Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Apprentissage de plusieurs représentations d'attributs au niveau de l'énoncé avec un encodeur de parole unifié

Maryem Bouziane Salima Mdhaffar Yannick Estève
Avignon Université, LIA, France
maryem.bouziane2@univ-avignon.fr, salima.mdhaffar@univ-avignon.fr,
yannick.esteve@univ-avignon.fr

RÉSUMÉ

Les modèles de fondation pour la parole entraînés par apprentissage auto-supervisé produisent des représentations génériques de la parole, capables de soutenir un large éventail de tâches de traitement de la parole. Lorsqu'ils sont ensuite adaptés par apprentissage supervisé, ces modèles peuvent atteindre de fortes performances sur des tâches aval spécifiques. Des approches récentes de post-entraînement, telles que SAMU-XLSR et SONAR, alignent les représentations de la parole sur des représentations sémantiques au niveau de l'énoncé, permettant des applications multimodales (parole–texte) et multilingues efficaces. Alors que les modèles de fondation pour la parole apprennent typiquement des embeddings contextuels au niveau de la trame acoustique, ces méthodes apprennent des représentations au niveau de l'énoncé. Dans ce travail, nous étendons ce paradigme à des attributs arbitraires au niveau de l'énoncé et proposons un cadre unifié de post-entraînement permettant à un unique modèle de fondation pour la parole de générer plusieurs types de représentations au niveau de l'énoncé. Nous démontrons l'efficacité de cette approche en apprenant conjointement des représentations sémantiques et des représentations de locuteur, puis en les évaluant sur des tâches de recherche multilingue à partir de la parole et de reconnaissance du locuteur.

ABSTRACT

Learning Multiple Utterance-Level Attribute Representations with a Unified Speech Encoder

Speech foundation models trained with self-supervised learning produce generic speech representations that support a wide range of speech processing tasks. When further adapted with supervised learning, these models can achieve strong performance on specific downstream tasks. Recent post-training approaches, such as SAMU-XLSR and SONAR, align speech representations with utterance-level semantic representations, enabling effective multimodal (speech–text) and multilingual applications. While speech foundation models typically learn contextual embeddings at the acoustic frame level, these methods learn representations at the utterance level. In this work, we extend this paradigm to arbitrary utterance-level attributes and propose a unified post-training framework that enables a single speech foundation model to generate multiple types of utterance-level representations. We demonstrate the effectiveness of this approach by jointly learning semantic and speaker representations and evaluating them on multilingual speech retrieval and speaker recognition tasks.

MOTS-CLÉS : apprentissage multitâche, encodeur de parole, représentation sémantique, représentation de locuteur.

KEYWORDS: Multi-task learning, speech encoder, semantic representation, speaker representation.

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 134–144.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

1 Introduction

Les avancées récentes en apprentissage auto-supervisé de représentations de la parole ont conduit à l'émergence de grands modèles de fondation pour la parole capables de capturer, à partir de l'audio brut, des informations acoustiques et linguistiques riches. Des modèles tels que wav2vec 2.0 (Baevski *et al.*, 2020), HuBERT (Hsu *et al.*, 2021) et w2v-BERT (Chung *et al.*, 2021) sont entraînés sur des volumes massifs de parole non annotée et apprennent des représentations contextualisées, qui peuvent être adaptées à un large éventail de tâches aval (Parcollet *et al.*, 2024; Yang *et al.*, 2021). Ces modèles produisent typiquement des représentations acoustiques au niveau de la trame, très efficaces pour des tâches telles que la reconnaissance automatique de la parole ou la traduction de la parole. Au-delà des représentations au niveau de la trame, l'apprentissage d'embeddings de parole au niveau de l'énoncé suscite un intérêt croissant, car ils permettent de capturer des informations de plus haut niveau, telles que la sémantique ou des caractéristiques du locuteur. De telles représentations sont particulièrement utiles pour des tâches incluant la recherche à partir de la parole, la recherche multimodale, la vérification du locuteur, ou la compréhension conversationnelle. Des approches récentes de post-entraînement ont montré que des encodeurs de parole peuvent être alignés sur des espaces d'embeddings sémantiques issus du texte, ce qui permet des applications multilingues et multimodales telles que la recherche parole–texte et la recherche de traduction de la parole.

Un exemple notable est le cadre SENSE (Mdhaffar *et al.*, 2025), proche du cadre SONAR de Meta (Duquenne *et al.*, 2023), qui apprend des embeddings sémantiques de parole selon un paradigme de distillation de connaissances enseignant–étudiant (Hinton *et al.*, 2015) : dans cette approche, un modèle d'embeddings textuels pré-entraîné fournit des cibles sémantiques, tandis qu'un encodeur de parole est entraîné à projeter des énoncés de parole dans le même espace sémantique. Cet alignement permet aux représentations de parole de capturer directement le sens d'un énoncé, indépendamment de la langue, les rendant ainsi adaptées aux tâches de recherche sémantique multilingue (Mdhaffar *et al.*, 2025; Duquenne *et al.*, 2023; Khurana *et al.*, 2022).

