorithm would take to align an ordering with the other, normalized by the maximum number of swaps to perform in the worst case, where τ_1 is the inverse of τ_2 , to have $K_n \in [0, 1]$. More precisely, K_n considers all possible pairs (i, j) of element indexes to

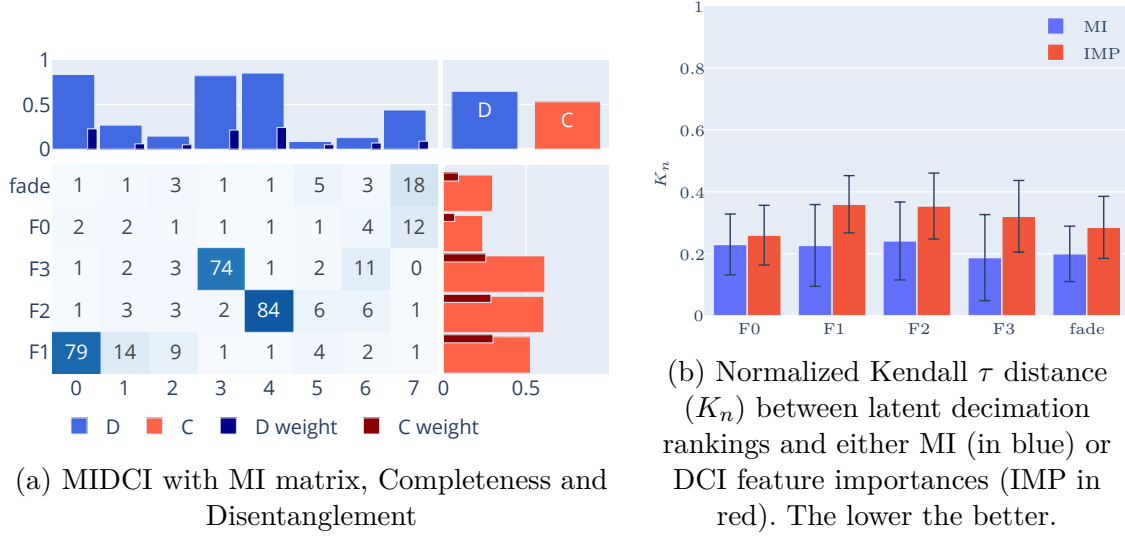


Figure 4.4 – MIDCI metric and importance consistency on diSpeech

order, and counts the number of cases when τ_1 and τ_2 disagree on the relative ordering of the pair, i.e., when $\tau_1(i) > \tau_1(j)$ and $\tau_2(i) < \tau_2(j)$, or $\tau_1(i) < \tau_1(j)$ and $\tau_2(i) > \tau_2(j)$. Accordingly, the lower $K_n(\tau_1, \tau_2)$, the better the accordance between τ_1 and τ_2 . Therefore, this measure of rank correlation is used to assess the agreement between the order by which latents are decimated, with the ordering forecasted by either MI or DCI predictor’s feature importances. Roughly speaking, each diagonal ordering in Figure 4.2 is compared through K_n distance with either its respective first column (the DCI feature importance ordering) or the ordering obtained from computing the MI of the concerned factor with each latent (as done in MIG). Figure 4.4b shows that for several models (used in Section 3.1 and trained on diSpeech), with several numbers of latents (8, 16, 32), K_n is, for each latent, on average smaller with MIDCI than with DCI. Hence, better accordance is achieved when using MI, demonstrating improved reliability.

The described experiments have been extended to the visual synthetic datasets described in Section 2.3: dSprites, Cars3D, and SmallNORB, and for each of them, K_n scores comparing DCI and MI-based importance consistency are displayed in Figure 4.5. Results are overall similar to what is observed with diSpeech, but one can notice some fluctuations for complex factors where DCI has a better K_n than MIDCI: the orientation on dSprites in Figure 4.5a, the azimuth on Cars3D in Figure 4.5b, and the azimuth and elevation on SmallNORB in Figure 4.5c, are better handled with DCI-based importance. They, however, share the common point of being complex and poorly disentangled, while

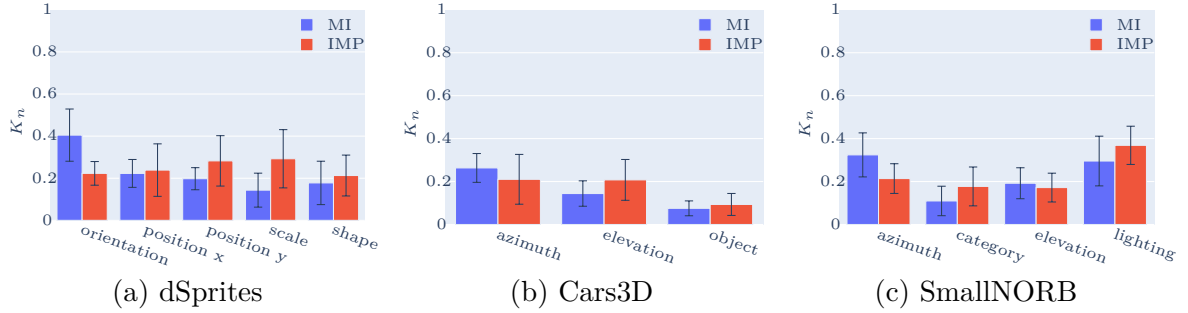


Figure 4.5 – Normalized Kendall τ distance (K_n) between latent decimation rankings and either MI or DCI feature importances (IMP). The lower the better.

other factors, better disentangled by models, are more reliably handled with a MI-based importance matrix.

4.2.3 Conclusions and discussions

As experienced in Chapter 3, investigating unsupervised disentanglement is a complex endeavor. Manually assessing disentanglement through traversal is a tedious task, especially for speech, and metrics still need to rely on ground-truth generative factors. It is hence crucial to ensure that the metrics used are reliable when they assert that a given factor is well disentangled.

Throughout the experiments described in this section, inconsistencies in the computation of the latent importance have been exposed: latents expected to have only one important latent were revealed to be still predictable when removing it, and the ordering of latents according to their importance has been demonstrated to be somehow unstable under the latent decimation sanity check procedure.

To address this inconsistency, MIDCI has been proposed, which exhibits better coherence under latent decimation on diSpeech and also on synthetic image corpora. Overall, a new analysis grid is proposed, through a more reliable version of DCI relying on mutual information, avoiding regressor or predictor biases. It is believed that this study serves as a preliminary research work, which can be beneficial to further investigations in the assessment of disentanglement.

It has also been disclosed through this work that Mutual Information, and more generally, information theory measures, are more reliable as quantifiers of latent / factor relationship strength. With this in mind, a new metric of completeness is proposed in Section 4.3, which builds on the direction initiated by the metric UNIBOUND [221] (de-

scribed in Subsection 2.4.2), and digs further in the decomposition of the information flows when multiple latents and factors are implied.

4.3 Decomposing information to quantify disentanglement

As advanced in Section 4.2, information theory-based quantities are reliable measures to assess disentanglement. Nevertheless, a redundant drawback of the proposed disentanglement metrics is that they do not consider correlations between latents and correlations between factors. This insight is unveiled by Tokui and Sato (2021) [221], which try to discard from MIG [26] redundant information and unique information coming from other latents when computing each latent importance. The definition of Informativeness proposed in Section 4.2 also follows this intuition, that correlations between factors should be taken into account. Hence, following Partial Information Decomposition (PID) framework described in Subsection 2.4.2, hints towards the definition of new metrics measuring completeness and disentanglement that deal with inter-latent and inter-factor correlations are provided.

It is thus proposed in Subsection 4.3.1 to go further in the decomposition of latent/factors information by considering the unique information as the only source of completeness for factors.

4.3.1 Disentangling pieces of information

The uniqueness-based completeness of factor f_i , denoted as \mathcal{C}_i , is defined as the sum of the unique information captured by each latent z_j , normalized by the entropy of factor f_i :

$$\mathcal{C}_i = \frac{\sum_j \mathcal{U}(f_i; z_j)}{\mathcal{H}(f_i)}. \quad (4.4)$$

In the more general scenario when multiple factors are involved, the global measure of completeness based on uniqueness is defined as the very same sum of unique information of each latent z_j , but regarding the joint distribution of factors $\{f_1, \dots, f_l\}$:

$$\mathcal{C} = \frac{\sum_j \mathcal{U}(f_1, \dots, f_l; z_j)}{\mathcal{H}(f_1, \dots, f_l)} \quad (4.5)$$

Both scenarios are illustrated in Figure 4.6. The main intuition behind factor-wise completeness, \mathcal{C}_i , is that only the parts of f_i captured by only one latent is considered disentangled. In other words, parts that are redundantly or synergistically captured are considered entangled². This definition also has the benefit of favoring cases when multiple latents convey information about a factor, as long as they do not convey the same kind of information, i.e., if several latents explain distinct parts of a factor, the latter is considered to exhibit good completeness. It is illustrated in Figure 4.6a, where the hatched part in the left Venn diagram corresponds to the region described by PID framework. Within this area, the decomposition is pictured in the right Venn diagram, with redundant \mathcal{R} , synergistic \mathcal{S} , and unique \mathcal{U} information pieces. Hence, the orange and blue regions are the unique information captured by z_1 and z_2 , respectively, about f_i , which are summed in the defined completeness \mathcal{C}_i . In the more general case of multiple factors, Figure 4.6b shows that the very same intuition is followed, except that factors are considered jointly. It is hence considered that, if correlations subsist between factors, latents capturing in a “unique” way those variations are counted positively.

