ArtSpeech
Appel à projet générique 2015
Synthèse articulatoire phonétique

Durée du projet : 42 mois
Aide totale demandée : 500118€

1. RESUME DE LA PROPOSITION DE PROJET

L'objectif est de synthétiser de la parole à partir du texte en simulant numériquement le processus physique de production de la parole chez un humain, c'est-à-dire les aspects articulatoires, aérodynamiques et acoustiques.
Les approches à base de corpus ont pris une place hégémonique en synthèse de la parole. Elles exploitent des bases de données acoustiques de très bonne qualité tout en couvrant un grand nombre d'expressions et de contextes phonétiques, ce qui suffit à produire de la parole intelligible. Malgré cela, ces approches font face à des obstacles presque insurmontables dès qu'il faut modifier des paramètres intimement liés au processus physique de production de la parole. Au contraire, une approche reposant sur la simulation du processus de production fait explicitement appel aux paramètres de la source, à l'anatomie et la géométrie du conduit vocal, ainsi qu'à une stratégie de supervision temporelle. Elle offre donc un contrôle direct de la nature de la parole synthétique.

Ce projet s'organise en 5 tâches :

1. **Simulations aérodynamiques et acoustiques** afin de produire le signal acoustique de parole connaissant l'aire transverse en tout point de toutes les cavités du conduit vocal.
2. **Source et scenarii de coordination afin de** coordonner les sources avec l'évolution temporelle de la forme du conduit vocal, ce qui est crucial lors de la production des consonnes pour assurer leur identification par des auditeurs humains.
3. **Contrôle de l'évolution temporelle de la géométrie du conduit vocal** afin d'anticiper la production des sons à venir et produire des gestes articulatoires réalistes.
4. **Acquisition de données de production de la parole** indispensables pour connaître l'activation des plis vocaux, les paramètres aérodynamiques, et la forme géométrique du conduit vocal (grâce à l'IRM à cadence élevée).
5. **Architecture générale** pour intégrer les différents niveaux et synthétiser un signal acoustique à partir du texte.

Le développement de simulations réalistes des processus de production de la parole sera un atout absolument déterminant pour comprendre les contributions respectives des caractéristiques anatomiques, des capacités de

coordination, et du contrôle des plis vocaux dans le signal produit. La portée de ce projet va bien au-delà de la compréhension des processus de la production de la parole et concerne la phonétique, le contrôle moteur, et dans le domaine du traitement automatique de la parole la synthèse à partir du texte.

Les applications sont très étendues. Elles concernent les situations dans lesquelles la synthèse de la parole standard n'est pas bien adaptée comme c'est le cas pour l'apprentissage des langues étrangères ou l'acquisition du langage. Ce projet ouvre aussi de nouvelles perspectives dans le domaine de la synthèse de parole expressive avec des répercutions attendues dans le cadre des agents conversationnels. Dans le domaine médical les applications portent sur les algorithmes d'acquisition IRM à cadence élevée qui concernent les organes se déformant rapidement au cours du temps, et sur les pathologies de la production de la parole, ou l'impact des interventions chirurgicales sur le conduit vocal.

Nous avons la conviction profonde que ArtSpeech réalisera des avancées scientifiques et techniques majeures et apportera ainsi la preuve de l'intérêt de l'approche physique qu'il s'agisse d'ouvrir de nouvelles perspectives de recherche, ou d'applications très innovantes dans le domaine de la production de la parole au sens large.
Le consortium est formé de quatre équipes de recherche remarquablement complémentaires avec des expériences théoriques et pratiques de premier plan international dans les domaines de :

- la simulation aérodynamique et acoustique de la production de la parole, et la modélisation de la source et de la géométrie du conduit vocal,
- l'imagerie par résonance magnétique et les autres techniques d'acquisition de données de parole.

Personnes impliquées dans le projet :

Partenaire	Nom	Prénom	Emploi actuel	Implication dans le projet en personne.mois**	Rôle & responsabilité dans le projet
LORIA	Laprie	Yves	DR2 CNRS	18	Coordinateur scientifique Modélisation et synthèse articulatoires, modèle de coarticulation
LORIA	Ouni	Slim	Maître de conférences / Université de Lorraine	12	Coordination des gestes articulatoires et mesures articulatoires multimodales
LORIA	Colotte	Vincent	Maître de conférences / Université de Lorraine	12	Synthèse de la parole à partir du texte, Analyse du langage naturel, Architecture du système de synthèse
GIPSA-lab	Pelorson	Xavier	DR2 CNRS	12	Acoustique du conduit vocal et modèle des plis vocaux
GIPSA-lab	Van Hirtum	Annemie	CR1 CNRS	8	Turbulence et modélisation des fricatives
GIPSA-lab	Laval	Xavier	IE G-INP	2	Support technique pour les mesures sur maquettes
IADI	Vuissoz	Pierre-André	Ingénieur de recherche – Université de Lorraine	12	Acquisition de données IRM du conduit vocal
IADI	Odille	Freddy	CR1 - INSERM	8	Acquisition de données IRM du conduit vocal
LPP	Demolin	Didier	Professeur/ Université Paris 3 Sorbonne nouvelle	12	Acquisition de données sur la source de parole Coordination entre les plis vocaux et le conduit vocal
LPP	Amelot	Angélique	Ingénieur de recherche CNRS	12	Production et Acquisition de données sur la source de parole

** à renseigner par rapport à la durée totale du projet

2. CONTEXTE, POSITIONNEMENT ET OBJECTIFS DE LA PROPOSITION

2.1. Objectifs et caractère ambitieux et novateur du projet

2.1.1. OBJECTIFS

L'objectif est de synthétiser de la parole à partir du texte en simulant numériquement le processus physique de production de la parole chez un humain, c'est-à-dire :

- les gestes des articulateurs de la parole (mandibule, langue, lèvres, voile du palais, larynx),
- le contrôle des plis vocaux dont la vibration est l'une des sources sonores qui excitent le conduit vocal,
- les phénomènes aérodynamiques qui modifient le flux d'air depuis les plis vocaux jusqu'aux lèvres,
- la propagation de l'onde acoustique dans le conduit vocal pour donner le signal acoustique tel qu'il est perçu par des auditeurs,
- et enfin les interactions entre ces différents éléments.