Cependant, aligner les représentations de parole exclusivement sur des embeddings sémantiques introduit une limitation importante : l'information paralinguistique présente dans la parole peut être atténuée. En particulier, des attributs tels que l'identité du locuteur, l'émotion ou le style de parole ne sont pas préservés lorsque la représentation est optimisée uniquement pour correspondre à des embeddings sémantiques textuels. Cela soulève une question importante : un unique encodeur de parole peut-il apprendre des représentations qui capturent simultanément plusieurs attributs au niveau de l'énoncé ?

Dans ce travail, nous proposons un cadre unifié de post-entraînement qui permet à un unique modèle de fondation pour la parole de produire plusieurs représentations au niveau de l'énoncé correspondant à différents attributs. Notre approche étend le paradigme enseignant–étudiant en introduisant plusieurs signaux de supervision spécifiques aux tâches, chacun définissant un espace d'embeddings cible pour un attribut particulier. Un encodeur de parole partagé est entraîné conjointement à s'aligner sur ces différentes cibles via des branches de projection spécifiques à chaque tâche. Pour démontrer l'efficacité de ce cadre, nous nous concentrons sur l'apprentissage de deux attributs complémentaires au niveau de l'énoncé : des représentations sémantiques, obtenues par alignement avec des embeddings textuels multilingues, et des représentations de locuteur, obtenues par supervision à partir d'un modèle de vérification du locuteur pré-entraîné.

Nous évaluons le modèle proposé sur deux tâches représentatives. Premièrement, nous mesurons la qualité de la représentation sémantique à l'aide de benchmarks de recherche de traduction multilingue

parole–parole et *parole–texte*. Deuxièmement, nous évaluons la représentation de locuteur sur la tâche de vérification du locuteur VoxCeleb.

Dans cet article, nous apportons les contributions suivantes :

1. Nous introduisons un cadre enseignant–étudiant multi-tâches général pour apprendre, à partir d'un encodeur de parole partagé, plusieurs représentations d'attributs au niveau de l'énoncé.
2. Nous montrons que les représentations sémantiques et de locuteur peuvent être apprises conjointement sans dégrader significativement les performances de l'un ou l'autre attribut.
3. Nous fournissons une analyse de l'utilisation des couches selon les tâches, montrant comment les informations sémantiques et de locuteur se répartissent différemment au sein de l'encodeur partagé.
4. Nous distribuons le code associé à cette approche ainsi que le modèle multi-tâche entraîné.

2 Méthode proposée

Notre travail étend le paradigme de distillation enseignant–étudiant utilisé dans le cadre SENSE (Mdhaffar *et al.*, 2025), une solution open source conçue pour apprendre des représentations sémantiques de la parole au niveau de l'énoncé. Le cadre SENSE est dérivé de SAMU-XLSR (Khurana *et al.*, 2022), une approche proche du cadre SONAR de Meta.

2.1 Le modèle SENSE

SENSE repose sur un paradigme de distillation de connaissances enseignant–étudiant, dont l'objectif principal est d'aligner la parole et le texte dans un espace sémantique partagé, indépendant de la langue. Plus précisément, l'architecture utilise un modèle pré-entraîné d'embeddings textuels jouant le rôle d'enseignant et un encodeur de parole jouant le rôle d'étudiant. L'encodeur de parole est initialisé à partir d'un encodeur de parole auto-supervisé (SSL). Côté texte, l'enseignant est un générateur d'embeddings de phrases indépendant de la langue, tel que LaBSE (Feng *et al.*, 2022) ou BGE-M3 (Chen *et al.*, 2024).

Pour aligner les représentations de parole avec des représentations sémantiques au niveau de l'énoncé, les représentations au niveau de la trame issues de la dernière couche de l'encodeur de parole sont d'abord agrégées par une couche de *pooling* attentionnel. Le vecteur obtenu est ensuite passé dans une projection linéaire et une fonction d'activation, ce qui produit une représentation de parole au niveau de l'énoncé pour chaque segment de parole.

L'ensemble du modèle, composé de l'encodeur de parole initial, de la couche de *pooling* attentionnel et de la projection linéaire, est entraîné à maximiser la similarité cosinus entre le vecteur de parole au niveau de l'énoncé et l'embedding textuel au niveau de la phrase fourni par le modèle enseignant. Tout au long de l'entraînement, l'encodeur de texte reste strictement gelé afin de préserver sa structure sémantique, tandis que l'encodeur de parole est optimisé. En optimisant directement la similarité cosinus, les connaissances sémantiques sont transférées de la modalité texte vers la modalité parole, ce qui permet à l'encodeur de parole d'apprendre des représentations orientées vers le sens au niveau de l'énoncé.