The same concerns, about correlations between latents and factors, can be derived for the measure of the modularity property, i.e., minimizing the number of latents informative about each factor. An uniqueness-based metric to quantify modularity, analogous to \mathcal{C}_i and \mathcal{C} , could be relevant and constitute a whole PID-based disentanglement analysis framework. But in contrast to completeness, there is no reason to allow a latent to capture multiple factors. Hence, it is believed that such a modularity metric cannot be expressed in the very same way as its completeness counterparts \mathcal{C}_i and \mathcal{C} . This direction calls for further investigations, in order to find a proper PID-based measure, fitting the definition of modularity.

In the literature, the UNIBOUND [221] metric is the closest proposition to the proposed uniqueness-based metric. However, the distinction from UNIBOUND is two-fold: on the one hand, in \mathcal{C}_i and \mathcal{C} , a factor can be disentangled by several latents, as long as they convey distinct information pieces, whereas UNIBOUND enforces the uniqueness of a single latent and penalizes other unique information coming from other latents; on the other hand, \mathcal{C} considers the correlations between factors, which are highly likely to be encountered in practice where the ground-truth generative factors are unknown, while UNIBOUND averages quantities computed for each factor, discarding their relationships.

2. Unique, redundant, and synergistic information pertaining to PID framework, are described in Subsection 2.4.2, with UNIBOUND metric.

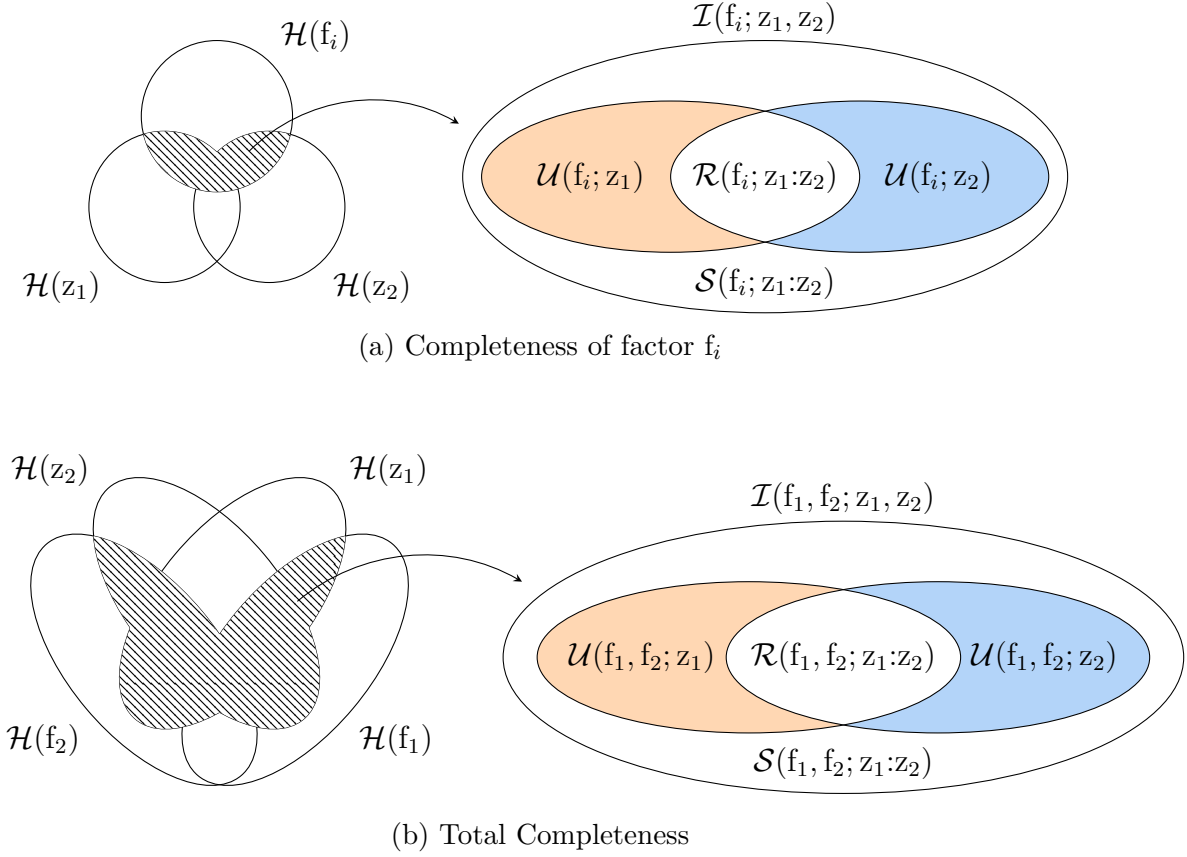


Figure 4.6 – Completeness Venn diagrams

4.3.2 Computing partial information pieces

It is believed that the proposed measure fits the completeness property. The challenge remains in the way to compute the partial information pieces. While the notion of PID is proposed by Williams and Beer (2010) [234], their definition of redundancy is argued by Ince (2017) [99] to not represent the actual amount of redundant information, but only the minimum amount captured. Interesting alternatives are then proposed by Ince with the common change in surprisal measure of redundancy (\mathcal{I}_{CCS}), or by Makkeh et al. (2021) [153] with the shared exclusion principle (\mathcal{I}_{\cap}^{sx}). Both measures of redundancy lead to different results, hence, the way to properly interpret those quantities and which one better fits disentanglement endeavor will be the concern of further investigations. Some preliminary experiments have already been conducted with simple and synthetic cases, i.e., Cartesian coordinates in a 2D space as factors and polar coordinates as latents, or a rotation being described by its angle as factor, and through its sin and cos components

as latents. However, neither \mathcal{I}_{CCS} nor \mathcal{I}_{\cap}^{sx} metrics have led to satisfying and easy-to-read results.

Their computation is also not straightforward, as information theory-based measures when many possible combinations are involved imply a great number of operations to cover distributions. Future work will then explore how to efficiently compute the proposed completeness measure, by discarding irrelevant partial information pieces.

A similar definition of modularity property, which might also benefit from PID framework granularity, is left for further investigation.

4.4 Conclusions and discussions

The difficulty of learning disentangled representations, especially for speech, has been experienced in Chapter 3. It is therefore of great importance to ensure that the evaluation protocols used are reliable. To this end, a sanity-check procedure, latent decimation, has been proposed. The metric DCI, acknowledged to be convenient and reliable, is demonstrated to exhibit inconsistent behaviors under latent decimation. The metric Mutual Information-based DCI (MIDCI) has then been proposed, to address the inconsistencies, thanks to a Mutual Information-based importance matrix.

However, the common drawback of the literature’s metrics is that correlations between factors and between latents are not considered. Thus, promising hints are introduced based on Partial Information Decomposition (PID), to dig deeper into the intricate interactions between factors and latents, and to advance a measure of completeness based on the unique information captured by latents.

All in all, measuring disentanglement properties remains convoluted, mainly due to the lack of a formal definition of disentanglement. Despite some attempts through symmetry [85] or causality [213], no natural way to quantify the degree of disentanglement of a learned model arises.

CONCLUSION

Speech is human’s most efficient, yet intricate, means of communication. Beyond the linguistic content, speech conveys way more information, about speaker’s emotional state or intents, and about the speaker himself. With the recent advances in machine learning, unprecedented performances of speech “understanding” have been achieved by computers in the past few years, approaching the near-human level in speech transcription and synthesis.

While deep learning demonstrated impressive results in very specific tasks through supervised learning, the recent trend to leverage huge amounts of data in a self-supervised approach settles a new era, where such pre-trained models fine-tuned on underlying tasks are exhibiting groundbreaking efficiency. Despite all this progress, speech attributes remain hard to comprehend for learning algorithms.