La synthèse articulatoire fait le lien entre la forme du conduit vocal, les plis vocaux et le signal de parole produit. Il s'agit donc d'un outil précieux pour contrôler toutes les facettes de la production de la parole qui donnent un timbre naturel à la parole humaine, et pour ***analyser en parallèle l'origine des indices acoustiques naturels et l'impact des gestes articulatoires***. Cette solution proposée au début des années quatre-vingt avait dû être abandonnée faute de disposer de techniques de simulation et de données d'une qualité suffisante fournissant des informations pertinentes. Cette situation s'expliquait par les obstacles auxquels était confrontée une approche physique en matière de modèles physiques, d'outils de simulation numérique et de dispositifs d'acquisition de données articulatoires.

Nous considérons que des progrès très significatifs ont été faits dans ces trois domaines sans apporter pour l'instant une preuve suffisamment claire de leur potentiel pour aborder la production et la synthèse de la parole. Nous avons la conviction profonde que ArtSpeech réalisera de nouvelles avancées dans ces trois domaines et apportera ainsi la preuve de l'intérêt de l'approche physique qu'il s'agisse d'ouvrir de nouvelles perspectives de recherche, ou d'applications très innovantes dans le domaine de la production de la parole au sens large. Il s'agira donc d'une avancée majeure.

Les approches à base de corpus ont pris une place hégémonique dans le domaine du traitement automatique de la parole. C'est en particulier le cas en synthèse car elles exploitent des bases de données acoustiques de très bonne qualité tout en couvrant un grand nombre de contextes phonétiques et d'expressions, ce qui suffit à produire de la parole de bonne qualité. Malgré cela, ces approches font face à des obstacles presque insurmontables dès qu'il faut modifier des paramètres intimement liés au processus physique de production de la parole. C'est en particulier le cas des paramètres d'activation des plis vocaux qui permettent par exemple de modifier le quotient entre phase ouverte et fermée de la glotte afin de modifier la qualité de la voix qu'il s'agisse de copier une pathologie des plis vocaux ou de transmettre une expression. L'objectif de développer un vrai modèle direct et performant, et de l'exploiter comme tel, marque une rupture forte par rapport aux techniques actuelles qui se fondent avant tout sur l'utilisation de corpus.

Ce projet fait le lien entre les domaines articulatoire et acoustique et pour cette raison il ouvrira de nouvelles perspectives de recherche dans les sciences de la parole. Il conduira aussi à de nouvelles approches de la synthèse de la parole, de nouvelles méthodes d'évaluation de l'impact acoustique des pathologies du conduit vocal ou des plis vocaux, et de nouveaux outils pour l'acquisition du langage ou des langues. Cette perspective de recherche est appelée à connaître un développement rapide pourvu que les outils de simulation puissent être utilisés facilement tout en offrant un niveau de réalisme suffisant du point de vue de la production de la parole.

Une série de travaux récents (dont certains menés par des partenaires de ce projet) concernant un modèle aérodynamique temporel discret du conduit vocal ont clairement démontré l'intérêt de la simulation pour étudier la production de la parole. L'un des avantages de la simulation est d'offrir ***un accès à un certain nombre de paramètres physiques qui sinon seraient très difficilement mesurables***.

Nous travaillerons d'abord sur le français parce qu'il est plus facile pour nous d'acquérir des données pour cette langue, mais bien sûr l'approche présente l'intérêt d'être très indépendante de la langue.

2.1.2. DÉFIS ET VERROUS SCIENTIFIQUES

Les défis que se propose de relever ArtSpeech sont de :

1. Réaliser la synthèse articulatoire de la parole, en s'approchant au mieux de la parole naturelle, ce qui nécessite d'élaborer de meilleures simulations acoustiques pour ce qui concerne les modèles physiques des bruits de friction notamment, et de prendre en compte la complexité géométrique du conduit vocal, en particulier les différentes cavités qui interviennent lors de l'articulation des consonnes,
2. Coordonner temporellement les gestes de la glotte et ceux des articulateurs de la parole,
3. Acquérir des données articulatoires et aérodynamiques qui permettent de valider les simulations acoustiques et algorithmes de contrôle de la forme du conduit vocal,
4. Développer un système de synthèse articulatoire opérationnel.

Faute d'une coopération étroite entre acousticiens et chercheurs en parole, le verrou le plus général est que les tentatives précédentes ont porté sur des points très précis, par exemple la simulation du comportement des plis vocaux sans prendre en compte un conduit vocal réaliste du point de vue géométrique et de son évolution temporelle, ou inversement la détermination des caractéristiques acoustiques du conduit vocal avec des modèles de plis vocaux insuffisants.

Le second verrou est l'absence d'une approche aéro-acoustique suffisamment élaborée, générale et bien fondée du point de vue de la physique. Les tentatives précédentes ont donné lieu à des modèles partiels, nécessitant des paramètres de contrôle artificiels, et donc sans rapport avec les commandes utilisées par l'être humain. Cela a donc conduit à des simulations insuffisamment réalistes et lourdes à mettre en œuvre.

Le troisième verrou est lié à l'absence de données couvrant tous les aspects de la production de parole, c'est-à-dire les paramètres aérodynamiques dont la pression sous-glottique, la pression intraorale, les débits à la bouche et aux narines, la vibration des plis vocaux, et d'autre part l'évolution temporelle de la géométrie du conduit vocal avec des résolutions temporelle et spatiale suffisamment élevées. L'acquisition des données est destinée à élaborer des modèles géométriques et physiques d'une part, et à valider des modèles physiques lors de la rupture brutale d'une occlusion dans le conduit vocal par exemple. Les partenaires du projet disposent des compétences couvrant toutes les technologies nécessaires.

Le quatrième verrou concerne le contrôle de l'évolution temporelle de la géométrie du conduit vocal. Plusieurs théories et modèles ont été proposés, en particulier la phonologie articulatoire associée à la dynamique des tâches qui apportent un certain nombre de réponses intéressantes. Leur point faible est de nécessiter un grand nombre de paramètres pour décrire les gestes articulatoires se recouvrant dans le temps, ce qui conduit à réduire leur flexibilité en appliquant des contraintes trop fortes au risque de ne pas garantir la réalisation des indices acoustiques observés en parole naturelle.

Les techniques d'acquisition de données sur la production de la parole, comme la modélisation numérique des phénomènes aérodynamiques, acoustiques et articulatoires de la parole ont fait des progrès importants ces dernières années, en particulier grâce aux contributions des partenaires de ce projet, et vont permettre de relever le défi de la synthèse articulatoire avec de bonnes chances de succès.