2.2 Apprentissage de plusieurs représentations d'attributs au niveau de l'énoncé

Nous étendons le paradigme enseignant–étudiant afin d'apprendre, à partir d'un encodeur de parole partagé, plusieurs représentations au niveau de l'énoncé correspondant à différents attributs. Soit $\mathcal{T}$ l'ensemble des attributs cibles et $\tau \in \mathcal{T}$ un attribut particulier. Étant donné un signal de parole en entrée, l'encodeur de parole SSL pré-entraîné produit une séquence de représentations cachées à différentes couches : $H^{(\ell)} \in R^{T \times D}$, où T est le nombre de trames temporelles, D la dimension cachée des représentations de l'encodeur, et ℓ l'indice de couche.

Pour chaque attribut τ, une branche spécifique à la tâche est attachée à l'encodeur partagé afin de produire un embedding au niveau de l'énoncé, aligné avec la représentation enseignante correspondante. Pour chaque représentation de couche sélectionnée $H^{(\ell)}$, on applique une projection linéaire spécifique à l'attribut τ : $\tilde{H}_\tau^{(\ell)} = H^{(\ell)} W_\tau^{(\ell)\top} + b_\tau^{(\ell)}$.

Cette projection projette les représentations de l'encodeur partagé dans un espace de caractéristiques associé à l'attribut τ. L'objectif de cette transformation est de limiter le degré d'adaptation requis de la part de l'encodeur de parole partagé. Plutôt que de contraindre l'encodeur à produire directement des représentations adaptées à tous les attributs cibles, le modèle apprend des projections spécifiques à chaque attribut, qui transforment l'espace de représentation partagé vers l'espace d'embeddings requis par chaque attribut. De cette manière, l'encodeur de parole peut conserver une représentation générique du signal d'entrée, tandis que chaque branche d'attribut adapte cette représentation partagée à son propre espace cible. Cette séparation aide l'encodeur à rester agnostique aux tâches et réduit les interférences potentielles entre attributs lorsque le modèle est entraîné avec plusieurs signaux de supervision. Cette transformation projette les représentations de l'encodeur dans un espace de caractéristiques spécifique à la tâche, associé à l'attribut τ.

Différents attributs peuvent reposer sur des régions différentes de l'encodeur. Pour capturer ce comportement, le modèle apprend un score d'importance scalaire $s_{\tau,\ell}$ pour chaque couche. Ce mécanisme de pondération des couches n'est pas présent dans le cadre SENSE. Ces scores sont convertis en poids d'interpolation normalisés au moyen d'une fonction softmax : $\lambda_{\tau,\ell} = \frac{\exp(s_{\tau,\ell})}{\sum_{j=1}^{n} \exp(s_{\tau,j})}, \qquad \sum_{\ell=1}^{n} \lambda_{\tau,\ell} = 1.$

Les représentations projetées sont ensuite combinées via une somme pondérée : $\hat{Z}_\tau = \sum_{\ell=1}^{n} \lambda_{\tau,\ell} \tilde{H}_\tau^{(\ell)}$.

Une normalisation de couche est ensuite appliquée : $Z_\tau = \mathrm{LayerNorm}(\hat{Z}_\tau)$. La séquence au niveau de la trame Z_τ est agrégée en une unique représentation au niveau de l'énoncé à l'aide d'un mécanisme de *pooling* attentionnel spécifique à l'attribut : $p_\tau = \mathrm{AttentionPooling}_\tau(Z_\tau)$.

Selon l'espace de représentation cible associé à l'attribut τ, une projection linéaire optionnelle peut être appliquée.

L'embedding final au niveau de l'énoncé produit par la branche d'attribut τ est normalisé en norme ℓ_2, puis aligné avec l'embedding enseignant correspondant à l'aide d'un objectif de similarité cosinus. La Figure 1 résume cette approche.

Le modèle est entraîné dans un cadre d'apprentissage multi-tâches où chaque attribut cible est considéré comme une tâche distincte. Pour chaque attribut τ, une branche dédiée est instanciée. Tous les paramètres, y compris ceux de l'encodeur SSL, sont optimisés conjointement pendant l'entraînement.