Speech production mechanisms are now well understood. But it remains fuzzy, even for humans, to formally link articulatory cues to one’s perceptual interpretations of prosody, emotion, or identity. It is even more difficult for machines, as no formal taxonomy of speech attributes has yet been acknowledged. Among the neural network-based approaches aiming to learn insightful abstract representations, disentanglement learning stands out as a promising paradigm, that has the purpose of separating independent factors of variations in distinct axes in learned latent space. It is therefore believed that captured independent variations pertaining to a set of observations are tied to its interpretable generative and explanatory factors. Such a model successfully trained on speech data has the potential to isolate speech characteristics, until then intricate in speech variabilities, and control them in speech synthesis scenarios.

Key contributions

Therefore, this thesis endeavors to explore the junction between disentanglement methods and speech processing, with the first key contribution being the proposition of a dataset of synthetic vowels, analogous to synthetic image corpora widely used in disentanglement studies. Hence, the proposed corpus diSpeech [247] is defined based on five

generative factors: the first three first formants F1, F2, and F3, the fundamental frequency F0, and its fade rate. French vowels are covered through formants, harmonic cues are controlled with pitch, and temporal variabilities are added with pitch’s fade rate. The disentanglement of diSpeech is subsequently experienced with β -VAE model, and evaluated with latent space traversals and disentanglement metrics. The disentanglement of TIMIT’s vowels is then tested with a β -VAE trained on TIMIT and inferred on diSpeech, with limited yet interesting results.

Investigating speech disentanglement is, however, shown to be tedious, as searching for the best model and hyperparameter setting involves a bunch of models that cannot all be traversed manually, and the various proposed disentanglement metrics rarely agree with each other and are hence hard to read. According to Carbonneau et al. (2021) [21], DCI [55] stands as a metric with good properties, as it measures distinct disentanglement properties, and can be leveraged for factor-wise and latent-wise analysis. It is however demonstrated that predictor and regressor-based importance matrix computed by DCI exhibit inconsistencies under latent decimation procedure [246]: factors stated well disentangled are still predictable when removing their most important factor, and the latent importance ordering is unstable when removing the most important latent and computing importance again. MIDCI is hence introduced, which advances a Mutual Information-based computation of the importance matrix, similarly to MIG. MIDCI scores are thus more mitigated, which better matches latent decimation forecasts, and Kendall τ distance demonstrates that the MI-based importance ordering is more consistent with latent decimation procedure ordering.

Discussions and perspectives

The redundant issue encountered throughout the experiments conducted in this thesis is the lack of a consensual definition of disentangled representation. It is thence difficult to know what to measure, how, and which metric to rely on. Interesting directions are provided from symmetry theories [85], causality [213], but require further developments to be followed practically in unsupervised scenarios. It is believed that a promising Partial Information Decomposition (PID)-based definition of disentanglement is hinted in Section 4.3, although it still does not provide explicit mechanisms to learn disentangled representations.

Disentanglement is a non-trivial task per se, and dealing with speech data, being

highly subjective and conveying multiple information at different time scales, makes speech disentanglement a very arduous desiderata. The development of sophisticated inductive biases, such as Capacitron [11] or GMVAE [93] with hierarchical latent space structure, is a promising research direction to enforce the disentanglement of speech latents, by introducing the prior knowledge of the multi-scale information organization in speech. Evaluation protocols, more tailored to speech data than traversals, would also be beneficial towards the assessment of disentangled speech representations.

All in all, the difficulty in disentangling speech attributes is deeply rooted in the lack of a widely acknowledged taxonomy of speech attributes. It is perilous to design speech disentanglement models, when one does not know exactly what to disentangle, and from what. Many approaches deal with this uncertainty by supervising the disentanglement of known labels (e.g. speaker identity, style, emotion) and leaving the neural network to handle the remaining variabilities which cannot be pigeonholed. The elaboration of techniques able to rely on available knowledge to better focus on the extraction of the missing information, in a semi-supervised fashion, is therefore a research topic with great potential. To another extent, it pertains to linguistic research efforts to establish proper frameworks to decompose and describe speech attributes, and their relationships.

In addition, taking a step back, one may have noticed that a great deal of disciplines are utilized throughout the work exposed in this manuscript: linguistics, acoustics, machine learning, information theory, Bayesian inference, and references to symmetry in physics and causality are also worth mentioning. This multidisciplinary of the thesis topic highlights how complex but fascinating the study of speech attribute disentanglement is.

To conclude, promising approaches are developed towards the disentanglement of data underlying factors of variations, and more generally towards structured latent spaces, with the properties advanced by Bengio et al. (2013) [13]. On the other hand, well-elaborated speech processing systems have recently been deployed with outstanding results in speech synthesis, voice cloning, dubbing, and so on. But such models are actually clueless about the hidden and intricate structure of speech attributes. Hopefully, this thesis will contribute to bridge the gap between speech and disentanglement, beyond hearing and towards listening machines. In an era where Large Language Models (LLMs) are already changing human's relation with artificial intelligence, deep learning systems are constantly raising the limits of neural networks' complexity and capacity to absorb information. But the challenge of the incoming groundbreaking advances resides in systems' capacity to disentangle world's intricacies, to build models more rational, reliable, explainable, robust

and aligned with human’s intuitions.

ACRONYMS

ASR Automatic Speech Recognition. 45, 174

ASV Automatic Speaker Recognition. 44, 174

CNN Convolutional Neural Network. 21, 41

DNN Deep Neural Network. 38, 41, 42

DTW Dynamic Time Warping. 43

ELBO Evidence Lower Bound. 67, 90, 120

FFT Fast Fourier Transform. 122

FHVAE Factorized Hierarchical VAE. 8, 14, 60, 74, 117, 118, 120, 122, 123, 124, 125, 126

HMM Hidden Markov Model. 40, 41

LLM Large Language Model. 10, 23, 47, 147

LSTM Long-Short-Term-Memory. 21, 41, 42, 120, 122

MI Mutual Information. 45, 65, 127, 131, 134, 136, 137, 138, 139, 143, 146

MIDCI Mutual Information-based DCI. 8, 10, 15, 129, 135, 137, 138, 139, 143, 146

PID Partial Information Decomposition. 8, 25, 81, 82, 127, 130, 140, 141, 142, 143, 146

PPG Phonetic Posteriorgram. 45

PSOLA Pitch-Synchronous Overlap and Add. 40

ResNet Residual Network. 21

RNN Recurrent Neural Network. 41, 42

SGD Stochastic Gradient Descent. 21

SOTA State-of-the-art. 20, 38, 45

SPSS Statistical Parametric Speech Synthesis. 20, 40

TD-PSOLA Time-Domain Pitch-Synchronous Overlap and Add. 46

TTS Text-to-Speech. 19, 20, 21, 39, 44, 45, 46, 47, 174

UDR Unsupervised Disentanglement Ranking. 84, 85, 123

VAE Variational Autoencoder. 6, 7, 8, 14, 25, 43, 45, 71, 73, 74, 89, 90, 91, 92, 93, 94, 101, 102, 110, 117, 118, 149

VC Voice Conversion. 22, 23, 24, 38, 39, 44, 45, 46, 47, 58, 59, 60

BIBLIOGRAPHY

- [1] Abdi, Amir H, Abolmaesumi, Purang, and Fels, Sidney, « A preliminary study of disentanglement with insights on the inadequacy of metrics », *in: arXiv preprint arXiv:1911.11791* (2019).
- [2] Alemi, Alexander et al., « Fixing a broken ELBO », *in: International conference on machine learning*, PMLR, 2018, pp. 159–168.
- [3] An, Xiaochun, Soong, Frank K, and Xie, Lei, « Disentangling style and speaker attributes for tts style transfer », *in: IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), pp. 646–658.
- [4] Appelbaum, I., « The Lack of Invariance Problem and the Goal of Speech Perception », *in: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, vol. 3, Oct. 1996, 1541–1544 vol.3, DOI: 10.1109/ICSLP.1996.607912.
- [5] Arik, Serkan et al., « Deep voice 2: Multi-speaker neural text-to-speech », *in: arXiv preprint arXiv:1705.08947* (2017).
- [6] Arik, Serkan Ö et al., « Deep voice: Real-time neural text-to-speech », *in: International conference on machine learning*, PMLR, 2017, pp. 195–204.
- [7] Asperti, Andrea and Trentin, Matteo, « Balancing reconstruction error and kullback-leibler divergence in variational autoencoders », *in: IEEE Access* 8 (2020), pp. 199440–199448.
- [8] Aubry, Mathieu et al., « Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models », *in: Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3762–3769.
- [9] Baevski, Alexei et al., « wav2vec 2.0: A framework for self-supervised learning of speech representations », *in: Advances in neural information processing systems* 33 (2020), pp. 12449–12460.

