AAPG2020	Full3DTalkingHead		PRC
Coordonné par :	Yves Laprie	42 mois	620 k€
CE23			

Fully 3D talking head with aero-acoustic simulations

Tableau récapitulatif des personnes impliquées dans le projet

Partenaire	Nom	Prénom	Position actuelle	Rôle & responsabilités dans le projet (4 lignes max)	Implication sur la durée du projet (personne.mois)
LORIA/CNRS	*LAPRIE*	*Yves*	*Directeur de recherche*	*Coordinateur scientifique Responsable tâche 5*	21
LORIA/Université de Lorraine	*OUNI*	*SLIM*	*MCF*	*Responsable Tâche 3*	*14*
LORIA/Université de Lorraine	*COLOTTE*	*Vincent*	*MCF*	*Participant Tâche 3*	*10*
IADI/Université de Lorraine	*VUISSOZ*	*Pierre-André*	*Ingénieur de Recherche*	*Responsable scientifique Tâche 2*	21
IADI/INSERM	*ODILLE*	*Freddy*	*Chargé de Recherche*	*participant Tâche 1,2*	*10*
LPP/CNRS Université Paris 3	*DEMOLIN*	*Didier*	*Professeur*	*Responsable scientifique Tâche 1*	*18*
LPP/CNRS Université Paris 3	*AMELOT*	*Angélique*	*Ingénieur de recherche*	*Participant tâche 1*	*18*
LEGI/CNRS	*VAN HIRTUM*	*Annemie*	*Chargée de Recherche*	*Responsable scientifique Tâche 4*	*18*
LEGI/CNRS	*PELORSON*	*Xavier*	*Directeur de Recherche*	*participant Tâche 4*	*10*

Evolution(s) éventuelle(s) de la proposition détaillée par rapport à la pré-proposition
Le budget initial était estimé à 545 k€. Il est passé à 620 k€ (+14%) parce que nous avions sous-estimé les coûts d'enregistrement et surtout de traitement des corpus (IRM temps réel, capture du mouvement pour le visage, mesures de pression et physiologiques pour la source glottique et enfin mesures physiques avec la maquette). Le reste dont le contenu scientifique est inchangé.

I. Contexte, positionnement et objectifs de la proposition

1) Objectifs et hypothèses de recherche

L'objectif est de réaliser une tête parlante numérique tridimensionnelle complète comprenant le conduit vocal depuis les plis vocaux jusqu'aux lèvres, le visage, et intégrant la simulation des phénomènes aéro-acoustiques. En soi la parole est l'expression la plus directe de l'intelligence humaine et par nature la synthèse de la parole relève donc naturellement de l'Intelligence Artificielle. En retour, les développements récents de l'IA bouleversent en profondeur les approches qui avaient été proposées jusqu'à présent.

Pour ce qui nous concerne le défi est d'apprendre à contrôler la forme du conduit vocal et du visage lors de la production de la parole, d'intégrer la dimension des expressions, de traiter des volumes de données importants dont des images IRM temps réel du conduit vocal pour lesquelles il n'existe pas encore d'approche efficace, et enfin de lier étroitement les simulations numériques de l'acoustique et l'apprentissage automatique.

Le **premier verrou** est la possibilité d'observer les déformations géométriques bidimensionnelles ou tridimensionnelles du conduit vocal à une fréquence suffisante, c'est-à-dire au moins 50 Hz. Pour le visage les systèmes de capture du mouvement fournissent déjà une solution efficace sans affecter la production de la parole. Pour le conduit vocal l'articulographie électromagnétique (EMA) ne fournit que la position de quelques capteurs collés dans la cavité buccale et l'échographie n'offre, quant à elle, qu'une vue très partielle de la langue. Le premier pilier de ce projet est donc l'arrivée réellement

opérationnelle de l'IRM temps réel (à 50 Hz) qui permet d'acquérir de très grandes bases de données sur l'évolution temporelle du conduit vocal. Au-delà du simple progrès technologique, l'IRM temps réel bouleverse la manière d'envisager la modélisation temporelle de la géométrie du conduit vocal (voir ici quelques exemples illustratifs accompagnés de résultats préliminaires de suivi). Mais les données d'IRM temps réel ne correspondent qu'à un plan du conduit vocal, le plus souvent le plan médiosagittal parce qu'il offre une vue sur tous les articulateurs de la parole. L'objectif est donc d'obtenir une information dans la direction transverse à l'axe du conduit vocal pour compléter la géométrie bidimensionnelle.

Le **second verrou** est la possibilité d'apprendre le contrôle temporel du conduit vocal et du visage. Il s'agit d'un double verrou. D'abord la possibilité de travailler sur les articulateurs de la parole en tant que tels ce qui nécessite le suivi du contour des articulateurs dans les images IRM, ensuite l'apprentissage du contrôle lui-même.

Le second pilier de ce projet est donc l'utilisation de techniques d'apprentissage profond qui seront utilisés dans les deux cas. Pour le suivi des articulateurs l'apprentissage profond peut être supervisé quand il existe un détourage manuel comme base d'entraînement, ou faiblement supervisé quand l'information fournie dans le plan médio-sagittal peut servir à contraindre l'apprentissage.

Pour le contrôle temporel de la tête parlante l'objectif est d'apprendre le passage d'une suite de phonèmes complétée par une information suprasegmentale – par exemple les expressions – à l'évolution temporelle de la position des articulateurs et des facteurs de déformation du visage. L'objectif est d'apprendre cette transformation en dissociant le contrôle de chacun des articulateurs du conduit vocal ou des modes de déformation du visage, tout en conservant la cohérence de l'ensemble. Ce point est important pour la prise en compte des phénomènes de compensation à la suite d'une pathologie qui n'affecte qu'un seul articulateur, ou encore de la réalisation d'une expression.

Le **troisième verrou** est la coordination entre les différentes contributions à la production de la parole, à savoir la source de pression – et donc la respiration –, l'ouverture à la glotte, la vibration des plis vocaux et les gestes du conduit vocal et du visage. Les points mal maîtrisés restent la précision que requiert cette coordination, l'apprentissage des commandes et les simplifications géométriques qui n'affectent pas le résultat des simulations mais réduisent la complexité des calculs. Le troisième pilier sur lequel ces travaux s'appuieront est l'avancée des simulations aéro-acoustiques pour lesquelles nous avons accompli des progrès importants du point de vue du réalisme de la source voisée, des gestes d'ouverture et de fermeture de la glotte, et de celui des simulations acoustiques impliquant les différentes cavités du conduit vocal. L'accès à des données couvrant simultanément toutes les facettes de la production de la parole est, et restera, difficile simplement parce qu'il n'est pas possible d'instrumenter un locuteur avec des technologies d'acquisition limitant souvent les gestes du locuteur et parfois incompatibles entre elles pour des raisons physiques. La piste que nous suivrons est de combiner des techniques d'apprentissage avec des simulations physiques et numériques. Les simulations physiques à l'aide de maquettes fourniront une vérité terrain et permettront d'évaluer le bien-fondé et l'efficacité de techniques d'apprentissage associées à des simplifications de l'architecture globale, notamment en termes de complexité géométrique pour le conduit et le visage.

2) Positionnement par rapport à l'état de l'art et caractère novateur

Jusqu'à présent la construction d'une tête parlante se fait en fusionnant plusieurs facettes quasi indépendantes : (i) le signal acoustique tel que produit par un système de synthèse maintenant réalisé par apprentissage profond à partir d'un corpus audio de très grande taille, (ii) le visage réalisé lui aussi par apprentissage à partir d'un corpus de positions de capteurs collés sur le visage. Bien que la parole des systèmes récents soit presque indiscernable de la parole naturelle, ces approches sont des

modélisations superficielles qui ne peuvent pas lier les gestes du conduit vocal et du visage au signal de parole . manque la chute

a) Positionnement par rapport à l'IA

Le contrôle de la forme du conduit vocal et du visage est un domaine complexe car il s'agit de générer une suite de formes du conduit vocal qui assure la bonne acoustique en intégrant les contraintes anatomiques et physiologiques, mais aussi l'expression que veut donner le locuteur. Cela nécessite un apprentissage de plusieurs années chez l'humain et aucun modèle théorique ou pratique n'apporte de solution satisfaisante, au point que de nombreuses approches font appel à des modèles anciens, particulièrement celui de (Öhman 1966) pour la coarticulation.