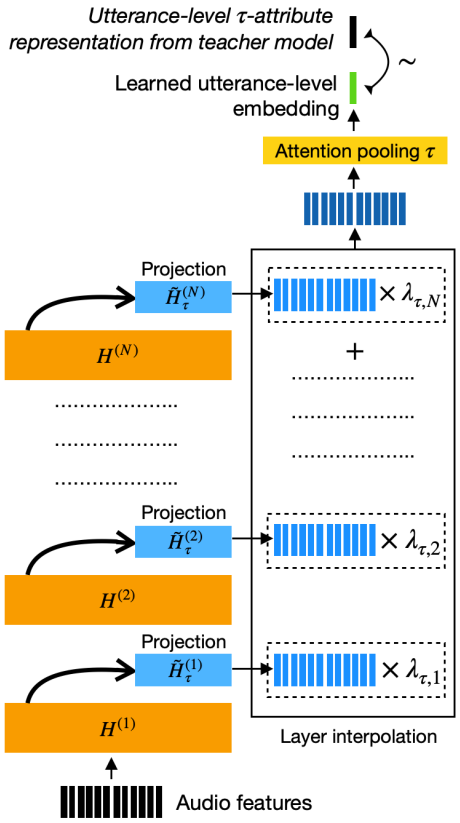

FIGURE 1 – Branche spécifique à l'attribut utilisée pour apprendre une représentation au niveau de l'énoncé pour l'attribut τ dans le cadre enseignant–étudiant. Les représentations de couches issues de l'encodeur SSL partagé sont projetées, combinées à l'aide de poids d'interpolation apprenables $\lambda_{\tau,\ell}$, puis agrégées par *pooling* attentionnel afin de produire un embedding aligné avec la représentation enseignante.

3 Expériences

3.1 Configuration expérimentale

Nous considérons deux tâches dans notre apprentissage multi-tâches enseignant–étudiant : (1) une tâche sémantique, centrée sur l'apprentissage de représentations du contenu sémantique agnostiques à la langue, et (2) une tâche orientée locuteur, visant à caractériser l'information de locuteur au sein de la même architecture unifiée. La branche sémantique est entraînée pour aligner la représentation de parole au niveau de l'énoncé avec l'embedding sémantique BGE-M3 (Chen *et al.*, 2024) correspondant, tandis que la branche locuteur est entraînée pour aligner la représentation de parole au niveau de l'énoncé avec un modèle d'embeddings de locuteur ECAPA-TDNN (Desplanques *et al.*, 2020) entraîné. Le modèle ECAPA-TDNN [1] utilisé dans ce travail est entraîné sur les jeux de données VoxCeleb 1 (Nagrani *et al.*, 2020) et VoxCeleb 2 (Chung *et al.*, 2018). Les deux modèles enseignants restent gelés pendant tout l'entraînement. L'encodeur de parole est initialisé avec w2v-BERT 2.0 (Barrault *et al.*, 2023). Nous avons implémenté notre modèle multi-tâches sous SpeechBrain (Ravanelli

1. https ://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

et al., 2024) à partir de l'implémentation open source de SENSE (Mdhaffar *et al.*, 2025) [2], puis l'avons entraîné sur le jeu de données Common Voice 19 (Ardila *et al.*, 2020) en utilisant les 83 langues supportées par BGE-M3, soit 8 250 heures de parole. Comme dans (Mdhaffar *et al.*, 2025), nous utilisons uniquement le sous-ensemble *validated* d'entraînement de Common Voice. Afin de tenir compte du déséquilibre entre langues, les exemples d'entraînement sont échantillonnés à l'aide d'une stratégie d'échantillonnage pondéré (Khurana *et al.*, 2022). Nous affinons l'encodeur w2v-BERT 2.0 partagé avec Adam, avec un taux d'apprentissage de 10^{-5}. Les modules spécifiques aux tâches sont optimisés séparément avec Adadelta, avec un taux d'apprentissage initial de 1.5. L'entraînement est réalisé avec une taille de lot de 20, pour l'entraînement comme pour la validation. Le modèle est entraîné pendant 350 000 itérations en utilisant 8 GPU H100.

3.2 Évaluation

Notre objectif principal est d'évaluer si le modèle multi-tâches peut apprendre efficacement à la fois des représentations sémantiques et des représentations de locuteur, sans dégrader les performances de l'une ou l'autre tâche par rapport à un entraînement mono-tâche. Afin de situer ses performances, nous évaluons le modèle sur des tâches sémantiques et des tâches liées au locuteur.

— **Tâche sémantique :** nous évaluons les performances du modèle en recherche multilingue et multimodale.
— **Tâche locuteur :** nous évaluons le modèle sur une tâche de vérification du locuteur, dont l'objectif est de déterminer si deux échantillons audio proviennent du même locuteur.

Pour l'évaluation sémantique, nous comparons notre modèle multi-tâches à deux modèles de l'état de l'art, SENSE et SONAR, tous deux conçus pour apprendre des représentations de parole porteuses de sémantique. Pour la vérification du locuteur, nous comparons notre modèle au modèle d'embeddings de locuteur ECAPA-TDNN. En complément, nous entraînons une base mono-tâche en ne conservant que la branche locuteur de notre architecture. Cette base nous permet d'évaluer l'apport de l'apprentissage multi-tâches et de quantifier l'impact d'une supervision conjointe sémantique et locuteur sur les performances de vérification du locuteur.