-
- [10] Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua, « Neural machine translation by jointly learning to align and translate », *in: arXiv preprint arXiv:1409.0473* (2014).
- [11] Battenberg, Eric et al., *Effective Use of Variational Embedding Capacity in Expressive End-to-End Speech Synthesis*, 2020, URL: <https://openreview.net/forum?id=SJgBQaVKwH>.
- [12] Bele, Irene Velsvik, « Reliability in perceptual analysis of voice quality », *in: Journal of Voice* 19.4 (2005), pp. 555–573.
- [13] Bengio, Yoshua, Courville, Aaron, and Vincent, Pascal, « Representation learning: A review and new perspectives », *in: IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.
- [14] Berckmoes, Celine and Vingerhoets, Guy, « Neural foundations of emotional speech processing », *in: Current Directions in Psychological Science* 13.5 (2004), pp. 182–185.
- [15] Bishop, Christopher M and Nasrabadi, Nasser M, *Pattern recognition and machine learning*, vol. 4, 4, Springer, 2006.
- [16] Bowman, Samuel R et al., « Generating sentences from a continuous space », *in: arXiv preprint arXiv:1511.06349* (2015).
- [17] Brown, Tom et al., « Language models are few-shot learners », *in: Advances in neural information processing systems* 33 (2020), pp. 1877–1901.
- [18] Burda, Yuri, Grosse, Roger, and Salakhutdinov, Ruslan, « Importance weighted autoencoders », *in: arXiv preprint arXiv:1509.00519* (2015).
- [19] Burgess, Chris and Kim, Hyunjik, *3D Shapes Dataset*, <https://github.com/deepmind/3dshapes-dataset/>, 2018.
- [20] Burgess, Christopher P et al., « Understanding disentangling in β -VAE », *in: arXiv preprint arXiv:1804.03599* (2018).
- [21] Carbonneau, Marc-André et al., « Measuring disentanglement: A review of metrics », *in: IEEE transactions on neural networks and learning systems* (2022).
- [22] Casanova, Edresson et al., « Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone », *in: International Conference on Machine Learning*, PMLR, 2022, pp. 2709–2720.

-
- [23] Chan, Chak Ho et al., « Speechsplit2. 0: Unsupervised speech disentanglement for voice conversion without tuning autoencoder bottlenecks », *in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6332–6336.
- [24] Charpentier, Francis and Stella, M., « Diphone synthesis using an overlap-add technique for speech waveforms concatenation », *in: ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 11, IEEE, 1986, pp. 2015–2018.
- [25] Chen, Meiyang and Duan, Zhiyao, « ControlVC: Zero-Shot Voice Conversion with Time-Varying Controls on Pitch and Rhythm », *in: arXiv preprint arXiv:2209.11866* (2022).
- [26] Chen, Ricky TQ et al., « Isolating sources of disentanglement in variational autoencoders », *in: Advances in neural information processing systems* 31 (2018).
- [27] Chen, Sanyuan et al., « Wavlm: Large-scale self-supervised pre-training for full stack speech processing », *in: IEEE Journal of Selected Topics in Signal Processing* 16.6 (2022), pp. 1505–1518.
- [28] Chen, Xi et al., « Variational lossy autoencoder », *in: arXiv preprint arXiv:1611.02731* (2016).
- [29] Cho, Taehong and Mücke, Doris, « Articulatory measures of prosody », *in:* (2020).
- [30] Choi, Hyeong-Seok et al., « NANSY++: Unified voice synthesis with neural analysis and synthesis », *in: arXiv preprint arXiv:2211.09407* (2022).
- [31] Chung, Junyoung et al., « Empirical evaluation of gated recurrent neural networks on sequence modeling », *in: arXiv preprint arXiv:1412.3555* (2014).
- [32] Conzález, Julio, « Correlations between speakers' body size and acoustic parameters of voice », *in: Perceptual and motor skills* 105.1 (2007), pp. 215–220.
- [33] Cox, Arnie, « Embodying music: Principles of the mimetic hypothesis », *in: Music Theory Online* 17.2 (2011).
- [34] Crystal, David, *A dictionary of linguistics and phonetics*, John Wiley & Sons, 2011.
- [35] Dai, Bin, Wang, Ziyu, and Wipf, David, « The usual suspects? Reassessing blame for VAE posterior collapse », *in: International conference on machine learning*, PMLR, 2020, pp. 2313–2322.

-
- [36] Dai, Bin and Wipf, David, « Diagnosing and Enhancing VAE Models », *in: International Conference on Learning Representations*, 2019, URL: <https://openreview.net/forum?id=B1e0X3C9tQ>.
- [37] Défossez, Alexandre et al., « High fidelity neural audio compression », *in: arXiv preprint arXiv:2210.13438* (2022).
- [38] Dejonckere, Philippe H et al., « Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements. », *in: Revue de laryngologie-otologie-rhinologie* 117.3 (1996), pp. 219–224.
- [39] Dellwo, Volker, Huckvale, Mark, and Ashby, Michael, « How is individuality expressed in voice? An introduction to speech production and description for speaker classification », *in: Speaker Classification I: Fundamentals, Features, and Methods* (2007), pp. 1–20.
- [40] Dellwo, Volker, Leemann, Adrian, and Kolly, Marie-José, « Speaker idiosyncratic rhythmic features in the speech signal », *in: Interspeech Conference Proceedings*, 2012.
- [41] Delvaux, Véronique and Pillot-Loiseau, Claire, « Perceptual judgment of voice quality in nondysphonic French speakers: effect of task-, speaker-and listener-related variables », *in: Journal of Voice* 34.5 (2020), pp. 682–693.
- [42] Desplanques, Brecht, Thienpondt, Jenthe, and Demuynck, Kris, « Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification », *in: arXiv preprint arXiv:2005.07143* (2020).
- [43] Devlin, Jacob et al., « Bert: Pre-training of deep bidirectional transformers for language understanding », *in: arXiv preprint arXiv:1810.04805* (2018).
- [44] Ding, Zheng et al., « Guided Variational Autoencoder for Disentanglement Learning », *in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [45] Dixon, N and Maxey, H, « Terminal analog synthesis of continuous speech using the diphone method of segment assembly », *in: IEEE transactions on Audio and Electroacoustics* 16.1 (1968), pp. 40–50.
- [46] Doersch, Carl, « Tutorial on variational autoencoders », *in: arXiv preprint arXiv:1606.05908* (2016).

-
- [47] Donahue, Chris, McAuley, Julian, and Puckette, Miller, « Adversarial audio synthesis », *in: arXiv preprint arXiv:1802.04208* (2018).
- [48] Donahue, Jeff et al., « End-to-end adversarial text-to-speech », *in: arXiv preprint arXiv:2006.03575* (2020).
- [49] Dos Santos, Priscila Campos Martins et al., « Effect of auditory-perceptual training with natural voice anchors on vocal quality evaluation », *in: Journal of Voice* 33.2 (2019), pp. 220–225.
- [50] Du, Zongyang et al., « Disentanglement of Emotional Style and Speaker Identity for Expressive Voice Conversion », *in: Proc. Interspeech 2022*, 2022, pp. 2603–2607, DOI: 10.21437/Interspeech.2022-10249.
- [51] Duan, Sunny et al., « Unsupervised model selection for variational disentangled representation learning », *in: arXiv preprint arXiv:1905.12614* (2019).
- [52] Duchi, John, « Derivations for linear algebra and optimization », *in: Berkeley, California* 3.1 (2007), pp. 2325–5870.
- [53] Duchi, John C., Hazan, Elad, and Singer, Yoram, « Adaptive Subgradient Methods for Online Learning and Stochastic Optimization », *in: Journal of machine learning research*, 2011.
- [54] Dudley, Homer G, « The Vocoder », *in: The Bell System Technical Journal* 18.3 (1939), pp. 126–136.
- [55] Eastwood, Cian and Williams, Christopher KI, « A framework for the quantitative evaluation of disentangled representations », *in: International conference on learning representations*, 2018.
- [56] Evain, Solène et al., « Task agnostic and task specific self-supervised learning from speech with lebenchmark », *in: Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021.
- [57] Fan, Yuchen et al., « Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis », *in: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 4475–4479.
- [58] Fan, Yuchen et al., « TTS synthesis with bidirectional LSTM based recurrent neural networks », *in: Fifteenth annual conference of the international speech communication association*, 2014.