2.1.3. RÉSULTATS ATTENDUS

La contribution de ArtSpeech relève de la recherche fondamentale sur la parole. Comme le Plan d'action 2015 de l'ANR le précise page 72, elle s'inscrit donc l'axe 7 « Interactions humain-machine, objets connectés, contenus numériques, données massives et connaissance » du défi 7 « Société de l'Information et de la communication ».
Le développement de simulations réalistes des processus de production de la parole sera un atout absolument déterminant pour comprendre les contributions respectives des caractéristiques anatomiques, des capacités de coordination, et du contrôle des plis vocaux dans le signal de parole produit. La portée de ce projet va bien au-delà de la compréhension des processus de la production de la parole et concerne la phonétique, le contrôle moteur et dans le domaine du traitement automatique de la parole, la synthèse à partir du texte.
La possibilité d'ajouter des expressions à une voix synthétique est en effet devenue un enjeu majeur, notamment dans le cadre de la mise en œuvre d'agents conversationnels. L'origine physiologique des expressions dépend notamment de la tension appliquée aux plis vocaux et de la modification des gestes

articulatoires, c'est-à-dire deux des paramètres d'entrée de la synthèse articulatoire. Or actuellement, tous les systèmes opèrent par concaténation ou modélisation stochastique, et l'une des seules solutions consiste à étendre la taille du corpus de parole qu'ils explorent en ajoutant plusieurs styles de parole. Au contraire, la synthèse articulatoire donne accès directement à ces paramètres ce qui représente un avantage considérable.

Les applications sont très étendues. Elles concernent les situations dans lesquelles la synthèse de la parole standard n'est pas bien adaptée comme c'est le cas pour l'apprentissage des langues étrangères ou l'acquisition du langage. Dans les deux cas la possibilité d'intervenir directement sur le geste articulatoire en le corrigeant ou, au contraire, en l'exagérant pour montrer son impact acoustique est un avantage déterminant. Cette flexibilité de la synthèse peut aussi être exploitée dans le domaine de la santé pour évaluer, avant ou après un acte chirurgical dans le conduit vocal, les capacités de production de la parole du patient par exemple.
La possibilité de contrôler l'expressivité de la parole synthétique concerne quant à elle toutes les applications dans lesquelles il faut maintenir la motivation de l'utilisateur, depuis les jeux jusqu'aux situations d'enseignement.

2.1.4. CARACTÈRE NOVATEUR

Le premier caractère novateur de ArtSpeech est d'utiliser un modèle physique pour aborder un problème de traitement automatique de la parole abordé actuellement quasi exclusivement par le biais d'approches à base de corpus.
Pour cela, et il s'agit là encore d'un point original, nous mettons la synthèse articulatoire au centre de problèmes (modélisation des plis vocaux, simulations aérodynamiques et acoustiques, modélisation articulatoire, coarticulation, acquisition de données dynamiques portant sur le conduit vocal, les plis vocaux et les paramètres aérodynamiques) habituellement abordés indépendamment les uns des autres. Cela est possible grâce à la remarquable complémentarité des partenaires du projet.
D'autres points importants présentés dans le programme de travail font l'originalité de notre projet : **(i)** l'utilisation d'un modèle physique bien-fondé pour limiter le nombre de paramètres de contrôle de la simulation aéroacoustique, **(ii)** la conception d'un modèle de source turbulente pour les fricatives et l'extension des modèles pour les plosives **(iii)** l'utilisation, ou le développement, de nouveaux dispositifs d'acquisition de données aérodynamiques et sur les plis vocaux qui ne perturbent pas la production de la parole, **(iv)** l'élaboration de scenarii de coordination source-conduit vocal pour les consonnes, **(v)** le développement de modèles articulatoires capables de produire des consonnes et l'adaptation de ces modèles à un locuteur quelconque, **(vi)** le développement de séquences d'acquisition IRM rapide pour étudier la coarticulation et la relâchement rapide des occlusions dans le conduit vocal.

2.2. État de l'art

Trois approches de la synthèse de la parole existent ou sont imaginables. La première, celle utilisée en pratique, consiste à exploiter de vastes corpus de parole enregistrée soit en concaténant de petits segments acoustiques extraits de ce corpus qui formeront le signal synthétique, soit en entraînant des modèles de Markov utilisés pour synthétiser des phrases. Malgré la très bonne qualité du signal synthétisé cette approche superficielle ne fournit aucune information utile sur la production de la parole elle-même.

L'approche biomécanique complétée par la résolution des équations de Navier-Stokes se situe aux antipodes de la synthèse concaténative car elle tente de copier fidèlement la déformation des muscles et organes impliqués dans la production de la parole ainsi que les détails des phénomènes aéro-acoustiques. En dépit de son intérêt à très long terme son utilisation dans le cadre d'applications se heurte d'une part à toute une série d'obstacles liés aux modèles numériques et simulations nécessaires, et d'autre part à la difficulté de recueillir toutes les données physiologiques et anatomiques que requièrent les simulations numériques.

La synthèse articulatoire, c'est-à-dire la troisième approche, présente l'intérêt majeur de fournir une modélisation profonde des processus de production de la parole tout en conservant des modèles et simulations numériques d'une complexité acceptable. Par ailleurs, et il s'agit là d'un avantage essentiel, la ***synthèse sera explicitement contrôlée par des paramètres anatomiques et physiologiques***. Cela contribuera à une meilleure

compréhension de la production de la parole. Qui plus est, tant ***les dispositifs d'acquisition de mesures géométriques ou temporelles du conduit vocal, que les modèles physiques et les simulations numériques associées ont fait des progrès tels qu'il semble très probable que la synthèse articulatoire conduise à des applications réelles.***

2.2.1. SYNTHÈSE ARTICULATOIRE

Aéroacoustique

Du point de vue de la physique, la production des sons de parole résulte de la conversion d'un écoulement aérien provenant des poumons en une perturbation acoustique. Ceci est réalisé notamment par l'auto-oscillation des plis vocaux, dans le cas des sons voisés, par le relâchement rapide d'une occlusion du conduit vocal, dans le cas des plosives ou bien encore par les fluctuations internes d'un écoulement turbulent en interaction avec les parois du conduit vocal, dans le cas des fricatives. La modélisation de ces évènements est complexe parce qu'elle implique la description conjointe de phénomènes relevant du domaine de la mécanique, de la mécanique des fluides, de l'acoustique ainsi que des leurs interactions c'est-à-dire de l'aérodynamique ou de l'aérodistorsion et de l'aéroacoustique.
Pour des raisons pratiques (puissance des calculateurs) et théoriques (validation des algorithmes, stabilité, définition des conditions aux limites), la simulation numérique directe de ces phénomènes reste à l'heure actuelle limitée à l'étude de configurations ponctuelles, souvent très simplifiées. Même ainsi restreinte, la simulation numérique de l'acoustique du conduit vocal nécessite typiquement une dizaine d'heures [6]. Celle de l'écoulement au travers des cordes vocales se compte en jours sur un supercalculateur [20]. La simulation numérique directe reste un outil coûteux et inadapté aux fins de la synthèse. Une autre approche consiste à développer des modèles théoriques simplifiés, visant à rendre compte des phénomènes physiques essentiels.