3.2.1 Recherche de traduction multilingue et multimodale

Afin d'évaluer si l'entraînement multi-tâches proposé préserve la qualité sémantique des représentations de parole apprises, nous évaluons le modèle sur une tâche de recherche de traduction multilingue. Pour chaque expérience, nous définissons un ensemble de requêtes (*query set*) et un ensemble de recherche (*search set*). L'ensemble de requêtes contient toujours des énoncés de parole, tandis que l'ensemble de recherche contient, selon la condition d'évaluation, soit de la parole, soit du texte dans une autre langue. L'objectif est de retrouver la traduction correcte de chaque requête parmi les candidats de l'ensemble de recherche. Chaque énoncé requête est encodé en un embedding au niveau de l'énoncé à l'aide de la branche sémantique du modèle multi-tâches. Les éléments de l'ensemble de recherche sont encodés en fonction de leur modalité. Lorsque l'ensemble de recherche contient de la parole, des embeddings de parole au niveau de l'énoncé sont extraits avec le même encodeur de parole. Lorsqu'il contient du texte, chaque phrase est représentée avec l'encodeur de texte gelé correspondant : BGE-M3 pour SENSE et pour notre modèle multi-tâches, et l'encodeur de texte SONAR pour SONAR, basé sur le modèle NLLB (Costa-Jussà *et al.*, 2022). Après normalisation par soustraction de la moyenne, la recherche est réalisée en comparant chaque embedding requête

2. https://github.com/speechbrain/speechbrain/tree/develop/recipes/CommonVoice/SENSE

à l'ensemble des embeddings candidats de l'ensemble de recherche via la similarité cosinus. Les performances sont rapportées avec Recall@1. Nous considérons les configurations d'évaluation suivantes. Pour (1) la recherche *parole* → *parole*, nous utilisons VoxPopuli (Wang *et al.*, 2021), où l'ensemble de requêtes et l'ensemble de recherche contiennent des énoncés de parole dans des langues différentes. Pour (2) la recherche *parole* → *texte*, nous utilisons MTEDx (Salesky *et al.*, 2021), où l'ensemble de requêtes contient des énoncés de parole et l'ensemble de recherche contient les phrases textuelles traduites correspondantes dans une autre langue. Afin d'évaluer la capacité de généralisation des modèles à des langues non vues et à faibles ressources, nous utilisons également le jeu de données FLEURS (Conneau *et al.*, 2023).

3.2.2 Vérification du locuteur

Nous évaluons la qualité de l'information liée au locuteur au moyen d'une expérience de vérification du locuteur, en suivant le protocole d'évaluation VoxCeleb1-O. Les embeddings de locuteur sont extraits à partir des énoncés d'enrôlement et de test, soit via les branches locuteur de nos modèles (entraînés en mode mono-tâche ou multi-tâches), soit via les représentations obtenues avec le modèle ECAPA-TDNN. Les scores de vérification pour chaque paire d'essai sont calculés par similarité cosinus. Les performances sont rapportées en termes de taux d'erreur égal (EER) et de fonction de coût minimale de détection normalisée (MinDCF), avec $P_{\text{target}} = 0.01$ et $C_{\text{FA}} = C_{\text{Miss}} = 1$.

3.3 Résultats

Cette section présente les résultats expérimentaux pour les évaluations sémantique et locuteur décrites ci-dessus. Dans les tableaux de résultats, nous utilisons la notation suivante pour plus de clarté : SONAR désigne le modèle SONAR de Meta AI introduit dans (Barrault *et al.*, 2023) ; Att(sem) correspond au modèle SENSE entraîné avec un objectif sémantique tel que décrit dans (Mdhaffar *et al.*, 2025) ; Att(spk) désigne la variante mono-tâche de notre architecture entraînée uniquement avec l'objectif locuteur ; et Att(sem+spk) représente le modèle multi-tâches proposé, entraîné conjointement avec une supervision sémantique et locuteur.

3.3.1 Recherche de traduction multilingue et multimodale

Le Tableau 1 présente les scores R@1 pour la recherche de traduction *parole* → *parole* sur VoxPopuli pour les trois modèles SONAR, Att(sem) et Att(sem+spk).
Att(sem) et Att(sem+spk) utilisent un unique encodeur de parole multilingue partagé pour l'ensemble des langues. SONAR, en revanche, utilise 37 encodeurs spécifiques à la langue. Dans nos expériences, nous utilisons l'encodeur SONAR correspondant lorsqu'il est disponible, et l'encodeur anglais sinon. Sur l'ensemble des paires de langues évaluées, notre modèle multi-tâches reste très proche de Att(sem), avec seulement de faibles différences en R@1, et surpasse systématiquement SONAR. Cela montre que l'ajout d'une supervision locuteur préserve, dans une large mesure, la capacité de recherche sémantique.