-
- [59] Fitch, W Tecumseh and Giedd, Jay, « Morphology and development of the human vocal tract: A study using magnetic resonance imaging », *in: The Journal of the Acoustical Society of America* 106.3 (1999), pp. 1511–1522.
 - [60] Fougeron, Cécile and Keating, Patricia A, « Articulatory strengthening at edges of prosodic domains », *in: The journal of the acoustical society of America* 101.6 (1997), pp. 3728–3740.
 - [61] Gagnon, R, « Votrax real time hardware for phoneme synthesis of speech », *in: ICASSP'78. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, IEEE, 1978, pp. 175–178.
 - [62] Gallardo, Laura Fernández and Weiss, Benjamin, « The nautilus speaker characterization corpus: Speech recordings and labels of speaker characteristics and voice descriptions », *in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.
 - [63] Ganin, Yaroslav et al., « Domain-adversarial training of neural networks », *in: The journal of machine learning research* 17.1 (2016), pp. 2096–2030.
 - [64] Garofolo, John S, « Timit acoustic phonetic continuous speech corpus », *in: Linguistic Data Consortium, 1993* (1993).
 - [65] Gelfer, Marylou Pausewang, « Perceptual attributes of voice: Development and use of rating scales », *in: Journal of Voice* 2.4 (1988), pp. 320–326.
 - [66] Georgeton, Laurianne et al., « Analyse formantique des voyelles orales du français en contexte isolé: à la recherche d’une référence pour les apprenants de FLE », *in: Conférence conjointe JEP-TALN-RECITAL 2012*, 2012, pp. 145–152.
 - [67] Gondal, Muhammad Waleed et al., « On the Transfer of Inductive Bias from Simulation to the Real World: a New Disentanglement Dataset », *in: Advances in Neural Information Processing Systems*, ed. by H. Wallach et al., vol. 32, Curran Associates, Inc., 2019, URL: <https://proceedings.neurips.cc/paper/2019/file/d97d404b6119214e4a7018391195240a-Paper.pdf>.
 - [68] Goodfellow, Ian, Bengio, Yoshua, and Courville, Aaron, *Deep learning*, MIT press, 2016.
 - [69] Goodfellow, Ian et al., « Generative adversarial nets », *in: Advances in neural information processing systems* 27 (2014).

-
- [70] Goodfellow, Ian J., Shlens, Jonathon, and Szegedy, Christian, « Explaining and Harnessing Adversarial Examples », *in: CoRR* abs/1412.6572 (2014), URL: <https://api.semanticscholar.org/CorpusID:6706414>.
- [71] Graves, Alex et al., « Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks », *in: Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.
- [72] Guo, Zhifang et al., « PromptTTS: Controllable text-to-speech with text descriptions », *in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.
- [73] Gussenhoven, Carlos, « The phonology of tone and intonation », *in:* (2004).
- [74] Habib, Raza et al., « Semi-Supervised Generative Modeling for Controllable Speech Synthesis », *in: International Conference on Learning Representations*, 2020, URL: <https://openreview.net/forum?id=rJeqeCEtvH>.
- [75] Hammarberg, Britta, « Voice research and clinical needs », *in: Folia phoniatrica et logopaedica* 52.1-3 (2000), pp. 93–102.
- [76] Hart, Johan't, Collier, René, and Cohen, Antonie, « A perceptual study of intonation », *in: (No Title)* (1990).
- [77] Hashimoto, Kei et al., « The effect of neural networks in statistical parametric speech synthesis », *in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, pp. 4455–4459.
- [78] He, Junxian et al., « Lagging inference networks and posterior collapse in variational autoencoders », *in: arXiv preprint arXiv:1901.05534* (2019).
- [79] He, Kaiming et al., « Deep residual learning for image recognition », *in: Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [80] Heidemann, Kate, « A System for Describing Vocal Timbre in Popular Song. », *in: Music Theory Online* 22.1 (2016).
- [81] Hellbernd, Nele and Sammler, Daniela, « Prosody conveys speaker's intentions: Acoustic cues for speech act perception », *in: Journal of Memory and Language* 88 (2016), pp. 70–86.

-
- [82] Heo, Hee Soo et al., « Clova baseline system for the voxceleb speaker recognition challenge 2020 », *in: arXiv preprint arXiv:2009.14153* (2020).
- [83] Hermansky, Hynek, « Perceptual linear predictive (PLP) analysis of speech », *in: the Journal of the Acoustical Society of America* 87.4 (1990), pp. 1738–1752.
- [84] Higgins, Irina et al., « beta-vae: Learning basic visual concepts with a constrained variational framework », *in: International conference on learning representations*, 2016.
- [85] Higgins, Irina et al., « Towards a definition of disentangled representations », *in: arXiv preprint arXiv:1812.02230* (2018).
- [86] Hirano, Minoru and McCormick, Karen R, *Clinical examination of voice by Minoru Hirano*, 1986.
- [87] Hirst, Daniel and Di Cristo, Albert, « Intonation systems », *in: Cambridge: CUP* (1998).
- [88] Hochreiter, Sepp and Schmidhuber, Jürgen, « Long short-term memory », *in: Neural computation* 9.8 (1997), pp. 1735–1780.
- [89] Hoerl, Arthur E and Kennard, Robert W, « Ridge regression: Biased estimation for nonorthogonal problems », *in: Technometrics* 12.1 (1970), pp. 55–67.
- [90] Hoffman, Matthew D and Johnson, Matthew J, « Elbo surgery: yet another way to carve up the variational evidence lower bound », *in: Workshop in Advances in Approximate Bayesian Inference, NIPS*, vol. 1, 2, 2016.
- [91] Hsu, Chin-Cheng et al., « Voice conversion from non-parallel corpora using variational auto-encoder », *in: 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, IEEE, 2016, pp. 1–6.
- [92] Hsu, Wei-Ning, Zhang, Yu, and Glass, James, « Unsupervised learning of disentangled and interpretable representations from sequential data », *in: Advances in neural information processing systems* 30 (2017).
- [93] Hsu, Wei-Ning et al., « Hierarchical Generative Modeling for Controllable Speech Synthesis », *in: International Conference on Learning Representations*, 2019, URL: <https://openreview.net/forum?id=rygkk305YQ>.

-
- [94] Hsu, Wei-Ning et al., « Hubert: Self-supervised speech representation learning by masked prediction of hidden units », *in: IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 3451–3460.
 - [95] Hunt, A.J. and Black, A.W., « Unit selection in a concatenative speech synthesis system using a large speech database », *in: 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1, 1996, 373–376 vol. 1, DOI: 10.1109/ICASSP.1996.541110.
 - [96] Hunt, Andrew J and Black, Alan W, « Unit selection in a concatenative speech synthesis system using a large speech database », *in: 1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*, vol. 1, IEEE, 1996, pp. 373–376.
 - [97] Huttar, George L, « Relations between prosodic variables and emotions in normal American English utterances », *in: Journal of Speech and Hearing Research* 11.3 (1968), pp. 481–487.
 - [98] Imai, Satoshi, Sumita, Kazuo, and Furuichi, Chieko, « Mel log spectrum approximation (MLSA) filter for speech synthesis », *in: Electronics and Communications in Japan (Part I: Communications)* 66.2 (1983), pp. 10–18.
 - [99] Ince, Robin AA, « Measuring multivariate redundant information with pointwise common change in surprisal », *in: Entropy* 19.7 (2017), p. 318.
 - [100] Jakobovits, Leon et al., « Effects of repeated stimulation on cognitive aspects of behavior: some experiments on the phenomenon of semantic satiation. », *in: (1962).*
 - [101] Jia, Ye et al., « Transfer learning from speaker verification to multispeaker text-to-speech synthesis », *in: Advances in neural information processing systems* 31 (2018).
 - [102] Kalchbrenner, Nal et al., « Efficient neural audio synthesis », *in: International Conference on Machine Learning*, PMLR, 2018, pp. 2410–2419.
 - [103] Kempelen, Wolfgang von, *Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine*, Vienna: J.V. Degen, 1791.
 - [104] Kempster, Gail B et al., « Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol », *in: (2009).*
 - [105] Kendall, Maurice G, « A new measure of rank correlation », *in: Biometrika* 30.1/2 (1938), pp. 81–93.

-
- [106] Kent, Ray D, « Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders », *in: American Journal of Speech-Language Pathology* 5.3 (1996), pp. 7–23.
- [107] Le-Khac, Phuc H, Healy, Graham, and Smeaton, Alan F, « Contrastive representation learning: A framework and review », *in: Ieee Access* 8 (2020), pp. 193907–193934.
- [108] Khemakhem, Ilyes et al., « Variational autoencoders and nonlinear ica: A unifying framework », *in: International Conference on Artificial Intelligence and Statistics*, PMLR, 2020, pp. 2207–2217.
- [109] Kim, Hyunjik and Mnih, Andriy, « Disentangling by factorising », *in: International Conference on Machine Learning*, PMLR, 2018, pp. 2649–2658.
- [110] Kim, Jaehyeon, Kong, Jungil, and Son, Juhee, « Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech », *in: International Conference on Machine Learning*, PMLR, 2021, pp. 5530–5540.
- [111] Kim, Jaehyeon et al., « Glow-tts: A generative flow for text-to-speech via monotonic alignment search », *in: Advances in Neural Information Processing Systems* 33 (2020), pp. 8067–8077.
- [112] Kim, Yoon and Rush, Alexander M, « Sequence-level knowledge distillation », *in: arXiv preprint arXiv:1606.07947* (2016).
- [113] King, Simon, « A beginners’ guide to statistical parametric speech synthesis », *in: The Centre for Speech Technology Research, University of Edinburgh, UK* (2010), pp. 28–29.
- [114] Kingma, Diederik P, Welling, Max, et al., « An introduction to variational autoencoders », *in: Foundations and Trends® in Machine Learning* 12.4 (2019), pp. 307–392.
- [115] Kingma, Diederik P and Welling, Max, « Auto-encoding variational bayes », *in: arXiv preprint arXiv:1312.6114* (2013).
- [116] Kingma, Diederik P. and Ba, Jimmy, « Adam: A Method for Stochastic Optimization », *in: CoRR* abs/1412.6980 (2014).
- [117] Kingma, Diederik P., Salimans, Tim, and Welling, Max, « Improved Variational Inference with Inverse Autoregressive Flow », *in: ArXiv* abs/1606.04934 (2016), URL: <https://api.semanticscholar.org/CorpusID:11514441>.