Ainsi, plutôt que de considérer une infinité de degrés de liberté, les modèles distribués de type masse-ressort [18] ne prennent en compte que les modes de vibration principaux des cordes vocales. De la même manière, la méthode modale en acoustique permet de modéliser, de manière très efficace et avec une précision comparable à la simulation numérique par Eléments Finis [6], des phénomènes tridimensionnels qui ne peuvent pas être prédits par les théories classiquement utilisées en parole (modèles d'ondes planes).

En ce qui concerne la modélisation de l'écoulement aérien, l'analyse théorique [42, 8], soutenue par des travaux expérimentaux [47, 48], permet d'identifier la nature des phénomènes physiques les plus importants et de rationaliser un certain nombre d'hypothèses simplificatrices. L'analyse dimensionnelle des équations de Navier Stokes appliquée au larynx, permet ainsi de justifier l'utilisation de modèles d'écoulement localement incompressible (faible nombre de Helmoltz), quasi-stationnaire (faible nombre de Strouhal) et à faible viscosité (nombre de Reynold modéré). De même, l'utilisation d'analogies aéroacoustiques permet de modéliser la production de son par turbulence sans nécessiter une description détaillée de l'écoulement [17, 22].

De fait, depuis une quarantaine d'années, un nombre considérable de travaux ont été réalisés dans le but de développer, de tester et d'améliorer ces modèles physiques simplifiés. Le champ d'applications est également vaste puisqu'il va de la recherche fondamentale à l'étude des pathologies [26, 51, 8], du contrôle moteur [27] et de la synthèse de la parole [43, 54], bien sûr.

Modélisation articulatoire

Les simulations aérodynamique et acoustique nécessitent la connaissance de la géométrie instantanée du conduit vocal depuis la glotte jusqu'aux lèvres, et plus précisément l'aire transverse à la propagation de l'onde sonore qui est appelée fonction d'aire.

La première approche consiste à donner directement la fonction d'aire sans modéliser la géométrie bi ou tridimensionnelle du conduit vocal. C'est la solution adoptée dans les premiers temps par Fant [11] et très récemment par Story [53] qui génère une fonction d'aire comme le produit de la fonction d'aire vocalique résultant de la transition entre deux voyelles consécutives, et des différentes constrictions appliquées pour

réaliser les consonnes. Cette solution demande une optimisation fine des deux contributions. Elle est donc bien adaptée au cas où l'on dispose d'une connaissance assez précise de l'évolution de la fonction d'aire qui sert à régler les paramètres des constrictions utilisées pour les consonnes, et de l'évolution des formes vocaliques. Cela est possible pour traiter un petit nombre de phrases mais très difficilement généralisable pour le problème de la synthèse à partir du texte.
La seconde approche s'appuie sur un modèle articulatoire qui consiste à décrire la forme du conduit vocal à partir d'un petit nombre de primitives. Le modèle peut être bidimensionnel, et représenter la coupe médiosagittale du conduit vocal, ou tridimensionnel. Il existe deux familles de modèles. La première issue des travaux de Mermelstein [34] fait appel à des primitives purement géométriques (arcs de cercle et segments de droite) et a donné lieu ensuite à des versions tridimensionnelles comme celle proposée par Birkholz [5]. Ces primitives n'ont d'autre justification qu'une analogie de forme avec les images du conduit vocal. La seconde famille de modèles fait appel à des techniques d'analyse de données appliquées à un corpus d'images médicales du conduit vocal. Le plus connu de ces modèles est celui de Maeda [30] initialement développé pour des voyelles. D'autres versions ont été développées soit pour prendre en compte les consonnes [23, 25] en deux dimensions, soit en trois dimensions [1]. Dans le cas d'un modèle tridimensionnel il faut utiliser des images IRM statiques ce qui d'une part risque de biaiser légèrement l'articulation et d'autre part nécessite de pouvoir acquérir des images d'une variabilité phonétique suffisante ce qui n'est pas simple.
Il est ensuite nécessaire de calculer la fonction d'aire à partir de cette représentation géométrique ce qui consiste à découper le conduit en petits tubes perpendiculairement à la ligne centrale du conduit vocal [32]. Cette ligne est déterminée de manière à correspondre au cheminement de l'onde plane dans le conduit. Sa détermination a donné lieu à de nombreux algorithmes car elle a un impact direct sur le découpage du conduit, et par conséquent les fréquences de résonance du conduit vocal.

2.2.2. COORDINATION DES GESTES ARTICULATOIRES ET DE LA SOURCE SONORE

Les articulateurs de la parole qui font évoluer la forme du conduit vocal, et par conséquent ses propriétés acoustiques, doivent se déplacer en fonction de la suite de sons à prononcer. Par ailleurs, il faut aussi que la réalisation des cavités et des constrictions supra-glottales soit coordonnée avec l'aire à la glotte pour que les propriétés aérodynamiques du conduit vocal soit compatibles avec la source d'excitation sonore bruitée ou voisée.
Ces deux mécanismes de coordination ont des échelles de temps différentes. Au niveau segmental, celui des sons, l'ordre de grandeur de la coordination entre la source et les cavités supra-glottales est la milliseconde, tandis qu'au niveau suprasegmental la coordination entre les articulateurs supra-glottaux est de l'ordre d'une centaine de millisecondes. Il est crucial de coordonner les deux niveaux pour obtenir un signal de parole réaliste.