Un courant de recherche très actif porte sur le contrôle moteur souvent sous l'angle des stratégies de compensation d'une perturbation (Guenther 2016) en altérant par exemple la perception de la parole que produit le locuteur, mais le modèle de production ne couvre généralement qu'une partie de la production (les voyelles et certaines suites consonnes voyelles).

Notre originalité est d'aborder l'apprentissage du contrôle avec une double approche consistant à apprendre le contrôle complet du conduit vocal et du visage dans le cadre de la parole continue naturelle, et d'autre part à apprendre le contrôle des simulations acoustiques afin de pouvoir explorer des situations de production éloignées de la parole standard ou encore l'interaction entre l'ouverture à la glotte et le reste du conduit vocal.

La première approche concerne l'apprentissage du passage d'une suite de phonèmes à articuler à l'évolution temporelle de la forme géométrique du conduit vocal et du visage en utilisant un corpus de données d'IRM temps réel et de données de capture du mouvement. L'originalité réside dans l'apprentissage direct du contrôle du conduit vocal à part des phonèmes à articuler sans passer par l'intermédiaire d'un modèle articulatoire qui approche la forme du conduit vocal à l'aide d'un petit nombre de modes de déformation. Cet apprentissage repose sur l'exploitation d'un vaste corpus d'IRM temps réel du français, mais cela pourrait être une autre langue. Il est possible parce que nous pouvons suivre automatiquement les contours des articulateurs, de la langue en particulier, dans un très grand nombre d'images et avec une excellente robustesse. Ce verrou a été levé très récemment fin 2019 (Isaieva et al. (2020), voir ici un exemple de suivi) pour la langue et nous allons maintenant étendre cette approche à l'ensemble du conduit vocal. Nous aborderons la question de la cohérence géométrique et temporelle des différents suivis pour éviter de récupérer des formes ou des gestes non réalistes. Le travail portera sur l'amélioration de la stratégie d'apprentissage (type de modèle, apprentissage par articulateur ou apprentissage global, régularisation...) de manière à accroître la robustesse à la variabilité intra et interlocuteur et aux situations de non parole mais intéressantes pour d'autres raisons (déglutition ou encore passage de la langue sur les lèvres).

La seconde approche consiste à apprendre le pilotage des outils de simulation acoustique. Jusqu'à présent le choix de nombreux paramètres de la source sonore, de sa coordination avec le reste du conduit vocal, du calcul de l'aire transverse à partir de la forme du conduit dans le plan médiosagittal et de la dynamique de la forme du conduit vocal relevait de plus du savoir faire que d'une approche systématique. Il y avait deux raisons. La première était que l'on n'avait pas de garantie que la forme du conduit vocal était correcte ce qui obligeait à la considérer comme une inconnue (qui plus est une inconnue à un grand nombre de degrés de liberté) et qu'il n'y avait pas de signal de parole représentant la vérité terrain. Ces deux limites fortes tombent en grande partie avec l'IRM temps réel débruitée (cf exemples) qui fournit la forme bidimensionnelle au cours du temps et un signal de qualité correcte, et complètement dans le cas de la maquette que le LEGI va développer. Il devient possible de recourir à une approche inspirée par exemple de l'architecture Reprise (REtrospective and Prospective Inference Scheme) (Butz et al. 2019) qui fait appel à des RNN pour intégrer dans le temps la dimension sensorimotrice.

Le positionnement de notre projets vis-à-vis de l'IA est donc le suivant.

1. Nous utiliserons l'IA comme outil pour le suivi du contour des articulateurs en visant à en améliorer la robustesse. Il faut souligner que cette question est restée jusqu'à 2019 un défi scientifique insurmontable, même si les modèles de forme actifs avaient commencé à apporter une réponse correcte (par exemple Labrunie et al. 2018).
2. L'apprentissage profond sera le cadre conceptuel pour le contrôle du conduit vocal et la découverte des variables latentes plus proches du comportement humain réel que ne le sont les paramètres d'un modèle articulatoire géométrique.
3. Nous explorerons la possibilité d'apprendre le contrôle direct des simulations acoustiques, soit dans le cadre d'un modèle simplifié d'une maquette physique afin de vérifier la pertinence de l'apprentissage, soit pour ce qui concerne l'interaction entre conduit vocal et plis vocaux. Dans les deux cas cela nous permettra de sortir du cadre restreint correspondant aux données présentes dans les corpus.
4. L'apprentissage du contrôle du conduit vocal, du visage et des plis vocaux nous donnera un point de vue nouveau pour aborder la question de l'existence d'un répertoire de gestes articulatoires et la mise en œuvre de stratégies de compensation.

b) Positionnement par rapport à la production et l'acoustique de la parole

Géométrie du conduit vocal

Dans le cadre de l'ANR ArtSpeech nous avions essentiellement travaillé sur la géométrie du conduit vocal par l'intermédiaire d'un modèle articulatoire construit à partir d'images IRM, les écoulements turbulents, le renforcement du réalisme de la modélisation des plis vocaux en prenant en compte leur humidification, un schéma numérique de simulations acoustiques numériques prenant en compte des topologies complexes (cavités en parallèle pour les latérales, vibration de la langue pour des trilles par exemple) et enfin le rôle de l'ouverture à la glotte.

Notre position pour ce projet est très profondément différente. Traditionnellement la modélisation aéro-acoustique et du contrôle moteur ne couvre que le conduit vocal et les plis vocaux. L'interaction entre la respiration, la pression sous-glottique et le contrôle des plis vocaux reste donc mal connue alors qu'il s'agit point d'un particulièrement important pour la prise en compte des expressions — parole plus visage. Pour cette raison nous proposons de mesurer les paramètres aéro-acoustiques profonds et des données directement liées aux expressions ce qui nécessite des EEG.

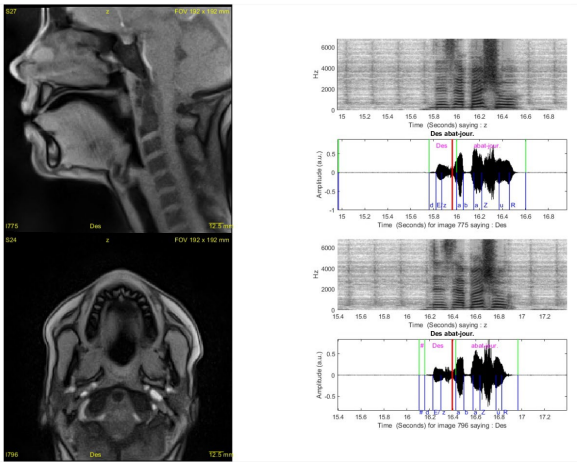

Fig 1: IRM temps réel sagittale et axiale synchronisée

Par ailleurs pour ce qui concerne la modélisation géométrique du conduit vocal nous considérons qu'il n'est plus nécessaire d'utiliser un modèle articulatoire construit à l'aide de techniques d'analyse de données (souvent dérivées de l'analyse en composantes principales) à partir d'un petit nombre d'images. Ces modèles se justifiaient parce qu'il était impossible d'acquérir un grand nombre d'images du conduit vocal et ensuite d'en extraire automatiquement les contours avec une fiabilité suffisante. Leur point faible est qu'une forme de conduit vocal peut être approchée par plusieurs vecteurs en tenant compte des compensations qui existent et que cela rend difficile la modélisation de la coarticulation. Dans notre projet l'apprentissage de la coarticulation sera apprise directement à partir des contours des articulateurs et des données du visage.

Contrôle de la source de pression

Le système phonatoire est étroitement lié au système respiratoire qui contrôle la source de pression à l'origine de la vibration des plis vocaux. Dans les simulations actuelles le contrôle de la pression est réalisé de manière tout à fait arbitraire en fixant la pression à une valeur réaliste mais qui ne repose sur aucune mesure. Nous nous démarquerons des approches actuelles en mesurant et en enregistrant la pression sous-glottique (à l'aide d'un pléthysmographe qui est un dispositif non-invasif pour évaluer la pression sous-glottique) et les données physiologiques du système respiratoire (à l'aide d'EMG et d'accéléromètres 3D) pour le même corpus de phrases utilisé pour l'IRM dynamique et le visage. Ces données permettront d'explorer et d'apprendre les mécanismes physiologiques qui déclenchent une phase aérodynamique et la production de la parole.
La deuxième phase concernera le niveau supérieur à la respiration en cherchant à comprendre les corrélations entre les mécanismes physiologiques, les processus aérodynamiques et l'acoustique des sons produits afin d'expliciter les phénomènes qui sont de l'ordre de l'automatisme et ceux qui sont contrôlés et produits intentionnellement. Il s'agira ici de d'utiliser des instrumentations permettant de faire le lien entre données physiologiques, aérodynamiques et acoustiques, avec les instruments suivants : EGG, ePGG, mesures aérodynamiques (pression et débit) et acoustiques.