-
- [118] Kingma, Durk P et al., « Semi-supervised learning with deep generative models », *in: Advances in neural information processing systems* 27 (2014).
 - [119] Klatt, Dennis H., « Software for a cascade/parallel formant synthesizer », *in: Journal of the Acoustical Society of America* 67 (1980), pp. 971–995, DOI: 10.1121/1.383940, URL: <http://asa.scitation.org/doi/10.1121/1.383940>.
 - [120] Kominek, John, Schultz, Tanja, and Black, Alan W, « Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. », *in: SLTU*, 2008, pp. 63–68.
 - [121] Kong, Jungil, Kim, Jaehyeon, and Bae, Jaekyoung, « HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis », *in: ArXiv abs/2010.05646* (2020), URL: <https://api.semanticscholar.org/CorpusID:222291664>.
 - [122] Kong, Jungil et al., « VITS2: Improving Quality and Efficiency of Single-Stage Text-to-Speech with Adversarial Learning and Architecture Design », *in: Proc. INTERSPEECH 2023*, 2023, pp. 4374–4378, DOI: 10.21437/Interspeech.2023-534.
 - [123] Kramer, Ernest, « Judgment of personal characteristics and emotions from non-verbal properties of speech. », *in: Psychological Bulletin* 60.4 (1963), p. 408.
 - [124] Krauss, Robert M, Freyberg, Robin, and Morsella, Ezequiel, « Inferring speakers’ physical attributes from their voices », *in: Journal of experimental social psychology* 38.6 (2002), pp. 618–625.
 - [125] Kreiman, Jody, Vanlancker-Sidtis, Diana, and Gerratt, Bruce R, « Defining and measuring voice quality », *in: ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, 2003.
 - [126] Kreiman, Jody et al., « Perceptual evaluation of voice quality: review, tutorial, and a framework for future research », *in: Journal of Speech, Language, and Hearing Research* 36.1 (1993), pp. 21–40.
 - [127] Krivokapić, Jelena, « Prosody in articulatory phonology », *in: Prosodic theory and practice* (2020), pp. 213–236.
 - [128] Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E, « Imagenet classification with deep convolutional neural networks », *in: Advances in neural information processing systems* 25 (2012).

-
- [129] Kuan, Chun-Yi et al., « Towards General-Purpose Text-Instruction-Guided Voice Conversion », *in: arXiv preprint arXiv:2309.14324* (2023).
- [130] Kumar, Abhishek, Sattigeri, Prasanna, and Balakrishnan, Avinash, « Variational inference of disentangled latent concepts from unlabeled observations », *in: arXiv preprint arXiv:1711.00848* (2017).
- [131] Kushner, Harold J. and Yin, George, « Stochastic Approximation Algorithms and Applications », *in: Applied Mathematics*, 1997.
- [132] Ladefoged, Peter and Johnson, Keith, *A course in phonetics*, Cengage learning, 2014.
- [133] Lamel, Lori F, Gauvain, Jean-Luc, Eskénazi, Mazcine, et al., « Bref, a large vocabulary spoken corpus for french », *in: training* 22.28 (1991), p. 50.
- [134] Laver, John, *Principles of phonetics*, Cambridge university press, 1994.
- [135] Laver, John, « The phonetic description of voice quality », *in: Cambridge Studies in Linguistics London* 31 (1980), pp. 1–186.
- [136] Lawson, A et al., « Survey and evaluation of acoustic features for speaker recognition », *in: 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2011, pp. 5444–5447.
- [137] LeCun, Yann, Huang, Fu Jie, and Bottou, Leon, « Learning methods for generic object recognition with invariance to pose and lighting », *in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*. Vol. 2, IEEE, 2004, pp. II–104.
- [138] Li, Lantian et al., « Deep factorization for speech signal », *in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 5094–5098.
- [139] Li, Naihan et al., « Neural speech synthesis with transformer network », *in: Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 01, 2019, pp. 6706–6713.
- [140] Li, Yuanchao et al., « Exploration of a self-supervised speech model: A study on emotional corpora », *in: 2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2023, pp. 868–875.

-
- [141] Li, Yuanning et al., « Human cortical encoding of pitch in tonal and non-tonal languages », *in: Nature communications* 12.1 (2021), p. 1161.
- [142] Liberman, Alvin M et al., « Perception of the speech code. », *in: Psychological review* 74.6 (1967), p. 431.
- [143] Lieberman, Philip and Michaels, Sheldon B, « Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech », *in: The Journal of the Acoustical Society of America* 34.7 (1962), pp. 922–927.
- [144] Lim, Dan, Jung, Sunghee, and Kim, Eesung, « JETS: Jointly training FastSpeech2 and HiFi-GAN for end to end text to speech », *in: arXiv preprint arXiv:2203.16852* (2022).
- [145] Lin, Shuyu et al., « Balancing reconstruction quality and regularisation in elbo for vaes », *in: arXiv preprint arXiv:1909.03765* (2019).
- [146] Liu, Guanghou et al., « PromptStyle: Controllable Style Transfer for Text-to-Speech with Natural Language Descriptions », *in: Proc. INTERSPEECH 2023*, 2023, pp. 4888–4892, DOI: 10.21437/Interspeech.2023-1779.
- [147] Liu, Ziwei et al., « Deep learning face attributes in the wild », *in: Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.
- [148] Livio, Mario, « Why symmetry matters », *in: Nature* 490.7421 (2012), pp. 472–473.
- [149] Locatello, Francesco et al., « Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations », *in: International Conference on Machine Learning*, 2019, pp. 4114–4124.
- [150] Lucas, James et al., « Don’t blame the elbo! a linear vae perspective on posterior collapse », *in: Advances in Neural Information Processing Systems* 32 (2019).
- [151] Lucas, James et al., *Understanding Posterior Collapse in Generative Latent Variable Models*, 2019, URL: <https://openreview.net/forum?id=r1xaVLUYuE>.
- [152] Ma, Youneng et al., « EdenTTS: A Simple and Efficient Parallel Text-to-speech Architecture with Collaborative Duration-alignment Learning », *in: Proc. INTERSPEECH 2023*, 2023, pp. 4449–4453, DOI: 10.21437/Interspeech.2023-700.

-
- [153] Makkeh, Abdullah, Gutknecht, Aaron J, and Wibral, Michael, « Introducing a differentiable measure of pointwise shared information », *in: Physical Review E* 103.3 (2021), p. 032149.
- [154] Markel, John D and Gray, AH Jr, *Linear prediction of speech*, vol. 12, Springer Science & Business Media, 2013.
- [155] Masuko, Takashi et al., « Speech synthesis using HMMs with dynamic features », *in: 1996 ieee international conference on acoustics, speech, and signal processing conference proceedings*, vol. 1, IEEE, 1996, pp. 389–392.
- [156] Mathieu, Emile et al., « Disentangling disentanglement in variational autoencoders », *in: International conference on machine learning*, PMLR, 2019, pp. 4402–4412.
- [157] Matthey, Loic et al., *dsprites: Disentanglement testing sprites dataset*, 2017.
- [158] Memon, Shahan Ali, « Acoustic Correlates of the Voice Qualifiers: A Survey », *in: arXiv preprint arXiv:2010.15869* (2020).
- [159] Moulines, Eric and Charpentier, Francis, « Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones », *in: Speech communication* 9.5-6 (1990), pp. 453–467.
- [160] Nickel, Maximillian and Kiela, Douwe, « Poincaré embeddings for learning hierarchical representations », *in: Advances in neural information processing systems* 30 (2017).
- [161] Obin, Nicolas, « MeLos: Analysis and modelling of speech prosody and speaking style », PhD thesis, Université Pierre et Marie Curie-Paris VI, 2011.
- [162] Obin, Nicolas and Roebel, Axel, « Similarity search of acted voices for automatic voice casting », *in: IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.9 (2016), pp. 1642–1651.
- [163] Obin, Nicolas, Roebel, Axel, and Bachman, Grégoire, « On automatic voice casting for expressive speech: Speaker recognition vs. speech classification », *in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2014, pp. 950–954.
- [164] Oord, Aäron van den et al., « Parallel WaveNet: Fast High-Fidelity Speech Synthesis », *in: International Conference on Machine Learning*, 2017, URL: <https://api.semanticscholar.org/CorpusID:27706557>.