Coarticulation

Le second niveau de coordination donne lieu aux phénomènes de coarticulation et implique les articulateurs de la parole. Ils peuvent anticiper les positions à venir ce qui donne lieu à la coarticulation anticipatrice, ou être influencés par leur position précédente, ce qui correspond à la coarticulation rétentrice.
De très nombreuses théories issues de la phonétique ou de la phonologie ont été avancées pour expliquer ces phénomènes (voir par exemple le panorama dressé par Farnetani [12]) et en prédire la direction ou l'ampleur. L'un des points les plus étudiés porte sur le début de l'anticipation qui intervient dès que possible en l'absence de contrainte imposée pour la théorie « Look ahead » [14, 3], ou avec un décalage temporel fixe par rapport au son à produire pour la théorie « Time locked » [10]. Ces modèles de base ont été déclinés en de nombreux modèles destinés à expliquer les données articulatoires collectées.
Par ailleurs, plusieurs modèles numériques ont été proposés cette fois pour prédire la coarticulation, soit dans le cas général du conduit vocal, soit dans le cas de la coarticulation labiale. Il faut noter que le modèle de Öhman proposé en 1966 et modifié en 1967 [39, 40] reste l'un des modèles les plus fréquemment utilisés soit sous sa forme initiale, soit sous une forme améliorée. Le travail récent de Birkholz [4] en est conceptuellement très proche.
La phonologie articulatoire proposée par Browman et Goldstein [41] en 1989 apporte une réponse intéressante aux questions soulevées par les théories de la coarticulation même si elle a été avant tout conçue comme une

théorie phonologique. Ses primitives et unités distinctives sont les gestes articulatoires destinés à réaliser une constriction dans le conduit vocal. Chaque constriction recrute en général plusieurs articulateurs (par exemple la mâchoire, le corps et la pointe de la langue pour réaliser une constriction avec la pointe de la langue). Les gestes peuvent se recouvrir dans le temps. S'il est assez facile de prévoir les gestes en fonction de la suite des phonèmes à articuler il est en revanche nettement plus difficile de calculer l'activation précise de chacun des articulateurs [35] puisqu'il n'existe pas d'étiquetage de la parole en termes d'activation des gestes articulatoires.

Coordination source – conduit vocal

Les propriétés aérodynamiques du conduit vocal depuis la glotte jusqu'aux lèvres déterminent la nature de la source d'excitation. En particulier, le flux d'air qui traverse le conduit vocal est déterminé par l'aire la plus petite entre la glotte et les constrictions du conduit vocal. Globalement, il est nécessaire que la pression sous-glottique soit plus élevée que la pression intra-orale pour que les plis vocaux vibrent. Mais il est possible de trouver des stratégies pour faire diminuer ponctuellement la pression intra-orale quand la différence devient trop faible, par exemple en gonflant légèrement les joues pour faire augmenter le volume intra-oral. Pour créer une source de bruit en avant d'une constriction il est nécessaire de créer une turbulence et donc que le flux d'air ne soit pas stoppé par les plis vocaux qui doivent donc être écartés. Il existe un certain nombre de travaux sur la coordination entre la source et les cavités du conduit vocal. Scully [50] propose des scenarii qui couvrent la plupart des sons et Maeda [31] associe les simulations aérodynamique et acoustique du conduit vocal en proposant des scenarii de coordination qui donnent des résultats assez proches de la parole naturelle. Nous avons d'ailleurs repris ces scenarii dans nos expériences de synthèse articulatoire par copie [24]. Même si les simulations ont fait de nombreux progrès elles ne suffisent en général pas à expliquer les données recueillies lors de la production de parole comme le montre les travaux de McGowan et al. [33], en particulier car les interactions entre le conduit vocal et les plis vocaux sont insuffisamment étudiés.

2.2.3. ACQUISITION DE DONNEES DYNAMIQUES SUR LA PRODUCTION DE LA PAROLE

L'acquisition de données sur la glotte est une tâche assez difficile. L'observation directe des plis vocaux requiert une source lumineuse et une caméra qui doit être introduite dans la cavité pharyngale pour les visualiser. Ceci réduit la variété de sons qui peuvent être étudiés à des voyelles qui ont une large cavité pharyngale.

Ces raisons ont fait qu'un photoglottographe non-invasif [15, 16] a été développé au LPP pour observer les changements d'ouverture glottale pendant la parole. Ceci peut être appliqué aussi bien à des d'études phonétiques qu'à des questions cliniques. Le système inclut une source lumineuse et des capteurs externes placés sur le cou. Une source lumineuse LED située sur le côté du cou illumine l'hypopharynx de manière diffuse, un photo-capteur sur la partie antérieure du cou, situé sous le cartilage cricoïde détecte la lumière qui passe à travers la glotte. Un circuit de rejet de la lumière ambiante a récemment été ajouté pour éviter l'effet de la lumière de la pièce. Le système du photoglottographe (ePGG) est libre de toute interférence due à la rétraction de la langue et est donc opérationnel pour les voyelles fermées et ouvertes, tandis qu'il est sensible aux mouvements articulatoires de la mâchoire et du larynx.

L'éclairage externe et le photoglottographe (ePGG) présentent donc trois traits qui ne sont pas partagés par les outils développés auparavant. L'appareil permet d'observer l'ouverture de la glotte d'une manière non invasive en parole continue pour tous les sons possibles.

De plus, le débit d'air peut être mesuré (en litre/s). Le débit est souvent mesuré en utilisant un masque rigide et un orifice dans lequel est placé un maillage en acier inoxydable qui donne une petite résistance au débit d'air. Ce type d'outil est un pneumotachographe à écran. Le principal inconvénient est que le masque rigide crée une cavité devant la bouche et les narines qui affecte le signal acoustique par l'addition d'une résonance. Le nouveau pneumotachographe développé au LPP utilise un masque fait de fibres synthétiques au lieu des masques rigides conventionnels. Ce masque est acoustiquement transparent et le son est donc propagé à travers le masque qui est quasiment libre de toute distorsion acoustique. Les mesures de débit d'air qui utilisent notre masque donnent des résultats comparables aux masques rigides mais avec des contraintes moindres.