Apprentissage du contrôle des actuateurs du conduit vocal

Du point de vue de la physique, la production de parole est un processus faisant intervenir des sources de son de natures différentes, leurs interactions ainsi que la propagation et le rayonnement d'ondes sonores dans des géométries complexes.
Notre approche repose sur la modélisation physique, et consiste à développer des modèles théoriques qui sont validés sur des maquettes de la totalité ou d'une partie du conduit vocal. Des simulation numériques peuvent également compléter les données expérimentales. Cette méthodologie est courante dans le cas de la source glottique mais reste très originale dans le cas des consonnes.
Un travail récent sur les séquences voyelle-plosive-voyelle (Delebecque L et al 2016) a été ainsi réalisé. Les résultats permettent de rendre compte de l'interaction entre la sources de plosive et la source glottique et de prédire l'extinction (offset) et l'apparition (onset) du voisement. Appliqués à la synthèse articulatoire de la parole, ces résultats sont particulièrement intéressants car ils permettent de s'affranchir d'un grand nombre de paramètres de commande ad-hoc et souvent irréalistes (contrôle de la glotte pendant la plosive, par exemple). Ces travaux et conclusions sont limités aux plosives bilabiales, l'extension à d'autre types de plosives et aux plosives voisées reste à réaliser.
En ce qui concerne les fricatives, si la compréhension des mécanismes de production a grandement progressé depuis peu, l'apparition et l'extinction de ces sons dans un contexte articulatoire reste à étudier. Il en est de même de l'interaction avec d'autres sources sonores (fricatives voisées).

Les observations expérimentales sont réalisées sur des maquettes du larynx et du conduit vocal de deux manières complémentaires. L'une consiste à considérer un conduit vocal géométriquement très simplifié mais dont la variation dans le temps est contrôlée, tel que le prototype décrit dans (Van Hirtum A et al 2016). La déformation imposée reproduit les gestes articulatoires et permet de générer spontanément les différentes sources de sons de parole. C'est déjà le cas de fricatives (Van Hirtum A et al 2011) et potentiellement de plosives. En conséquence la mesure simultanée des quantités géométriques, acoustiques et aérodynamiques permet la validation de modèles physiques de la production de ces sons, de leur apparition (onset) et de leur extinction (offset) ainsi que de leurs interactions notamment avec la source glottique.
Ces données expérimentales, uniques au monde, sont d'une grande utilité pour l'approche IA proposée. Dans un premier temps, une base de données mesurées permet de contribuer au développement des stratégies par apprentissage profond. Dans une second temps, l'IA sera utilisée pour prédire les commandes de la maquette et pourra donc être validée par comparaison avec les géométries, les sources sonores et les variables d'écoulement effectivement imposées sur la maquette. La question de la stabilité des commandes est essentielle ; des techniques proches de la

régularisation utilisée notamment par les auto-encodeurs variationnels (Kingma & Welling 2014) seront étudiées afin d'apporter une réponse à cette question.

L'autre approche utilise des géométries détaillées ("réalistes") du conduit vocal obtenues par impression 3-D de géométries obtenues par IRM. Ce type d'approche a été réalisée dans le cas de voyelles, mais seules quelques fricatives ont pu être étudiées jusqu'à présent. Les données aérodynamiques sur des géométries réalistes de fricatives voisées, ou non, sont rares et d'une très grande importance.
La simulation acoustique multimodale a été développée et validée dans le cas des voyelles (Blandin R et al 2018). Cette approche permet de prédire avec précision l'acoustique de conduits vocaux réalistes avec un faible coût de calcul. Il est donc intéressant d'utiliser cette approche à l'articulation et à la descriptions des différentes sources sonores, de manière systématique. Récemment, il a été montré que les sources de sons de fricatives pouvaient être représentées par des sources dipolaires (Yoshinaga T et al 2020). Il est donc prévu de développer plus avant l'approche multimodale afin d'améliorer la description des sources de fricatives ainsi que prendre en compte l'interaction avec d'autres sources.
Le modèle aéroacoustique ainsi développé pourra être validé par confrontation aux mesures sur les maquettes évoquées ci dessus et sera implémenté dans le synthétiseur.
La simulation numérique, particulièrement dans le cas des fricatives, est un outil intéressant car elle permet d'accéder à l'ensemble des caractéristiques de l'écoulement et du champ acoustique (Pont A et al 2019; Cisonni J et al 2013; Cisonni et al 2011). Les coût élevés en temps de calculs nécessitent cependant de se restreindre à une sélection réduite de géométries. Les simulations numériques peuvent être partiellement validées par les mesures sur maquettes et les prédictions de la théorie multimodale.

c) Positionnement par rapport à l'imagerie IRM du conduit vocal

L'imagerie temps réel IRM (rtMRI) possède un avantage par rapport aux autres techniques (rayons X, ultrasons, électromagnétographie …) car elle produit une vue du conduit vocal dans son ensemble en incluant les structures pharyngales d'une manière sûre et non invasive. Avec la rtMRI une série dynamiques d'images peut être acquise dans n'importe quel plan anatomique, en particulier le plan médio-sagittal du conduit vocal de la glotte aux lèvres. Dans ces images on peut tracer l'interface air tissu des articulateurs d'intérêt pour la recherche sur la production vocale et ainsi obtenir un modèle du conduit vocal.

L'imagerie IRM étant peut sensible, les protocoles d'acquisition avec cette technique nécessite toujours une optimisation entre le rapport signal sur bruit (SNR) et la couverture de la zone d'intérêt (2D/3D) et la résolution spatiale et temporelle. L'IRM est utilisée depuis les années 90 dans l'étude du conduit vocal (Demolin D. et al 1996), mais un regain d'intérêt a été provoqué ces dernières années (Bresch E. et al 2008), (Scott A. D. et al 2014) , (Lingala S. G. et al 2016) avec la dissémination de techniques multi antennes de reconstruction itérative (Odille F. et al 2008) et non cartésiennes nécessitant d'importantes puissances de calcul (Uecker M. et al 2010).

Les techniques d'acquisition actuelles font encore l'objet de recherches et c'est dans ce cadre que le système IRM 3T du CHRU de Nancy a été équipé d'une séquence rtMRI (licence du Max Planck Institut) permettant l'acquisition de séries dynamiques d'images à 50 images par secondes (Uecker M. et al 2010). Les techniques de séparation de sources développées au Loria permettent en outre de disposer d'enregistrements sonores synchronisés avec suppression du bruit IRM (le bruit peut être de 120 db pendant l'acquisition) (Ozerov A. et al 2012). La quantité de données produites par un tel système nécessite d'automatiser le post-traitement, en particulier la segmentation.

L'IRM permet l'acquisition de séries dynamiques en coupe épaisse (8 mm) pour garantir un SNR suffisant. Les nouveaux développements de reconstruction IRM utilisant la super résolution (Bustin A.

et al 2018) permettent d'envisager d'associer des acquisitions réalisées dans différentes orientation pour produire une acquisition 3D ciné de la tête complète (Zhu Y et al 2013).

3) Méthodologie et gestion des risques

a) Programme et structuration générale du projet

Le projet est organisé en 5 tâches scientifiques et techniques plus une tâche de coordination (tâche 0). Les tâches ont été conçues de manière à isoler les risques et éviter qu'un retard se répercute sur l'ensemble du projet.