-
- [165] Oord, Aaron van den et al., « Wavenet: A generative model for raw audio », *in: arXiv preprint arXiv:1609.03499* (2016).
- [166] OpenAI, « GPT-4 Technical Report », *in: ArXiv abs/2303.08774* (2023).
- [167] Paige, Brooks et al., « Learning disentangled representations with semi-supervised deep generative models », *in: Advances in neural information processing systems* 30 (2017).
- [168] Pasad, Ankita, Chou, Ju-Chieh, and Livescu, Karen, « Layer-wise analysis of a self-supervised speech representation model », *in: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2021, pp. 914–921.
- [169] Pasad, Ankita, Shi, Bowen, and Livescu, Karen, « Comparative layer-wise analysis of self-supervised speech models », *in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.
- [170] Pasad, Ankita et al., « What do self-supervised speech models know about words? », *in: arXiv preprint arXiv:2307.00162* (2023).
- [171] Paysan, Pascal et al., « A 3D face model for pose and illumination invariant face recognition », *in: 2009 sixth IEEE international conference on advanced video and signal based surveillance*, Ieee, 2009, pp. 296–301.
- [172] Paz, Karoline Evangelista da Silva et al., « Training for perceptive-auditory voice analysis: scope review », *in: Audiology-Communication Research* 28 (2023), e2768.
- [173] Pearl, Judea, *Causality*, Cambridge university press, 2009.
- [174] Peebles, William et al., « The hessian penalty: A weak prior for unsupervised disentanglement », *in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, Springer, 2020, pp. 581–597.
- [175] Ping, Wei, Peng, Kainan, and Chen, Jitong, « Clarinet: Parallel wave generation in end-to-end text-to-speech », *in: arXiv preprint arXiv:1807.07281* (2018).
- [176] Ping, Wei et al., « Deep voice 3: Scaling text-to-speech with convolutional sequence learning », *in: arXiv preprint arXiv:1710.07654* (2017).

-
- [177] Polyak, Adam et al., « Speech Resynthesis from Discrete Disentangled Self-Supervised Representations », *in: Proc. Interspeech 2021*, 2021, pp. 3615–3619, DOI: 10 . 21437/Interspeech.2021-475.
- [178] Poyatos, Fernando, « Paralinguistic qualifiers: Our many voices », *in: Language & Communication* 11.3 (1991), pp. 181–195.
- [179] Qian, Kaizhi et al., « Autovc: Zero-shot voice style transfer with only autoencoder loss », *in: International Conference on Machine Learning*, PMLR, 2019, pp. 5210–5219.
- [180] Qian, Kaizhi et al., « Unsupervised speech decomposition via triple information bottleneck », *in: International Conference on Machine Learning*, PMLR, 2020, pp. 7836–7846.
- [181] Raitio, Tuomo, Li, Jiangchuan, and Seshadri, Shreyas, « Hierarchical prosody modeling and control in non-autoregressive parallel neural TTS », *in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7587–7591.
- [182] Raitio, Tuomo, Rasipuram, Ramya, and Castellani, Dan, « Controllable neural text-to-speech synthesis using intuitive prosodic features », *in: arXiv preprint arXiv:2009.06775* (2020).
- [183] Ramesh, Aditya et al., « Hierarchical Text-Conditional Image Generation with CLIP Latents », *in: ArXiv abs/2204.06125* (2022).
- [184] Reed, Scott E et al., « Deep visual analogy-making », *in: Advances in neural information processing systems* 28 (2015).
- [185] Ren, Yi et al., « FastSpeech 2: Fast and high-quality end-to-end text to speech », *in: arXiv preprint arXiv:2006.04558* (2020).
- [186] Ren, Yi et al., « FastSpeech: Fast, robust and controllable text to speech », *in: Advances in neural information processing systems* 32 (2019).
- [187] Rezende, Danilo Jimenez, Mohamed, Shakir, and Wierstra, Daan, « Stochastic backpropagation and approximate inference in deep generative models », *in: International conference on machine learning*, PMLR, 2014, pp. 1278–1286.
- [188] Ridgeway, Karl and Mozer, Michael C, « Learning deep disentangled embeddings with the f-statistic loss », *in: Advances in neural information processing systems* 31 (2018).

-
- [189] Rolinek, Michal, Zietlow, Dominik, and Martius, Georg, « Variational autoencoders pursue pca directions (by accident) », *in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12406–12415.
- [190] Rose, Phil, *Forensic speaker identification*, cRc Press, 2002.
- [191] Rybkin, Oleh, Daniilidis, Kostas, and Levine, Sergey, « Simple and effective VAE training with calibrated decoders », *in: International Conference on Machine Learning*, PMLR, 2021, pp. 9179–9189.
- [192] Sagisaka, Yoshinori, « Speech synthesis by rule using an optimal selection of non-uniform synthesis units », *in: ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, IEEE Computer Society, 1988, pp. 679–680.
- [193] Saito, Yuki et al., « Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors », *in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 5274–5278.
- [194] Sambur, Marvin, « Selection of acoustic features for speaker identification », *in: IEEE Transactions on Acoustics, Speech, and Signal Processing* 23.2 (1975), pp. 176–182.
- [195] San Segundo, Eugenia and Mompean, Jose A, « A simplified vocal profile analysis protocol for the assessment of voice quality and speaker similarity », *in: Journal of Voice* 31.5 (2017), 644–e11.
- [196] San Segundo, Eugenia et al., « The use of the Vocal Profile Analysis for speaker characterization: Methodological proposals », *in: Journal of the International Phonetic Association* 49.3 (2019), pp. 353–380.
- [197] Scherer, Klaus R, « Personality inference from voice quality: The loud voice of extroversion », *in: European Journal of Social Psychology* 8.4 (1978), pp. 467–487.
- [198] Schneider, Steffen et al., « wav2vec: Unsupervised pre-training for speech recognition », *in: arXiv preprint arXiv:1904.05862* (2019).
- [199] Schuller, Björn et al., « Medium-term speaker states—A review on intoxication, sleepiness and the first challenge », *in: Computer Speech & Language* 28.2 (2014), pp. 346–374.
- [200] Schuller, Björn et al., « The INTERSPEECH 2011 speaker state challenge », *in: Proc. INTERSPEECH 2011, Florence, Italy*, 2011.

-
- [201] Schuller, Björn et al., « The interspeech 2012 speaker trait challenge », *in: INTER-SPEECH 2012, Portland, OR, USA*, 2012.
- [202] Schultz, Tanja, « Speaker characteristics », *in: Speaker Classification I: Fundamentals, Features, and Methods* (2007), pp. 47–74.
- [203] Schweinberger, Stefan R. et al., « Speaker Perception », *in: Wiley Interdisciplinary Reviews: Cognitive Science* 5.1 (Jan. 2014), pp. 15–25, ISSN: 19395078, DOI: 10.1002/wcs.1261, (visited on 07/23/2023).
- [204] Sepiarskaia, Anna, Kiseleva, Julia, Rijke, Maarten de, et al., « Evaluating disentangled representations », *in: arXiv preprint arXiv:1910.05587* (2019).
- [205] Shen, Jonathan et al., « Natural tts synthesis by conditioning wavenet on mel spectrogram predictions », *in: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 4779–4783.
- [206] Simonyan, Karen and Zisserman, Andrew, « Very deep convolutional networks for large-scale image recognition », *in: arXiv preprint arXiv:1409.1556* (2014).
- [207] Skerry-Ryan, RJ et al., « Towards end-to-end prosody transfer for expressive speech synthesis with tacotron », *in: international conference on machine learning*, PMLR, 2018, pp. 4693–4702.
- [208] Snyder, David et al., « X-vectors: Robust dnn embeddings for speaker recognition », *in: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 5329–5333.
- [209] Song, Lin, Langfelder, Peter, and Horvath, Steve, « Comparison of co-expression measures: mutual information, correlation, and model based indices », *in: BMC bioinformatics* 13.1 (2012), pp. 1–21.
- [210] Stevens, Kenneth N, « Sources of inter-and intra-speaker variability in the acoustic properties of speech sounds », *in: Proceedings of the Seventh International Cons. Phonetic Sciences* (1971), pp. 206–232.
- [211] Sun, Guangzhi et al., « Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis », *in: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2020, pp. 6264–6268.
- [212] Sun, Lifa et al., « Phonetic posteriorgrams for many-to-one voice conversion without parallel data training », *in: 2016 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2016, pp. 1–6.