Les Tableaux 2 et 3 rapportent les scores R@1 pour la recherche de traduction *parole* → *texte* sur MTEDx et FLEURS, respectivement. Sur MTEDx, le modèle multi-tâches reste proche de Att(sem) pour la plupart des paires de langues et se maintient au-dessus de SONAR pour la majorité d'entre elles. Sur FLEURS, notre modèle reste proche de Att(sem) et surpasse systématiquement SONAR. De plus, le modèle surpasse légèrement Att(sem) sur la paire my-en (16,38 vs 14,11), ce qui suggère que la capacité de généralisation sémantique est maintenue même pour les langues peu dotées.

Recherche X Parole → EN Parole								
R@1 ↑	**fr-en**	**pl-en**	**nl-en**	**es-en**	**hr-en**	**de-en**	**ro-en**	**cs-en**
SONAR	91,91	95,79	95,16	95,30	52,16	94,18	94,45	95,62
Att(sem)	96,55	96,46	95,75	96,48	96,50	94,71	96,83	96,70
Att(sem+spk)	95,94	95,67	95,37	96,01	95,90	93,91	96,49	96,32
Recherche EN Parole → Y Parole								
R@1 ↑	**en-fr**	**en-pl**	**en-nl**	**en-es**	**en-hr**	**en-de**	**en-ro**	**en-cs**
SONAR	91,43	95,57	94,65	95,10	52,36	93,82	74,29	95,50
Att(sem)	96,54	96,25	95,71	96,37	96,31	94,12	97,16	97,09
Att(sem+spk)	95,96	95,75	95,39	95,96	95,79	93,46	96,58	96,49
Recherche X Parole → Y Parole								
R@1 ↑	**fr-de**	**hr-cs**	**ro-fr**	**hu-da**	**de-fr**	**cs-hr**	**fr-ro**	**da-hu**
SONAR	91,39	52,93	92,01	3,32	92,73	53,65	92,15	4,31
Att(sem)	95,20	94,75	96,55	92,69	95,36	94,69	96,90	92,63
Att(sem+spk)	93,83	94,01	96,18	91,07	93,72	93,91	96,49	90,79

TABLE 1 – Scores R@1 pour la recherche de traduction parole → parole pour diverses paires de langues (VoxPopuli)

	X Parole → EN Texte				X Parole → Y Texte				
R@1 ↑	**it-en**	**fr-en**	**pt-en**	**ru-en**	**it-es**	**es-it**	**es-fr**	**fr-pt**	**pt-es**
SONAR	89,01	82,45	85,05	84,76	92,25	88,57	87,76	84,27	87,08
Att(sem)	90,69	87,01	86,69	83,06	94,35	86,83	87,35	90,08	89,31
Att(sem+spk)	90,10	86,14	85,68	82,27	94,15	86,57	86,91	89,68	88,63

TABLE 2 – Scores R@1 pour la recherche de traduction parole → texte pour diverses paires de langues (MTEDx)

	X Parole → EN Texte					X Parole → Y Texte		
R@1 ↑	**ml-en**	**lb-en**	**uz-en**	**bs-en**	**my-en**	**ny-ces**	**sd-fr**	**xh-ar**
SONAR	22,36	39,98	54,12	27,27	8,19	18,18	21,20	16,07
Att(sem)	62,59	60,88	55,40	60,39	14,11	27,00	59,74	36,67
Att(sem+spk)	61,55	59,59	55,17	60,28	16,38	25,24	58,24	35,71

TABLE 3 – Scores R@1 pour la recherche de traduction parole → texte pour diverses paires de langues (FLEURS)

3.3.2 Vérification du locuteur

Le Tableau 4 présente les résultats de vérification du locuteur sur VoxCeleb1-O en termes d'EER (Equal Error Rate) et de fonction de coût minimale de détection (minDCF). Les résultats montrent que le modèle multi-tâches atteint un EER de 0,91%, très proche du modèle enseignant ECAPA-TDNN (0,90%), avec un écart également faible en minDCF. Cela indique que la branche locuteur parvient à reproduire les représentations de locuteur du modèle enseignant, alors même que l'encodeur partagé est simultanément optimisé pour l'alignement sémantique. De plus, l'optimisation conjointe peut même bénéficier aux représentations de locuteur, puisque Att(sem+spk) surpasse légèrement Att(spk).

Globalement, l'entraînement multi-tâches proposé préserve efficacement l'information discriminante pour le locuteur.