Le projet pourra démarrer dans de très bonnes conditions puisque la plupart des équipements sont disponibles (dont le système d'acquisition d'IRM temps réel notamment). Les partenaires pourront disposer des données acquises dans le cadre d'autres projets pour commencer leurs travaux sans attendre la fin de l'acquisition du corpus. Il s'agit en particulier de données de capture du mouvement du visage (en lien avec la parole et l'expressivité), du corpus IRM (statique et dynamique) acquis à la fin du projet ArtSpeech, de données de la pression sous-glottique pour plusieurs locuteurs, d'une maquette physique d'un conduit vocal très simplifié (voir vidéo ici) et des simulations numériques de l'acoustique du conduit vocal.

b) Description des tâches et calendrier

Tâche 1 - Constitution du corpus – Conduit vocal – Visage – Plis vocaux - Resp. LPP

Objectif
L'objectif est de construire et collecter un corpus associant l'IRM temps réel (complétée par un petit nombre d'IRM statiques de meilleure résolution), le suivi de capteurs collés sur le visage à l'aide d'un système Optitrack, et enfin la pléthismographie et l'électro-photo-glottographie (ePGG) pour la pression sous-glottique et l'ouverture glottique. Le but est d'acquérir un corpus de 3 heures pour une locutrice et un locuteur.
Défis
Le défi est de collecter plusieurs modalités sur plusieurs sessions d'enregistrement en sachant que la qualité des données dépend souvent des locuteurs, de les annoter et les utiliser conjointement ce qui impose notamment de définir un même référentiel spatial et temporel.
Livrables
Corpus IRM statiques (environ 80 volumes statiques) et dynamiques (3 heures dans le plan médiosagittal pour 3000 phrases avec une expression standard et 500 phrases pour les expressions de la joie, la colère et la tristesse) et une ou deux heures pour la meilleure locutrice et le meilleur locuteur dans d'autres orientations pour récupérer la dynamique 3D. Corpus de données du visage sous la forme de nuages de points capturés à l'aide d'un dispositif de capture du mouvement (Optitrack) pour le même ensemble de phrases et les deux meilleurs locuteurs. Corpus de données aéroacoustiques intégrant la pression sous-glottique et l'ouverture à la glotte pour les 2 meilleurs locuteurs.
Programme détaillé
T1.1 Outre la conception du corpus à enregistrer, le travail consistera à assurer la répétabilité des acquisitions IRM (en contrôlant la posture de la tête dans l'antenne IRM), la segmentation phonétique du corpus et l'alignement des données du conduit vocal et du visage puisque les deux acquisitions ne peuvent avoir lieu simultanément. De la même façon, le dispositif d'ePGG sera amélioré pour faciliter son utilisation et généraliser sont utilisation à un plus grand nombre de sujets.

T1.2 Compte tenu de la difficulté de trouver une locutrice et un locuteur pour lesquels toutes les modalités donnent des données de très bonne qualité, nous commencerons par sélectionner rigoureusement 4 locuteurs (2 locutrices et 2 locuteurs) compatibles avec l'IRM parce que c'est la modalité la plus exigeante en termes du choix des sujets. Pour les deux meilleurs sujets nous acquérons des IRM dynamiques dans d'autres plans pour réaliser la reconstruction tridimensionnelle et l'ensemble des données pour les autres modalités.

Pour l'IRM cela représente 5 séances d'IRM d'une heure et demie par locuteur, plus 2 ou 3 pour les deux meilleurs afin de réaliser des IRM 2D dynamiques dans différentes orientations afin de contraindre le modèle 3D complet, soit au total 30 séances d'IRM.

T1.3 : Les acquisitions de données sur le visage se dérouleront au Loria et utiliseront le même corpus que le corpus d'IRM dynamique.

T1.4 : Les acquisitions de données aéroacoustiques se dérouleront au LPP à Paris. Elles porteront sur un sous corpus parce qu'elles sont plus contraignantes pour les sujets, même si elles ne présentent aucun risque. Il s'agira d'enregistrer des données de pression sous-glottique avec un plethymographe et de coupler ces enregistrements avec un ePGG pour connaître l'évolution de l'ouverture glottique durant la production de parole.

Choix technologiques

IRM statique 3D, dynamique 2Dl (M. Uecker et al. (2008))

Système de capture du mouvement Optitrack.

Pléthismographe, électrophotoglottographie (Amelot et al. (2018))

Pour les trois dispositifs nous sommes capables d'enregistrer le son produit par le sujet ce qui permettra la création d'un référentiel temporel. Une IRM 3D statique haute résolution, sera réalisée pour chaque session et toutes les images seront liées à ce référentiel spatial.

Contributions (qui fait quoi)

Le **IADI** assurera l'acquisition des IRM, le laboratoire a accès à l'IRM 3T Siemens Prisma du CHRU de Nancy dédiée à 40% à la recherche, cet IRM est équipé de la séquence de recherche 2D RT-MRI, un protocole générique de recherche clinique sur les acquisitions du conduit vocal est développé dès cet été 2020 en collaboration avec le **CIC-IT de Nancy** et le **Loria** les sujets seront inclus dans le cadre de ce protocole.

Le **LPP** assurera les acquisitions aéroacoustiques, le LPP est équipé d'un système de pléthymographie et d'un électrophotoglottographe. Une partie de ce projet sera aussi l'occasion d'améliorer l'électrophoto-glottographie et de permette d'acquérir des données sur un échantillon moins restreint de locuteurs.

Le **Loria** assurera l'acquisition des données dynamiques du visage grâce au système de capture de mouvement de la plateforme MultiMod du Loria.

Risques

Peu de risque à ce niveau mais nous devrons faire une sélection rigoureuse des locuteurs en vérifiant qu'ils ne présentent pas d'incompatibilité avec les différentes modalités d'acquisition.

Alternatives

Nous disposons déjà d'une base de données IRM 2D dynamique et 3D statique sur 10 volontaires même si la durée enregistrée est moindre, des captures de mouvement du visage pour un autre locuteur, et enfin des données électrophotoglottographe pour d'autres sujets. Ces données permettront de commencer la tâche 2 avant la finalisation de la tâche 1.

Tâche 2 - Prétraitement des données IRM, visage, glotte et passage 2D-3D - resp. IADI

Objectif

AAPG2020	**Full3DTalkingHead**		PRC
Coordonné par :	Yves Laprie	42 mois	620 k€
CE23			

L'objectif est d'homogénéiser un ensemble de données multimodales acquises par la tâche T1, de développer des outils et algorithmes permettant un apprentissage automatisé d'un modèle de locuteur personnalisé à partir de l'ensemble des données multimodales pour fournir à la tâche T3 un modèle de tête 3D incluant le conduit vocal. Les séries dynamiques en coupe (en général dans le plan médio-sagittal) acquises à l'aide de l'IRM temps réel constituent les informations essentielles qui permettront le suivi du contour des articulateurs.

La première partie du travail consiste à développer les algorithmes de suivi automatique pour chacune des modalités (rtIRM, optitrack ...). Dans une seconde partie les outils permettant l'obtention d'une géométrie tridimensionnelle dynamique du conduit vocal seront développés. Cette extension vise le développement d'un outil automatique.

Il est possible d'acquérir des données dynamiques simultanément dans plusieurs plans en abaissant la fréquence d'acquisition. L'idée est d'utiliser ces données en supplément des acquisitions statiques 3D, pour calculer le champ de déformation et reconstruire la géométrie 3D dynamique.

Défis

Segmentation de plus de 540k images (87GB dicom + 2GB wav) par locuteur ~~non~~ supervisée des différents articulateurs, extension de la dynamique 2D à la 3D à partir d'un nombre de contraintes restreintes (petit nombre d'acquisition 2D perpendiculaires et 3D statique)

Indicateur de succès

Corpus prêt à l'utilisation pour l'apprentissage du modèle de contrôle (données annotées et recalées entre elles, segmentations indépendantes pour chaque articulateur pour tout le corpus...).

Livrables

Publication décrivant les outils de segmentations automatique robuste et leur application sur le corpus.

Publication décrivant les outils d'extension de la dynamique 2D à un modèle dynamique 3D.

Programme détaillé

T2.1 Constitution de la base de données contenant les différentes modalités, débruitage des enregistrements sonores sous IRM, Annotation et synchronisation des différentes modalités.

T2.2 Développement de réseaux auto-apprenants (par exemple : antagonistes génératifs) pour la segmentation non supervisée des articulateurs.

T2.3 Développement de système de reconstruction super-résolution permettant d'étendre à partir des données dynamique 2D vers un modèle dynamique 3D.

Choix technologiques

Bien qu'intrinsèquement instables les GAN (Réseaux antagonistes génératifs) (M. Eslami et al. 2020), (Joyce T. et al. 2018) peuvent être appliqués à la segmentation. La segmentation du conduit vocal a déjà été réalisée à l'aide de cette technique et la segmentation non supervisée semble réalisable. La clef sera de déterminer la fonction de coût permettant d'étendre l'apprentissage par un transfert à partir de la petite base de segmentation déjà existante.

Nous avons des paires 2D, 3D pour une quantité de phonèmes ce qui va nous permettre une approche Image to Volume Translation. Nous envisageons l'utilisation de la constitution d'un atlas pour étendre la dynamique 2D avec une approche Image to Volume Translationavec cGAN (Isola P. et al. 2018), (Kwon G. et al 2019).