-
- [213] Suter, Raphael et al., « Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness », *in: International Conference on Machine Learning*, PMLR, 2019, pp. 6056–6065.
- [214] Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc V, « Sequence to sequence learning with neural networks », *in: Advances in neural information processing systems* 27 (2014).
- [215] Tabet, Youcef and Boughazi, Mohamed, « Speech synthesis techniques. A survey », *in: International Workshop on Systems, Signal Processing and their Applications, WOSSPA*, IEEE, 2011, pp. 67–70.
- [216] Tan, Xu et al., « A survey on neural speech synthesis », *in: arXiv preprint arXiv:2106.15561* (2021).
- [217] Taylor, Paul, *Text-to-speech synthesis*, Cambridge university press, 2009.
- [218] Tinsley, Howard E and Weiss, David J, « Interrater reliability and agreement of subjective judgments. », *in: Journal of Counseling Psychology* 22.4 (1975), p. 358.
- [219] Tishby, Naftali, Pereira, Fernando C, and Bialek, William, « The information bottleneck method », *in: arXiv preprint physics/0004057* (2000).
- [220] Tokuda, Keiichi, Kobayashi, Takao, and Imai, Satoshi, « Speech parameter generation from HMM using dynamic features », *in: 1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, IEEE, 1995, pp. 660–663.
- [221] Tokui, Seiya and Sato, Issei, « Disentanglement analysis with partial information decomposition », *in: arXiv preprint arXiv:2108.13753* (2021).
- [222] Trask, Robert Lawrence, *A dictionary of phonetics and phonology*, Routledge, 1996.
- [223] Variani, Ehsan et al., « Deep neural networks for small footprint text-dependent speaker verification », *in: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2014, pp. 4052–4056.
- [224] Vaswani, Ashish et al., « Attention is All you Need », *in: NIPS*, 2017.
- [225] Voiers, William D, « Perceptual bases of speaker identity », *in: The Journal of the Acoustical Society of America* 36.6 (1964), pp. 1065–1073.
- [226] Wang, Chengyi et al., « Neural codec language models are zero-shot text to speech synthesizers », *in: arXiv preprint arXiv:2301.02111* (2023).

-
- [227] Wang, Disong et al., « VQMIVC: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion », *in: arXiv preprint arXiv:2106.10132* (2021).
- [228] Wang, Qiqi et al., « Drvc: A framework of any-to-any voice conversion with self-supervised learning », *in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 3184–3188.
- [229] Wang, Yuxuan et al., « Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis », *in: International conference on machine learning*, PMLR, 2018, pp. 5180–5189.
- [230] Wang, Yuxuan et al., « Tacotron: Towards end-to-end speech synthesis », *in: arXiv preprint arXiv:1703.10135* (2017).
- [231] Warren, Richard M, « Verbal transformation effect and auditory perceptual mechanisms. », *in: Psychological Bulletin* 70.4 (1968), p. 261.
- [232] Weiss, Benjamin, Estival, Dominique, and Stiefelhagen, Ulrike, « Non-Experts’ Perceptual Dimensions of Voice Assessed by Using Direct Comparisons », *in: Acta Acustica united with Acustica* 104.1 (2018), pp. 174–184.
- [233] Williams, Jennifer et al., « Learning disentangled phone and speaker representations in a semi-supervised VQ-VAE paradigm », *in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 7053–7057.
- [234] Williams, Paul L and Beer, Randall D, « Nonnegative decomposition of multivariate information », *in: arXiv preprint arXiv:1004.2515* (2010).
- [235] Wilson, D Kenneth, « Voice problems of children », *in: (No Title)* (1987).
- [236] Yamamoto, Ryuichi, Song, Eunwoo, and Kim, Jae-Min, « Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram », *in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6199–6203.
- [237] Yang, Shu-wen et al., « SUPERB: Speech Processing Universal PERformance Benchmark », *in: Proc. Interspeech 2021*, 2021, pp. 1194–1198, DOI: 10.21437/Interspeech.2021-1775.

-
- [238] Yoshimura, Takayoshi et al., « Duration modeling for HMM-based speech synthesis », *in: ICSLP*, vol. 98, 1998, pp. 29–32.
- [239] Yoshimura, Takayoshi et al., « Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis », *in: Sixth European conference on speech communication and technology*, 1999.
- [240] Yuan, Siyang et al., « Improving Zero-Shot Voice Style Transfer via Disentangled Representation Learning », *in: International Conference on Learning Representations*, 2021, URL: <https://openreview.net/forum?id=TgSVWxw22FQ>.
- [241] Zen, Heiga and Sak, Haşim, « Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis », *in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, pp. 4470–4474.
- [242] Zen, Heiga, Senior, Andrew, and Schuster, Mike, « Statistical parametric speech synthesis using deep neural networks », *in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2013, pp. 7962–7966.
- [243] Zen, Heiga, Tokuda, Keiichi, and Black, Alan W, « Statistical parametric speech synthesis », *in: speech communication* 51.11 (2009), pp. 1039–1064.
- [244] Zhang, Jing-Xuan, Ling, Zhen-Hua, and Dai, Li-Rong, « Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations », *in: IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2019), pp. 540–552.
- [245] Zhang, Mingyang et al., « Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet », *in: arXiv preprint arXiv:1903.12389* (2019).
- [246] Zhang, Olivier et al., « An extension of disentanglement metrics and its application to voice », *in: Proc. INTERSPEECH 2023*, 2023, pp. 2878–2882, DOI: 10.21437/Interspeech.2023-383.
- [247] Zhang, Olivier et al., « diSpeech: A Synthetic Toy Dataset for Speech Disentangling », *in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 8557–8561.
- [248] Zhao, Shengjia, Song, Jiaming, and Ermon, Stefano, « Infovae: Information maximizing variational autoencoders », *in: arXiv preprint arXiv:1706.02262* (2017).

-
- [249] Zhu, Xinqi, Xu, Chang, and Tao, Dacheng, « Learning disentangled representations with latent variation predictability », *in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, Springer, 2020, pp. 684–700.
- [250] Zhu, Xinqi, Xu, Chang, and Tao, Dacheng, « Where and What? Examining Interpretable Disentangled Representations », *in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5861–5870.

Titre : Méthodes neuronales pour le traitement de la parole : vers le démêlage des attributs de la voix

Mot clés : analyse de la voix, réseaux de neurones, démêlage

Résumé : Les récentes avancées de l'apprentissage profond ont mené à des résultats sans précédent dans une grande variété de tâches et de modalités. Un nombre grandissant de systèmes parvenant à des performances proches de l'humain en analyse (transcription, reconnaissance du locuteur) et en génération de la parole (conversion, synthèse vocale) sont proposés. De telles solutions émergent dans l'industrie, et commencent à atteindre le grand public. Cependant, la complexité et la taille grandissantes des réseaux de neurones induisent un manque important d'interprétabilité. De plus, les représentations profondes ne sont pas encouragées pour être structurées. C'est pourquoi l'apprentissage de représentations dites

démêlées a fait son apparition, et a pour priorité la structuration des représentations apprises, en rapport avec les facteurs génératifs des données, et si possible alignées avec la perception humaine. Un tel paradigme a le potentiel pour reconnaître les attributs de la parole (identité du locuteur, émotion), pouvant alors être exploité dans la synthèse vocale. À noter que le démêlage est encore un domaine de recherche récent, nécessitant des données simples et synthétiques pour être développé. Ainsi, cette thèse vise à combler le fossé entre le traitement de la parole et le démêlage, en exploitant des modèles de démêlage à l'état de l'art pour identifier les attributs de la parole de manière automatique, et à terme améliorer le contrôle de la synthèse vocale.

Title: Neural methods for speech processing: towards speech attributes disentanglement

Keywords: voice analysis, neural networks, disentanglement

Abstract: The past few years' advances in deep learning have brought unprecedented performances in a wide range of tasks and modalities. An increasing number of close-to-human accuracy speech analysis (e.g., ASR, ASV) and near-natural speech generation (e.g., conversion, TTS) models are proposed. Such solutions are emerging in the industry and are reaching the public. Nevertheless, the growing complexity and size of neural networks are causing a significant lack of interpretability. Moreover, well-structured representations are, by design, not enforced. Hence, disentangled representations have emerged, which aim to prioritize

representation structuring, related to data explanatory factors, hopefully aligned with human perceptions. Such a paradigm can be expected to properly recognize speech attributes (e.g., speaker identity, gender, and emotion), which may be leveraged for speech synthesis. However, disentanglement learning is a research topic still in its early stages, needing simple and synthetic data to be developed. Thus, this thesis endeavors to bridge the gap between speech processing and disentanglement, examining how state-of-the-art disentangling models can be leveraged to automatically recover speech attributes, and ultimately improve control over synthesized speech.