Les images par résonnance magnétique (IRM) fournissent un excellent contraste des tissus mous d'une coupe placée dans une orientation quelconque et sans utilisation de radiation ionisante. L'IRM dynamique est reconnue comme un outil puissant pour l'imagerie de la parole [49] tout particulièrement pour observer les changements dynamiques dans les régions oro-pharyngienne et naso-pharyngienne. Elle permet de capturer les mouvements articulatoires pendant la production de la parole. Elle a aussi le potentiel de rendre visible les structures en tissu mou come la cavité du pharynx et la musculature interne, ainsi que la dynamique articulatoire sous forme d'une tomographie d'orientation arbitraire. Ces capacités de l'IRM ont déjà fait leurs preuves dans une variété d'études publiées tel que des études des mouvements articulatoires pendant la production vocale [57, 9, 7], des études sur la production du chant sous diverses formes [55, 44] et des travaux plus méthodologiques sur l'augmentation de la résolution spatio-temporelle par l'utilisation de contraintes de parcimonie [13].
Idéalement, une méthode d'imagerie dynamique de la parole doit avoir les trois propriétés suivantes.
Premièrement elle doit utiliser une résolution temporelle suffisamment rapide pour enregistrer la dynamique de la parole. Le but commun de bien des applications de l'imagerie de la parole est un enregistrement précis du processus de production de la parole, dans le quelle la position et la forme des articulateurs varient rapidement dans le temps. Il a par exemple été montré qu'une résolution temporelle de 30 à 50 images par seconde est nécessaire pour observer la dynamique de rétraction du voile du palais ou pour mettre en évidence les effets de coarticulation pendant la production de la parole [2, 37, 56].
Deuxièmement, la méthode d'imagerie doit offrir une haute résolution spatiale pour permettre la segmentation des structures fines des articulateurs, en particulier les détails fins de la pointe de la langue et de l'orifice vélopharyngien, par exemple, une résolution spatiale de 1.9 mm a été utilisée pour observer et modéliser le mouvement de la pointe de la langue [52].
Troisièmement, la méthode d'imagerie doit permettre une couverture complète de la zone du conduit vocal à explorer, par exemple plusieurs coupes ont été utilisées pour l'étude des fricatives anglaises [21]. Pour obtenir une couverture complète du conduit vocal, les techniques d'acquisition 3D ont récemment été utilisées avec une production d'un son continu pendant toute l'acquisition [46].
Bien qu'il soit possible de faire des compromis pour satisfaire chacune de ces propriétés indépendamment c'est nécessairement au détriment des autres propriétés. Cela reste un défi majeur de pouvoir proposer une acquisition d'IRM dynamique du conduit vocal pouvant satisfaire ces trois contraintes simultanément.
En utilisant les dernier progrès de l'IRM en particulier l'imagerie parallèle [45] et les reconstructions sous contraintes parcimonieuse « compress sensing » [28], le laboratoire IADI a développé des techniques IRM de reconstruction compensée en mouvement [38] et des reconstructions dynamiques multi coupes permettant l'application de reconstruction par super résolution. Ces techniques ont déjà été appliquées au mouvement cardio-respiratoire en utilisant les signaux physiologiques (ECG et respiration) comme contraintes pour les algorithmes de reconstruction. Des résultats préliminaires du transfert de ces techniques aux domaines de l'imagerie de la production vocale ont déjà fait l'objet d'une communication dans un atelier dédié à l'IRM pour l'étude des mouvements [59].

2.3. Positionnement aux niveaux national, européen et international

Il existe peu de tentatives de systèmes de synthèse articulatoire et celles qui existent n'apportent que des solutions partielles. Le système CASY [19] (Configurable Articulatory Synthesizer) développé au laboratoire Haskins (New Haven, USA) combine une simulation fréquentielle simpliste appliquée à une coupe médiosagittale du conduit vocal et un synthétiseur à formants qui génère le signal acoustique. Il ne s'agit donc pas à proprement parler de synthèse articulatoire, et en particulier aucune des interactions entre plis vocaux et conduit vocal ne peut être prise en compte. La solution adoptée par Brad Story (Univ. Arizona) est plus conforme du point de vue de la simulation des plis vocaux [53] mais le contrôle de la fonction d'aire est très artificiel puisque chaque constriction géométrique est modélisée sous la forme d'une gaussienne. Le système VocalTractLab développé par Peter Birkholz (Technische Universität Dresden) est sans doute ce qui se rapproche le plus de notre projet avec néanmoins plusieurs différences importantes sur la nature des phénomènes physiques, du modèle physique, du modèle articulatoire et de la coarticulation. La modélisation aéroacoustique est assez incomplète ce qui oblige à introduire des paramètres de contrôle ad-hoc. Par ailleurs, le modèle articulatoire est aussi assez artificiel puisqu'il repose sur des primitives géométriques (c'est en fait

une extension tridimensionnelle du modèle de Mermelstein [34]). L'adaptation du modèle à un locuteur se fait donc pour chaque image articulatoire et sans garantie sur la cohérence du modèle articulatoire global, ce qui rend l'évolution temporelle de la forme du conduit vocal, et donc la prise en compte de la coarticulation, assez difficile.
Il s'agit là des trois systèmes qui se rapprochent le plus de notre projet qui propose pour chacun des points des avancées décisives, et qui surtout propose une approche qui associe étroitement la prise en compte simultanée des meilleurs modèles de plis vocaux et du conduit vocal.

Ce projet ne pourrait pas avoir lieu sans l'acquisition de données sur la production de la parole et validation des modèles physiques. Sur ce dernier point il faut souligner que le Gipsa-Lab a développé des dispositifs expérimentaux probablement uniques au monde qui lui permettent maintenant de pouvoir tester des modèles physiques dans des conditions très proches de celles de la parole réelle.
Dans le domaine des acquisitions IRM de nombreux laboratoires ont acquis des données tridimensionnelles statiques, mais souvent elles ne sont ni très complètes, ni consacrées au français que nous étudierons en premier pour des raisons de disponibilités des sujets et de connaissance de la phonétique. Pour ce qui concerne les données dynamiques le laboratoire SAIL (Signal Analysis and Interpretation Lab) à USC (USA) est le laboratoire de référence pour l'acquisition de données en temps réel. Il existe par ailleurs de nombreuses tentatives destinées à dépasser la vitesse atteinte par le SAIL qui est légèrement inférieure à 25 Hz. L'expérience du laboratoire IADI, son expertise dans les algorithmes de reconstruction d'images sont un atout absolument essentiel.

Dans le domaine des données aérodynamiques enfin, de nombreux systèmes ont été développés, en particulier le système EVA au LPL à Aix-en-Provence. L'intérêt du système développé au LPP est de pouvoir utiliser un masque souple (du type masque de bricolage) qui ne crée pas de résonance parasite. De la même façon le système d'électro photoglottographie est non invasif ce qui est un avantage décisif puisqu'il peut être utilisé sans restriction. Enfin, nous disposons de données de pression sous-glottique sans doute uniques au monde.