Contributions (qui fait quoi)

L'expertise en apprentissage profond du **Loria** sera étroitement associée à l' expertise sur l'IRM des organes en mouvement du **IADI.** L'unité de lieu facilitera la définition des spécifications nécessaire pour la réalisation de la tache T3. Le **LPP** contribuera à la constitution de la base de données globales.
Risques
Nécessiter de réaliser beaucoup plus de segmentation que prévues et d'utiliser un apprentissage supervisé.
Alternatives
Apprentissage entièrement supervisé. Le passage 2D à 3D peut être réalisé par superposition d'acquisition 2D synchronisée (Zhu Y et al 2013)

Tâche 3 - Apprentissage du modèle de génération - conduit vocal + visage + source. resp. Loria

Objectif
L'objectif est de contrôler l'évolution temporelle de la forme du conduit vocal, du visage et de l'ouverture à la glotte à partir de la suite des phonèmes à articuler et d'informations supra-segmentales. L'approche reposera sur un apprentissage profond. Une approche similaire a été utilisée avec succès pour la prédiction du visage (Biasutto–Lervat et al. 2019), ce qui est un résultat important et relativement inattendu. Nous souhaitons étendre cette approche à toute la tête (conduit vocal plus visage) afin de produire la géométrie instantanée complète de la tête. Le second objectif est de faire émerger les variables latentes permettant de contrôler un articulateur, les lèvres par exemple, aussi indépendamment que possible des autres articulateurs. Il s'agit d'un point essentiel dans la perspective d'étudier les répercutions acoustiques du dysfonctionnement de l'un des articulateurs, ou de la réalisation d'expressions
Défis
L'hétérogénéité des données et leur quantité peuvent avoir un impact sur la technique de deep learning qui sera considérée.
Livrables
Un modèle de contrôle d'une tête parlante animée complète et finement contrôlable qui permet d'observer un mécanisme de parole réaliste.
Programme détaillé
T3.1 Modèle de contrôle dynamique de la tête parlante - Des techniques d'apprentissage profond seront utilisées pour estimer les déformations du conduit vocal et du visage à partir des informations phonétiques. Une modélisation à base d'auto-encodeur variationnel (ou des techniques similaires, si elles facilitent l'interprétabilité) sera utilisée pour contrôler conjointement le conduit vocal et la dynamique du visage. **T3.2 Coordination avec la source vocale dont la pression sous-glottique** - Un apprentissage similaire sera réalisé à partir des données de pression sous-glottique, de la segmentation phonétique et des données de l'expressions. Les informations du contrôle du conduit et du visage pourront êtes utilisées dans le cadre d'un processus d'apprentissage par transfert (Jia et al. 2018). Cela consiste à injecter un espace latent pré-calculé dans le réseau, de sorte que l'architecture neurale dispose déjà d'une représentation interne des

données articulatoires pour initier l'apprentissage. En lien avec la tâche 4.3 un apprentissage plus fin temporellement sera réalisé spécifiquement pour les occlusives et fricatives en intégrant les simulations acoustiques numériques atteindre le niveau de précision temporel suffisant.

T3.3 Interprétabilité du modèle - Afin de faire émerger les variables latentes permettant de contrôler un articulateur, nous allons explorer des techniques d'interprétation des réseaux neuronaux profonds (*Bach et al.*, 2015 ; Zeiler & Fergus, 2014). Ces outils seront utilisés pour trouver les meilleures explications liées au domaine de la production de la parole.

Choix technologiques

Plusieurs techniques de l'apprentissage profond seront utilisées.

Contributions (qui fait quoi)

Cette tâche sera réalisée principalement par le **Loria** avec des contributions du **LEGI** et du **LPP.**

Risques

Comme la résolution temporelle de chaque modalité n'est pas la même, une architecture à base d'auto-encodeurs peut ne pas être en mesure de capturer finement la spécificité de l'ensemble des informations multimodales.

Alternatives

Le contrôle du conduit vocal et de la dynamique du visage sera fait indépendamment l'un de l'autre en utilisant DNN, et nous étudierons comment les fusionner pour contrôler l'animation d'une tête parlante complète. Cette fusion peut profiter également des techniques d'apprentissage par transfert (*transfer learning*).

Tâche 4 - Mesures et simulations aéro-acoustiques — **resp. LEGI**

Objectif

Le premier objectif de cette tâche est de développer banc expérimental qui permettra de générer un grand nombre de mesures précises des paramètres aérodynamiques et de la variation de géométrie qui pourront servir de test de validation des stratégies développées dans la tâche 3. Le banc expérimental inclue une source auto-oscillante couplée à un résonateur acoustique dont la géométrie est contrôlée mécaniquement.
Ce banc expérimental permettra en outre de valider des modèles physiques de la production de la parole, au niveau des sources (sons voisés, plosives, fricatives) et de leurs interactions.
Enfin, une sélection de géométries réalistes, issues de la tâche 2, sera utilisée pour faire des mesures aéro-acoustiques sur des maquettes obtenues par impression 3-D ainsi que des simulations numériques d'écoulement.

Défis

Construire un banc expérimental unique au monde permettant de générer de manière contrôlée et précise des déformations comparables en amplitude et en durée à celles observées dans le conduit vocal lors de la production de parole.

Indicateur de succès

Les performances mécaniques et aérodynamiques de la maquette seront comparées aux données de la littérature. $$ couvrir l'ensemble des réalisations dynamiques possibles par un locuteur, voire au delà.

Livrables

Un banc expérimental complet, fonctionnel et calibré.

Série de mesures complètes pour tests en lien avec la Tâche 3.

Programme détaillé

T4.1 Un prototype fonctionnel présent au LEGI constitué d'un tube en silicone peut être déformé en différents endroits par des actuateurs. Couplé à une source (maquette de cordes vocales) il est donc possible de mesurer une grande quantité de réalisations acoustiques issues de déformations dont les ordres de grandeur (variation d'aire, vitesse de déformation) sont comparables à celles observées chez un locuteur. Ce dispositif sera amélioré pour prendre en compte des déformations plus rapides comme celles rencontrées lors des plosives et des formes plus complexes sans chercher toutefois à réaliser une machine parlante.

T4.2 Les données générées sont en effet destinées à tester l'apprentissage profond sur des configurations plus simples que des géométries de conduits vocaux humains, ce qui est une innovation importante sur la stratégie d'utilisation des simulations acoustiques.

T4.3 Par ailleurs, la modélisation physique de la production de parole constitue un second enjeu, en particulier pour ce qui concerne les consonnes. Dans la continuité des travaux déjà amorcés notre approche consistera à tester différents modèles théoriques aéro-acoustiques sur des maquettes simplifiées ou d'impression 3D de géométries issues d'IRM afin d'en évaluer la pertinence.
Quelques simulations numériques seront réalisées en complément grâce au code LES développé au LEGI. L'intégration de ces modèles théoriques à la tête parlante sera réalisée en collaboration étroite avec le LORIA.

Choix technologiques

Les choix sont guidés par l'expérience obtenue sur le prototype. L'amélioration portera sur la conception d'une carte électronique dédiée à la place d'arduinos et sur l'utilisation d'actuateurs plus rapides et plus précis permettant de dépasser la limite actuelle de 500 mm/s, nécessaire pour simuler des occlusions comparables à des plosives.

Les simulations numériques seront réalisées avec les codes et moyens de calculs (HPC) du LEGI.

Contributions (qui fait quoi)

Cette tâche sera réalisée principalement par le **LEGI** avec des contributions du **Loria** et du **LPP**

Risques

1. Du point de vue expérimental, le risque peut être de ne pas avoir le banc prêt à temps pour réaliser la base de données nécessaire à la Tâche 3.
2. Les simulations numériques d'écoulements qui, dans des géométries aussi complexes que le conduit vocal peuvent présenter des problèmes de convergence.

Alternatives

1. Utiliser le prototype existant, bien que limité en performance dynamique.
2. Réaliser les simulations numériques sur des géométries simplifiées.