Modèle	**EER** ↓	**MinDCF**$_{0.01}$ ↓
ECAPA-TDNN	0,90	0,1104
Att(spk)	0,93	0,1285
Att(sem+spk)	0,91	0,1253

TABLE 4 – Résultats de vérification du locuteur sur VoxCeleb1-O

4 Analyse

Afin de mieux comprendre comment le modèle multi-tâches exploite l'information sémantique et l'information liée au locuteur au sein de l'encodeur de parole unifié et partagé, nous analysons les poids d'interpolation entre couches appris $\lambda_{\tau,\ell}$ (Eq. 2.2) pour les deux branches spécifiques aux tâches.

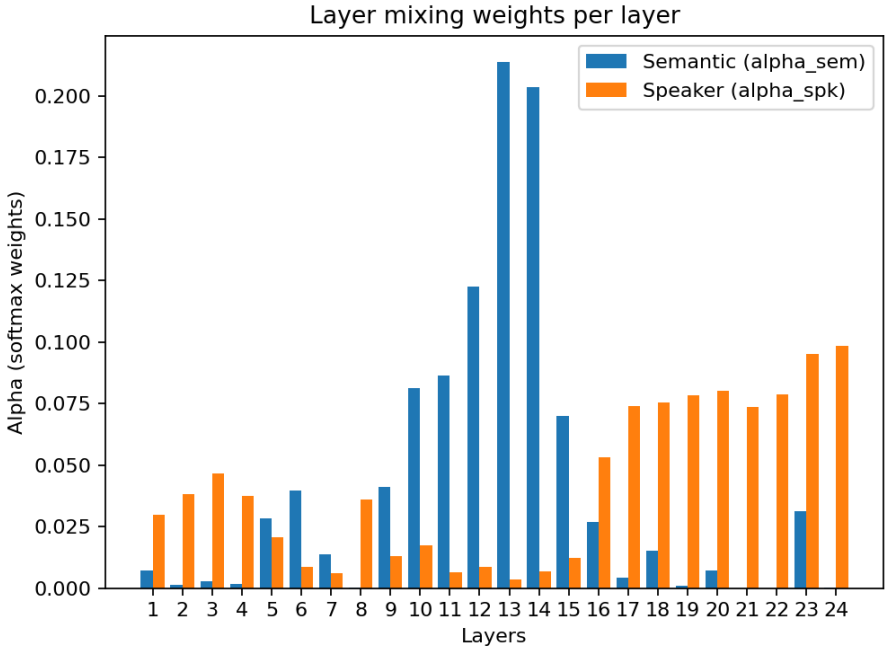

FIGURE 2 – Poids d'interpolation entre couches appris pour les branches sémantique et locuteur de l'encodeur de parole unifié.

La Figure 2 présente les poids de combinaison appris. Les deux branches montrent des profils de sélection de couches clairement différents. La branche sémantique concentre l'essentiel de sa masse de poids sur un intervalle restreint de couches intermédiaires, avec un pic marqué autour des couches 13 et 14, ce qui indique que la tâche sémantique s'appuie principalement sur une région localisée de l'encodeur. À l'inverse, la branche locuteur répartit ses poids de manière plus large sur l'ensemble de l'encodeur, avec une augmentation progressive vers les couches les plus hautes et un maximum aux couches 23 et 24, ce qui suggère que la tâche locuteur mobilise une portion plus étendue du réseau.

5 Conclusion

Dans ce travail, nous avons introduit un cadre unifié de post-entraînement permettant à un unique modèle de fondation pour la parole d'apprendre simultanément plusieurs représentations d'attributs au niveau de l'énoncé, au moyen de branches spécifiques aux tâches connectées à un encodeur partagé.

Nos expériences montrent que plusieurs attributs, tels que des représentations sémantiques et des représentations de locuteur, peuvent être appris conjointement sans dégradation majeure de l'une ou l'autre tâche : le modèle multi-tâches reste proche de la base sémantique mono-tâche, à la fois en recherche *parole* $\rightarrow$ *parole* et *parole* $\rightarrow$ *texte*, y compris pour des langues peu dotées, tandis que les performances en vérification du locuteur demeurent quasiment au niveau du modèle enseignant ECAPA-TDNN. L'analyse des poids d'interpolation entre couches appris révèle que chaque tâche sélectionne des couches différentes de l'encodeur selon un motif complémentaire, ce qui met en évidence la capacité du modèle à identifier automatiquement les couches les plus pertinentes pour chaque tâche. Afin de favoriser la reproductibilité et le partage scientifique, le code associé à cette approche sera rendu public. Dans des travaux futurs, nous prévoyons d'étendre ce cadre en intégrant des attributs supplémentaires, tels que l'émotion, la langue et l'accent, afin de construire des représentations de parole plus riches et plus polyvalentes à partir d'un encodeur de parole unifié.