3. PROGRAMME SCIENTIFIQUE ET TECHNIQUE, ORGANISATION DU PROJET

3.1. Programme scientifique et structuration du projet

Les partenaires de ArtSpeech ont déjà travaillé sur de nombreux aspects du projet et il existe donc un certain nombre de données, simulations et algorithmes disponibles. Il s'agit en particulier de données aérodynamiques et de la pression sous-glottique pour une dizaine de locuteurs, d'IRM statiques tridimensionnelles très complètes sur deux locuteurs, de fonctions d'aires dérivées d'une quinzaine de phrases issues de cinéradiographies, des algorithmes de simulation acoustique utilisées pour la synthèse par copie articulatoire issus des travaux de S. Maeda [31], [29], et des simulations des plis vocaux et des occlusives.
Ces données et algorithmes seront mis à la disposition du consortium de manière à amorcer les travaux, qu'il s'agisse de simulations qui disposeront donc de données de bonne qualité dès le début, ou inversement de l'exploitation des données recueillies à l'aide des simulations existantes.

3.2. Description des travaux par tâche

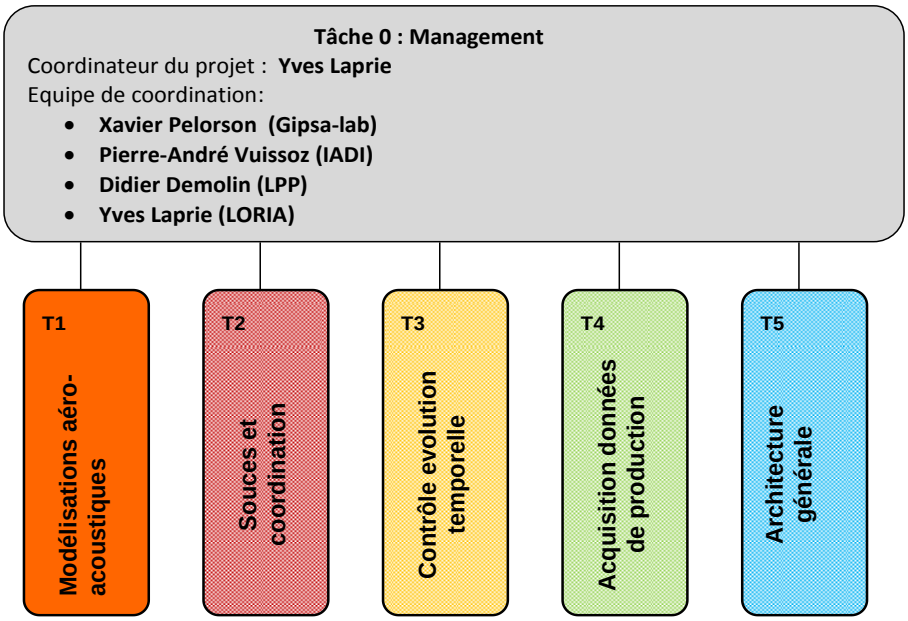
3.2.1. TÂCHE 0: COORDINATION

Synergie
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, forge…) que nous utiliserons, nous renforcerons la synergie grâce à des séjours de travail des deux doctorants et post-doctorants afin qu'ils maîtrisent parfaitement leur contribution et les interactions avec les autres aspects du projet, par exemple les liens qui existent entre contrôle

de la géométrie et acoustique, ou encore entre données articulatoires et simulation aérodynamique et acoustique. Par ailleurs, le doctorant travaillant au Gipsa-lab, respectivement au LORIA, sera co-encadré par un membre senior du projet au LORIA, respectivement au Gipsa-lab.

Equipe de coordination La coordination du projet sera assurée par l'équipe formée de Yves Laprie (LORIA), Xavier Pelorson (Gipsa-lab), Pierre-André Vuissoz (IADI) et Didier Demolin (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.



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 longue expérience des responsabilités collectives puisqu'il a animé l'équipe Parole (environ 30 personnes) entre les années 1997 et 2014, et plusieurs projets dont un projet européen et un projet ANR. Le service de gestion du LORIA assurera le suivi des coûts, des documents budgétaires et de justifications des dépenses.

Stratégie de communication L'objectif est que tous les partenaires du projet soient parfaitement informés de l'état d'avancement du projet, du programme de travail et tout autre point important pour assurer une parfaite synergie entre les partenaires. Le consortium organisera des réunions d'avancement tous les 4 mois, et des réunions techniques dès que cela sera nécessaire pour résoudre une difficulté technique. Afin de faciliter la coopération effective entre les partenaires les réunions physiques seront privilégiées.
L'idée générale de la gestion du projet est d'anticiper les problèmes par une bonne communication, et le cas échéant de réagir dès qu'une difficulté survient. Le coordinateur sera prévenu dans les meilleurs délais afin qu'il soit possible de trouver la solution la plus appropriée. Un site web sera aussi construit pour présenter le projet et ses résultats au grand public et aux scientifiques.

3.2.2. TÂCHE 1 : SIMULATIONS AÉRODYNAMIQUES ET ACOUSTIQUES

Responsable : Gipsa-Lab

Les partenaires de ArtSpeech ont déjà réalisé des simulations acoustiques et aérodynamiques du conduit vocal qui ont été validées soit par comparaison avec des mesures in-vitro sur des copies du conduit vocal, soit d'un point de vue perceptif sur des logatomes. Récemment, nous avons copié avec succès des signaux de parole à

partir de films aux rayons-x du conduit vocal enregistrés durant la production de la parole. Les images du film ont été exploitées pour estimer la géométrie du conduit vocal à chaque instant et en déduire les fonctions d'aire qui ont été utilisées en entrée des simulations aérodynamique et acoustique.
Le travail concernera la simulation acoustique de manière à améliorer la qualité acoustique tout en garantissant une mise en œuvre efficace et proche du temps réel. L'amélioration du réalisme du modèle géométrique du conduit vocal n'a de sens que si elle s'accompagne d'un effort similaire sur la modélisation des sources sonores. Cette modélisation physique visera en premier lieu à développer des théories d'une complexité croissante pour expliquer et prédire les phénomènes aéro-acoustiques liés à la production de la parole. La pertinence et la précision de ces modèles théoriques seront ensuite systématiquement confrontées à des mesures réalisées sur des maquettes physiques du conduit vocal. Une fois validés ces modèles théoriques seront mis en œuvre dans le synthétiseur articulatoire.
Contrairement aux simulations numériques directes cette approche n'altère pas l'efficacité computationnelle du synthétiseur. Des recherches récentes sur des sons voisés et des occlusives ont montré qu'une telle approche conduit à une réduction significative du nombre de paramètres et commandes ad-hoc nécessaires, et donc à un gain d'efficacité.