Tâche 5 - Exploration des capacités d'adaptation de la tête parlante à d'autres situations de production - **resp. LORIA**

AAPG2020	**Full3DTalkingHead**		PRC
Coordonné par :	Yves Laprie	42 mois	620 k€
CE23			

Objectif
L'objectif est de développer les possibilités de la tête parlante au-delà de ses fonctions de synthèse articulatoire directement issues de l'apprentissage à partir des corpus et des outils de simulation aéro-acoustique. Nous voulons pouvoir adapter la tête parlante à un nouveau locuteur, pouvoir modifier les expressions et explorer la compensation de perturbations imposées en bloquant une variable articulatoire latente.
Défis
- adapter la génération de la forme du conduit vocal et du visage à un nouveau locuteur à partir de points anatomiques issues d'IRM et d'un scan 3D du visage. - évaluation de l'impact de contraintes au niveau des variables latentes sur le comportement global de la tête parlante et étude de la compensation.
Indicateur de succès
- Possibilité d'adapter la tête parlante à un autre locuteur par morphing géométrique à partir d'une IRM statique et d'un scan du visage. - Possibilité de modifier les expressions et la forme du conduit vocal de la tête parlante aux niveaux acoustique et visuel
Livrables
Algorithmes et publications correspondant aux trois objectifs
Programme détaillé
T5.1 Recherche de points anatomiques sur les images IRM statiques et le visage et développement du morphing destiné à l'adaptation. **T5.2** Utilisation des variables latentes (cf. T3.3) appliquées aux émotions et traits articulatoires des phonèmes sur les corpus des deux locuteurs et modification de l'un des paramètres en simulant une perturbation (due à une contrainte anatomique ou physiologique) pour une resynthèse par simulation aéro-acoustique.
Choix technologiques
- Exploitation d'images IRM statiques et d'images du visage recalées sur les IRM pour partie du visage commune aux deux images. - Utilisation d'auto-encodeurs variationnels pour la mise en évidence de variables latentes
Contributions (qui fait quoi)
IADI pour l'adaptation anatomique, LORIA, LPP et LEGI pour la mise en évidence et la modification de variable latentes pour la mise en œuvre de compensations.
Risques
Assez faibles pour l'adaptation, plus élevés pour les variables latentes et la prise en compte de contraintes anatomiques ou physiologiques.
Alternatives
Il s'agit d'une partie plus exploratoire et la recherche d'alternatives n'est pas absolument indispensable.

Calendrier (personne×mois intégrant les permanents et non permanents hors stagiaires)

	Tâches	LORIA	IADI	LEGI	LPP	Chronogramme			
		Partenaires				**Année 1**	**Année 2**	**Année 3**	**Année 4**

AAPG2020	Full3DTalkingHead		PRC
Coordonné par :	Yves Laprie	42 mois	620 k€
CE23			

						I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II
T0	**Coordination**	5	2	2	2														
T 1	**Constitution du corpus**																		
1.1	Conception du corpus et sélection des locuteurs	6	4		4														
1.2 -4	Acquisition des données	12	14		12														
T2	**Prétraitement des données IRM, visage, glotte et passage 2D-3D**																		
2.1	Constitution de la base de données et débruitage	6	7		7														
2.2	Développement des réseaux de segmentation automatique	7	8																
2.3	Extension 2D —> 3D		12																
T 3	**Apprentissage du modèle de génération**																		
3.1	Modèle de contrôle de la tête parlante	16																	
3.2	Coordination de la source et les cordes vocales	12		8	9														
3.3	Interprétabilité du modèle	12			4														
T 4	**Apprentissage du contrôle acoustique**																		
4.1	Construction et validation de la maquette du conduit vocal			12	4														
4.2	Mesures aéro-acoustiques et apprentissage du contrôle	5		24															
4.3	Modèles et simulations aéro-acoustiques	4		12															
T 5	**Exploration des capacités d'adaptation de la tête parlante à d'autres situations de production**																		
5.1	Adaptation anatomique à une autre locuteur	4	4		6														
5.2	Simulation de perturbations	6		6	6														

II. Organisation et réalisation du projet

a. Coordinateur scientifique et son consortium / son équipe

Notre consortium est formé de quatre équipes de recherche remarquablement complémentaires avec des expériences théoriques et pratiques de premier plan international dans les domaines de l'IA, l'acoustique, de la phonétique expérimentale, de l'imagerie par IRM et du traitement automatique de la parole:

Equipe MultiSpeech du LORIA UMR 7503 (Nancy) **Yves Laprie (DR CNRS), Slim Ouni (MCf, Univ. Lorraine) et Vincent Colotte (MCf, Univ. Lorraine)** spécialistes en production et synthèse acoustique et audiovisuelle de la parole.
Laboratoire IADI (Imagerie Adaptative Diagnostique et Interventionnelle, unité INSERM U947) (Nancy) **Pierre-André Vuissoz (Ingénieur de recherche, UL) et Freddy Odille (CR, INSERM)** *spécialiste de l'IRM et de la reconstruction d'images.*
Laboratoire LPP (Laboratoire de Phonétique et de Phonologie) (Paris). **Didier Demolin (Prof. Université Sorbonne Nouvelle) et Angélique Amelot (IR CNRS)** spécialistes de phonétique expérimentale, d'aérodynamique et des plis vocaux.
Équipe EDT : Écoulements Diphasiques et Turbulences du LEGI (Laboratoire des Écoulements Géophysiques et Industriels) UMR 5519 (Grenoble) **Xavier Pelorson (DR CNRS), Annemie Van Hirtum (CR1 CNRS)** spécialistes de l'acoustique de la parole.
L'équipe Multispeech et IADI ont des contributions importantes en IA dans les domaines du traitement de la parole et de l'ECG en IRM.

Coordination du projet : Yves Laprie (CNRS-Loria) sera responsable de la coordination régulière du projet et assurera l'interface entre le consortium et l'ANR. Yves Laprie a une grande expérience des responsabilités collectives puisqu'il a animé l'équipe Parole (environ 30 personnes), plusieurs projets dont un projet européen et deux projets ANR dans le domaine de la production et de l'analyse de la parole, et qu'il est maintenant directeur de la structure de coordination des laboratoires du numérique de l'Université de Lorraine. Le service de gestion du Loria assurera le suivi des coûts, des documents budgétaires et de justifications des dépenses.
Équipe de coordination : La coordination du projet sera assurée par l'équipe formée de Yves Laprie (LORIA), Annemie Van Hirtum (LEGI), Pierre-André Vuissoz (IADI) et Angélique Amelot (LPP).

Responsables des tâches : Les responsables des tâches sont des membres permanents seniors des partenaires du projet. Ils seront responsables pour la coordination détaillée, la planification des travaux, le suivi et les interactions avec les autres tâches du projet.

Une part importante de la valeur ajoutée vient de la synergie qui existe à l'intérieur du consortium, et c'est un aspect essentiel à sa réussite. Au-delà des outils classiques de communication et de travail coopératif (réunions physiques ou à distance, mail, site web, gitlab…) que nous utiliserons, nous renforcerons la synergie grâce à des séjours de travail des deux doctorants, post-doctorants et chercheurs afin qu'ils maîtrisent parfaitement leur contribution et les interactions avec les autres acteurs du projet concernant les approches d'apprentissage profond, l'obtention des données géométriques de l'IRM ou physiologiques et les simulations acoustiques.

Tableau d'implication du coordinateur et des responsables scientifiques des partenaires dans d'autres projets en cours

Nom du participant au projet	Personne.mois	Intitulé de l'appel à projets, agence de financement, montant attribué	Titre du projet	Nom du coordinateur du projet	Date début - Date fin
Yves Laprie	6	ANR	ANR Benephidire	Fabrice Hirsch	2019-2022
Slim Ouni	6	ANR	ANR Benephidire	Fabrice Hirsch	2019-2022
P.-A. Vuissoz	6	ANR	ANR BRACOIL	Jacques Felblinger	2017-2020

b. Moyens mis en œuvre et demandés pour atteindre les objectifs

Partenaire 1 : LORIA/CNRS

Frais de personnel

Un ingénieur recruté pour 12 mois (47693€) dans le cadre de ce projet travaillera sur la préparation, l'enregistrement à l'aide du système de capture du mouvement et le prétraitement du corpus de données du visage.

Un doctorant (116 332€) travaillera sur l'apprentissage du contrôle temporel du conduit vocal et du visage à l'aide d'apprentissage profond et du suivi automatique des contours.

Coûts des instruments et du matériel

1 portables et 1 poste de travail fixe pour le doctorant (5000€).

Frais généraux additionnels et autres frais d'exploitation

Réunions d'avancement (5000€)

Conférences internationales (2 personnes par an x 3,5 ans) (11875€)

Frais de gestion 4% et frais de laboratoire 4% soit au total (14872€)

Partenaire 2 : IADI/Université de Lorraine

Frais de personnel

Un chercheur post doctorant recruté pour 20 mois (81000€) dans le cadre de projet travaillera sur l'acquisition des données IRM temps réel, leur synchronisation spatiale et temporelle ainsi que leur segmentation automatisée par DNN.