Références

ARDILA R., BRANSON M., DAVIS K., KOHLER M., MEYER J., HENRETTY M., MORAIS R., SAUNDERS L., TYERS F. & WEBER G. (2020). Common Voice : A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, p. 4218–4222.

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). Wav2vec 2.0 : A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems (NeurIPS)*.

BARRAULT L., CHUNG Y.-A., MEGLIOLI M. C., DALE D., DONG N., DUPPENTHALER M., DUQUENNE P.-A., ELLIS B., ELSAHAR H., HAAHEIM J. *et al.* (2023). Seamless : Multilingual expressive and streaming speech translation. *arXiv preprint arXiv :2312.05187*.

CHEN J., XIAO S., ZHANG P., LUO K., LIAN D. & LIU Z. (2024). Bge m3-embedding : Multilingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv e-prints*, p. arXiv–2402.

CHUNG J. S., NAGRANI A. & ZISSERMAN A. (2018). Voxceleb2 : Deep speaker recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2018, p. 1086–1090.

CHUNG Y.-A., ZHANG Y., HAN W., CHIU C.-C., QIN J., PANG R. & WU Y. (2021). W2v-bert : Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, p. 244–250 : IEEE.

CONNEAU A., MA M., KHANUJA S., ZHANG Y., AXELROD V., DALMIA S., RIESA J., RIVERA C. & BAPNA A. (2023). Fleurs : Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, p. 798–805 : IEEE.

COSTA-JUSSÀ M. R., CROSS J., ÇELEBI O., ELBAYAD M., HEAFIELD K., HEFFERNAN K., KALBASSI E., LAM J., LICHT D., MAILLARD J. *et al.* (2022). No language left behind : Scaling human-centered machine translation. *arXiv preprint arXiv :2207.04672*.

DESPLANQUES B., THIENPONDT J. & DEMUYNCK K. (2020). Ecapa-tdnn : Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *Interspeech 2020*.

DUQUENNE P.-A., SCHWENK H. & SAGOT B. (2023). SONAR : sentence-level multimodal and language-agnostic representations.

FENG F., YANG Y., CER D., ARIVAZHAGAN N. & WANG W. (2022). Language-agnostic BERT sentence embedding. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1 : Long papers)*, p. 878–891.

HINTON G., VINYALS O. & DEAN J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv :1503.02531*.

HSU W.-N., BOLTE B., TSAI Y.-H. H., LAKHOTIA K., SALAKHUTDINOV R. & MOHAMED A. (2021). Hubert : Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

KHURANA S., LAURENT A. & GLASS J. (2022). SAMU-XLSR : Semantically-aligned multimodal utterance-level cross-lingual speech representation. *IEEE Journal of Selected Topics in Signal Processing*, **16**(6), 1493–1504.

MDHAFFAR S., ELLEUCH H., CHELLAF C., NGUYEN H. & ESTÈVE Y. (2025). SENSE models : an open source solution for multilingual and multimodal semantic-based tasks. *arXiv preprint arXiv :2509.12093*.

NAGRANI A., CHUNG J. S., XIE W. & ZISSERMAN A. (2020). Voxceleb : Large-scale speaker verification in the wild. *Computer Speech & Language*, **60**, 101027.

PARCOLLET T., NGUYEN H., EVAIN S., BOITO M. Z., PUPIER A., MDHAFFAR S., LE H., ALISAMIR S., TOMASHENKO N., DINARELLI M. *et al.* (2024). Lebenchmark 2.0 : A standardized, replicable and enhanced framework for self-supervised representations of french speech. *Computer Speech and Language*, **86**, 101622.

RAVANELLI M., PARCOLLET T., MOUMEN A., DE LANGEN S., SUBAKAN C., PLANTINGA P., WANG Y., MOUSAVI P., DELLA LIBERA L., PLOUJNIKOV A. *et al.* (2024). Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*, **25**(333), 1–11.

SALESKY E., WIESNER M., BREMERMAN J., CATTONI R., NEGRI M., TURCHI M., OARD D. W. & POST M. (2021). The Multilingual TEDx Corpus for Speech Recognition and Translation. In *Interspeech 2021*, p. 3655–3659. DOI : 10.21437/Interspeech.2021-11.

WANG C., RIVIÈRE M., LEE A., WU A., TALNIKAR C., HAZIZA D., WILLIAMSON M., PINO J. & DUPOUX E. (2021). Voxpopuli : A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *ACL*.

YANG S.-W., CHI P.-H., CHUANG Y.-S., LAI C.-I. J., LAKHOTIA K., LIN Y. Y., LIU A. T., SHI J., CHANG X., LIN G.-T. *et al.* (2021). Superb : Speech processing universal performance benchmark. *arXiv preprint arXiv :2105.01051*.