Coûts du recours aux prestations de service (et droits de propriété intellectuelle)

Le CIC-IT de Nancy fournira tout le support nécessaire pour inclure les sujets dans le protocole de recherche clinique permettant les acquisitions IRM (30 sessions IRM à 670 € la session + participations aux frais) (21100€).

Frais généraux additionnels et autres frais d'exploitation

Réunions d'avancement (2 personnes, 4 réunion/ans, pour 2 ans) (4800€)

Conférences internationales (2 personnes par an x 2 ans) (8000€)

Frais de gestion 4% et frais de laboratoire 4% soit au total (9192€)

Partenaire 3 : LEGI

Frais de personnel

Un doctorant (116 300 €) travaillera sur la modélisation physique (théorique et expérimentale) des consonnes et des interactions avec la source glottique.

4 stagiaires, niveau Master, sont prévus sur la durée du projet pour un total de 12 000 €. Ils travailleront sur la partie mécanique et électronique du banc expérimental et sur la réalisation des mesures.

Coûts des instruments et du matériel

15 000 € de matériel mécanique, électronique ont été prévus ainsi que 10 000 € de consommables (licences de logiciels).

Frais généraux additionnels et autres frais d'exploitation

Réunions d'avancement (2 personnes, 4 réunion/ans, pour 2 ans) (4800€)

Conférences internationales (2 personnes par an x 2 ans) (8000€)

Frais de gestion 4% et frais de laboratoire 4% soit au total (13944€)

Partenaire 4 : LPP

Frais de personnel

Un chercheur post doctorant recruté pour 20 mois (81000€) dans le cadre de projet travaillera sur blabla blabla

Coûts des instruments et du matériel

5000€ pour le matériel d'acquisition EPGG et d'accéléromètres et de connexion au pléthismographe.

Frais généraux additionnels et autres frais d'exploitation

Réunions d'avancement (2 personnes, 4 réunion/ans, pour 2 ans) (4800€)

Conférences internationales (2 personnes par an x 2 ans) (8000€)

Frais de gestion 4% et frais de laboratoire 4% soit au total (7899,36)

Moyens demandés par grand poste de dépense et par partenaire

		Partenaire *LORIA/CNRS*	**Partenaire** *IADI*	**Partenaire** *LEGI*	**Partenaire** *LPP*
Personnel		164024,95	81000,00	128300,00	83742,00
Coûts des instruments et du matériel (dont consommables scientifiques)		5000,00		25000,00	4000,00
Prestation de service et droits de propriété intellectuelle			21100,00		
Frais généraux additionnels et autres frais d'exploitation	Missions	16875,00	12800,00	21000,00	11000,00
	Frais d'environnement**	14872,00	9192,00	13944,00	7899,36
Sous-total		**200771,95**	**124092,00**	**188244,00**	**106641,36**
Aide demandée		**619749,31 €**			

III. Impact et retombées du projet

Relations entre science et société

Quel que puisse être leur niveau de complexité les recherches sur la production de la parole et la synthèse peuvent facilement être présentées au grand public à travers des interviews dans les chaînes radio ou télé régionales notamment. Le site web du projet présentera de nombreux documents (films IRM, suivi des articulateurs, dispositifs de capture du mouvement, maquettes acoustiques…) et nous ferons appel au soutien de nos tutelles pour réaliser un petit film documentaire (par exemple le dispositif Avant scène recherche de l'Université de Lorraine). Par ailleurs nous participerons aux actions de communication classiques (fête de la Science, accueil de lycéens et collégiens...).

Impact scientifique

La première retombée est une approche radicalement différente de la modélisation de la production de la parole. Jusqu'à présent ces approches, et singulièrement celle de la synthèse articulatoire, utilisaient des modèles numériques (géométrie du conduit vocal, coarticulation…) dont le cadre formel limitaient l'ajustement à des données réelles. Ici nous conservons la puissance de la contribution de la physique pour faire le lien entre les niveaux anatomiques, physiologiques et acoustique mais en utilisant les outils d'apprentissage de l'IA pour modéliser une grande partie du contrôle des niveaux anatomiques et physiologiques.

L'utilisation de l'apprentissage pour faire émerger des variables latentes contrôlant la forme du conduit vocal en fonction des phonèmes à articuler est une retombée majeure car l'utilisation d'un modèle articulatoire posait plus de questions qu'elle n'en résolvait. Que ce modèle soit purement géométrique ou construit à partir de données d'imagerie il fallait déterminer un vecteur de paramètres pour représenter chaque forme, ce qui posait la question de choisir un vecteur parmi d'autres puisqu'un grand nombre de vecteurs permettent d'approcher chaque forme du conduit vocal avec une

assez bonne précision. Au contraire, les variables latentes issues de l'apprentissage profond sont très proches du concept de variables articulatoires au sens de la phonologie articulatoire, mais avec l'avantage de reposer sur un vaste corpus de parole continue couvrant tous les gestes.

La seconde retombée scientifique concerne le couplage entre la résolution d'équations physiques et l'apprentissage profond. Dans notre cas il s'agit des équations de l'acoustique et l'enjeu est de pouvoir utiliser une maquette et des mesures physiques comme données d'apprentissage pour ensuite contrôler un conduit vocal simplifié. L'idée est de s'affranchir des incertitudes des paramètres physiques mal connus, des simplifications de la modélisation physique et bien sûr d'éviter des calculs lourds que nécessite la résolution d'équations complexes. Les travaux que nous mènerons dans ce projet peuvent aussi déboucher sur un couplage plus étroit entre la notion de fonction de perte de l'apprentissage profond et la résolution d'EDP à l'image des travaux de Michoski et al. (2020).

Dans ce projet nous sommes limités à deux locuteurs pour garantir la faisabilité en termes d'acquisition et de traitement des données et valider l'infrastructure de calcul. L'objectif est d'étendre ces travaux à un plus grand nombre de locuteurs et de langues.

Retombées technologiques et valorisation économique

La première retombée est le suivi des articulateurs dans les films IRM. Il s'agit d'une retombée technologique importante parce qu'elle ouvre de nouvelles possibilités d'investigation médicale pour de nombreuses pathologies de la parole d'un point de vue dynamique alors que jusqu'à présent il fallait se contenter d'une investigation essentiellement statique. Le fait de disposer du contour des articulateurs facilite une évaluation numérique des paramètres de vitesse, d'accélération et de contact avec les dents, le palais, ou les lèvres ce qui est essentiel puisque la parole est avant tout un processus dynamique.

Au-delà du diagnostic il deviendra possible de fournir un retour au sujet afin de le guider pour corriger ou apprendre des gestes articulatoires. Cette possibilité qui semblait trop coûteuse pour se concrétiser devient plus réaliste car le bénéfice est substantiel pour les patients et l'infrastructure de calcul existe déjà puisque l'imagerie fait appel à 8 cartes GPU. Il serait donc facile d'ajouter la puissance de calcul pour réaliser le suivi en temps réel. Le passage à l'imagerie dynamique et au suivi d'organes en temps réel et de manière fiable est une révolution en termes de l'utilisation de l'imagerie IRM dont on ne ne fait qu'entrevoir les applications. La première application concrète sera sans doute la kinésithérapie maxillo-faciale destinée à corriger le geste de déglutition ou la rééducation des gestes de la langue (Chauvoi et al. 1991) après un traumatisme maxillo-facial ou une réanimations longue.

La seconde retombée concerne la modélisation des lèvres dans les têtes parlantes. Actuellement, la modélisation n'est que partielle puisque les systèmes de capture du mouvement ne « voient » que l'extérieur des lèvres. Les modèles géométriques construits ou ajustés à partir de l'analyse des capteurs ont donc tendance à restituer la géométrie à partir de la partie visible, géométrie qui est donc souvent perçue comme insuffisamment réaliste. L'IRM statique et dynamique offre une vue complète qui pourra être fusionnée avec celle des capteurs données du visage afin d'atteindre un meilleur réalisme. Il s'agit d'un point extrêmement important du point de vue de déploiement d'applications de têtes parlantes et Slim Ouni est en train de monter un projet de startup qui pourra aussi exploiter ces résultats.

La troisième retombée technologique concerne la possibilité d'offrir un retour articulatoire dans le cadre de de la rééducation vocale ou de l'apprentissage des langues. Cette retombée devient réaliste parce que l'ensemble des gestes du conduit vocal et du visage ainsi que le signal de parole que pourra produire la tête parlante atteindront un réalisme suffisant.

Les partenaires du projet acceptent de conclure un accord de consortium basé sur le modèle unicANR (http://www.anrt.asso.fr/sites/default/files/unicanr-elucidation-mai-2010.pdf), afin de mettre en place les conditions et modalités de la réalisation du projet, de fixer leurs droits et devoirs respectifs et de préciser comment la propriété des résultats reviendra aux partenaires. La valorisation économique concernera le suivi des articulateurs et la synthèse de la forme géométrique du conduit vocal et du visage. Les autres algorithmes concernant l'acoustique seront diffusés sous la forme de logiciels libres pour une utilisation non commerciale et dans l'objectif de valoriser l'image du consortium et de déboucher sur de nouvelles coopérations scientifiques et des applications plus directement commerciales.

Le plan de gestion des données prendra en compte les aspects RIPH pour les acquisitions IRM (un CPP plus large que ce projet est en cours de montage pour couvrir toutes les acquisitions concernant la parole) et les aspects RGPD en sachant qu'un petit nombre de sujets volontaires seront concernés (mais pour un volume de données important) ce qui simplifiera la question des consentements.

IV. Bibliographie

Amelot, A., Sathiyanarayanan, D., Maeda, S., Honda, K., & Crevier-, L. (2018). Validation of a Noninvasive System to Observe Glottal Opening and Closing : External PhotoGlottoGraph (ePGG). 11th International Conference on Voice Physiology & Biomechanics, 10–11.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, *10*(7), e0130140.

Biasutto–Lervat, T., Dahmani S., & Ouni S. (2019) "Modeling Labial Coarticulation with Bidirectional Gated Recurrent Networks and Transfer Learning."

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7), e0130140.

Blandin R., Van Hirtum A., Pelorson X., Laboissière R. (2018) The effect on vowel directivity patterns of higher order propagation modes. J Sound Vib, 432:621-632.

Bustin A, Voilliot D, Menini A, Felblinger J, de Chillou C, Burschka D, Bonnemains L, Odille F (2018) Isotropic Reconstruction of MR Images Using 3D Patch-Based Self-Similarity Learning IEEE Trans Med Imaging. 2018 Aug;37(8):1932-1942. doi: 10.1109/TMI.2018.2807451 .

Butz, M. V., Bilkey, D., Humaidan, D., Knott, A., & Otte, S. (2019). *Learning, planning, and control in a monolithic neural event inference architecture. Neural Networks.*

Bresch E. , Kim Y.-C. , Nayak K., Byrd D. , Narayanan S. (2008), « Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging [Exploratory DSP] », IEEE Signal Process. Mag., vol. 25, no3, p. 123-132, mai 2008, doi: 10.1109/MSP.2008.918034.

Butz, M. V., Bilkey, D., Humaidan, D., Knott, A., & Otte, S. (2019). Learning, planning, and control in a monolithic neural event inference architecture. Neural Networks.

Chauvoi, A., Fournier M, Girardin F. (1991) Rééducation des fonctions dans la thérapeutique orthodontique, Vanves Éditions S.I.D

Cisonni J., Nozaki K., Van Hirtum A., Grandchamp X., Wada S. (2013) Numerical simulation of the influence of the orifice aperture on the flow around a teeth-shaped obstacle. Fluid Dynamics Research, 45:1-19.

Cisonni J., Nozaki K., Van Hirtum A., Wada S. (2011) A parameterized geometric model of the oral tract for aeroacoustic simulation of the fricatives. Int. J. of Information and Electronics Engineering (IJIEE), 1:223-228

Delebecque L., Pelorson X., Beautemps D. (2016) Modeling of aerodynamic interaction between vocal folds and vocal tract during production of a vowel–voiceless plosive–vowel sequence. The Journal of the Acoustical Society of America 139, 350-60

Demolin D., Metens T., Soquet A., (1996) « Three-dimensional measurement of the vocal tract by MRI », in Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96, Philadelphia, PA, USA, 1996, vol. 1, p. 272-275, doi: 10.1109/ICSLP.1996.607098.

Eslami M., Neuschaefer-Rube C., Serrurier A. (2020) « Automatic vocal tract landmark localization from midsagittal MRI data », Sci. Rep., vol. 10, no 1, p. 1468, doi: 10.1038/s41598-020-58103-6.

Guenther, F.H. (2016). Neural Control of Speech. Cambridge, MA: MIT Press.

Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., ... & Wu, Y. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in Neural Information Processing Systems* (pp. 4480-4490).

Isaieva, K, Laprie Y., Houssard A., Felbinger, J., Vuissoz, P.-A. (2020). Tracking the tongue contours in rt-MRI films with an autoencoder DNN approach. Submitted to ISSP 2020

Isola P. , Zhu J.-Y., Zhou T., Efros A. A. (2018), « Image-to-Image Translation with Conditional Adversarial Networks », ArXiv161107004 Cs, nov. 2018, http://arxiv.org/abs/1611.07004

Joyce T. , Chartsias A. , Tsaftaris S. A. (2018), « Deep Multi-Class Segmentation Without Ground-Truth Labels», p. 9. 1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam.

Kingma, Diederik & Welling, Max. (2014). Auto-Encoding Variational Bayes. https://arxiv.org/pdf/1312.6114.pdf

Kwon G. , Han C. , Kim D. (2019), « Generation of 3D Brain MRI Using Auto-Encoding Generative Adversarial Networks », ArXiv190802498 Cs Eess, août 2019. http://arxiv.org/abs/1908.02498.

Labrunie, M., Badin, P., Voit, D., Joseph, A. A., Frahm, J., Lamalle, L., Boë, L.-J. (2018). Automatic segmentation of speech articulators from real-time midsagittal MRI based on supervised learning. Speech Communication, 99, 27–46.doi:10.1016/j.specom.2018.02.004

Lingala S. G. , Sutton B. P. , Miquel M. E. , Nayak K. S. (2016) « Recommendations for real-time speech MRI: Real-Time Speech MRI », J. Magn. Reson. Imaging, vol. 43, no 1, p. 28-44, janv. 2016, doi: 10.1002/jmri.24997.

Öhman, S. E. G. (1966). Coarticulation in VCV Utterances: Spectrographic Measurements. The Journal of the Acoustical Society of America, 39(1), 151–168. doi:10.1121/1.1909864

Ozerov, A.; Vincent, E.; Bimbot, F. A (2012), General flexible framework for the handling of prior information in audio source separation. IEEE Trans. Audio Speech Lang. Process. 2012, 20, 1118–1133. doi: 10.1109/TASL.2011.2172425

Michoski, C., Milosavljević, M., Oliver, T., & Hatch, D. (2020). Solving Differential Equations using Deep Neural Networks. Neurocomputing. doi:10.1016/j.neucom.2020.02.015

Odille F., Vuissoz P.-A. , Marie P.-Y. , Felblinger J. (2008), « Generalized Reconstruction by Inversion of Coupled Systems (GRICS) applied to free-breathing MRI », Magn. Reson. Med., vol. 60, no 1, p. 146-157, juill. 2008, doi: 10.1002/mrm.21623.

Öhman, S. E. G. (1966). Coarticulation in VCV Utterances: Spectrographic Measurements. The Journal of the Acoustical Society of America, 39(1), 151–168. doi:10.1121/1.1909864

Pont A., Guasch O., Baiges J., Codina R. Van Hirtum A. (2019) Computational aeroacoustics to identify sound sources in the generation of sibilant /s/. Int. J. for Numerical Methods in Biomedical Engineering, 35:e3153

Van Hirtum A., Pelorson X., Estienne O., Bailliet H. (2011) Experimental validation of flow models for a rigid vocal tract replica. J. Acoust. Soc. Am., 130(4):2128-2138.

Van Hirtum A., Blandin R., Pelorson X. (2016) A setup to study aero-acoustics for finite length ducts with time-varying shape. Applied Acoustics, 105:83-92.

Yoshinaga T., Van Hirtum A., Nozaki K., Wada S. (2020) Acoustic modeling of fricative /s/ for an oral tract with rectangular cross-sections. J Sound Vib, 476:115337.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833).

Zhu Y. et al. (2013) « Dynamic 3D Visualization of Vocal Tract Shaping During Speech », IEEE Trans. MEDICALIMAGING, vol. 32, no 5, p. 838-848, mai 2013. doi: 10.1109/TMI.2012.2230017

AAPG2020	**Full3DTalkingHead**		PRC
Coordonné par :	Yves Laprie	42 mois	620 k€
CE23			