Objectifs
Fournir une modélisation des sources de son et de leurs interactions précise et adaptée à la synthèse (Tâche 5).
Défis
Le premier défi est d'ordre théorique puisqu'il s'agit de proposer et de valider des modèles physiques des sources de sons de parole, notamment des fricatives et plosives, ainsi que de leurs interactions. Le second défi est expérimental, puisqu'il s'agit de recréer en laboratoire des conditions géométriques, acoustiques et aérodynamiques comparables aux données mesurées in-vivo.
Indicateurs de succès
Deux indicateurs de succès des modèles physiques : - vis-à-vis de des mesures sur maquettes, - par comparaison avec les données mesurées in-vivo
Livrables
Modèles théoriques de sources Maquettes permettant de simuler l'aérodynamique de séquence voyelle-plosive ou voyelle-fricative.
Programme détaillé
Ce travail s'appuie sur les données expérimentales fournies par la tâche 4. Ces données permettent de valider un certain nombre d'hypothèses théoriques, de définir les paramètres de contrôle des maquettes et serviront de référence pour l'évaluation finale de la modélisation. 1. **Modèle de source vocale**. Un modèle de source vocale déjà existant sera fourni à la tâche 5 afin d'avoir dès le démarrage du projet une base commune de travail. Ce modèle pourra ensuite être enrichi, si nécessaire, par une description plus fine de la mécanique des plis vocaux (modélisation de la collision, en particulier) et de l'aérodynamique (modélisation du souffle, du chuchotement …). Le couplage acoustique avec les cavités sous- et supra-glottiques sera réalisé dans un premier temps au moyen d'une approximation unidimensionnelle (ondes planes). Une série limitée de mesures sur des impressions 3D de conduits vocaux sera réalisée afin de tester l'approximation unidimensionnelle en particulier en présence de cavités (cf. tâche 4) ainsi qu'aux fréquences élevées. Selon les résultats de cette étude une modélisation bi- ou tri-dimensionnelle pourra être proposée (méthode modale). **2. Modèle de plosives et interaction avec la source de voisement.** Il s'agira d'étendre les résultats d'une première étude limitée aux plosives bilabiales à l'ensemble des plosives du français. L'interaction aérodynamique entre les plis vocaux et le lieu de la plosive fera l'objet d'une attention particulière du point de vue de la dynamique et de ses conséquences sur le VOT. Une maquette comprenant des plis vocaux auto-oscillants couplés à une constriction motorisée permettra de reproduire en laboratoire différentes configurations typiques des plosives du français. La mesure de la pression supra-glottique, intra-orale et acoustique permettra de tester les modèles théoriques.

3. Modèle de fricatives.

Le modèle de source turbulente sera basé sur des analogies aéroacoustiques. Une analyse théorique et expérimentale sera menée afin d'étudier l'influence de paramètres géométriques (conditions en amont de la constriction), aérodynamiques (interaction d'un jet turbulent avec les dents ou les parois du conduit vocal) et acoustique (effet des modes non plans sur les sources). L'étude expérimentale sera réalisée sur une maquette simplifiée de la cavité buccale inspirée des données anatomiques fournies par la tâche 4. De la même manière que pour les plosives, la coordination avec la source de voisement sera étudiée.

Choix technologiques

La validation des modèles théoriques sera réalisée sur les bancs expérimentaux existant au Gipsa-lab.

Contributions (Qui fait quoi)

Xavier Pelorson : acoustique, sons voisés et plosives
Annemie Van Hirtum : fricatives
Un étudiant en thèse : modélisation complète.
Xavier Laval : support technique pour les maquettes

Risques

La complexité de la modélisation, en particulier pour les fricatives, peut nuire à l'efficacité du synthétiseur en termes de temps de calculs.

Alternatives

Adopter un compromis entre la précision de la modélisation et le coût numérique.

3.2.3. TACHE 2: SOURCE ET SCENARII DE COORDINATION

Responsable : LPP

Objectifs

L'objectif est de concevoir les scenarii de coordination entre les sources et les articulateurs du conduit vocal afin de produire des consonnes ou groupes de consonnes et les indices acoustiques rendant possible leur identification par des auditeurs humains.
Il faut souligner que l'impact acoustique des erreurs de coordination peut être très fort. L'objectif est d'atteindre une meilleure robustesse de la coordination temporelle.

Défis

Le défi est de piloter la coordination entre le contrôle des plis vocaux et les déformations rapides du conduit vocal qui accompagnent la production des consonnes avec des paramètres en petit nombre, pertinents du point de vue physique et en produisant les indices acoustiques attendus.

Indicateurs de succès

Réalisation des bons indices acoustiques sur les bruits d'explosion, le voice onset time (VOT), les bruits de friction, les transitions formantiques…

Programme détaillé

1. Recherche des paramètres de contrôle des modèles physiques

La recherche des paramètres de contrôle des modèles physiques développés dans la tâche précédente est destinée à l'élaboration des scenarii de coordination. Quand la modélisation aéro-acoustique est insuffisamment fondée du point de vue physique, on risque de faire appel à un trop grand nombre de paramètres, qui de plus ne sont pas accessibles à un être humain. Ce travail a déjà été fait pour les plosives labiales et doit être étendu aux autres points d'articulation des occlusives.
L'articulation des fricatives demande une plus grande précision afin de contrôler la turbulence à l'origine du bruit de friction et il est vraisemblable que le contrôle locuteur soit plus important. L'utilisation des simulations aéroacoustiques permettra de mettre à jour ces paramètres.

2. Intégration des mesures aérodynamiques et des plis vocaux dans les modèles physiques

Dans un second temps il s'agira d'intégrer les mesures aérodynamiques et des plis vocaux)acquises sur de la parole réelle dans les simulations aéro-acoustiques afin d'accroître nos connaissances sur les aspects dynamiques de la production de la parole comme la coordination entre la source glottale et les autres sources sonores d'origine aérodynamique (plosives et fricatives) qui est rarement étudiée théoriquement et

expérimentalement. Ce travail nécessite le dépouillement des mesures (aérodynamiques et plis vocaux), la préparation des fonctions d'aire obtenues soit directement par IRM 2D dynamiques complétées pour récupérer la 3ème dimension, soit construites à partir d'IRM statiques 3D et interpolation temporelle. Ensuite, il sera aussi nécessaire d'adapter certaines simulations et d'interpréter les résultats notamment afin d'étudier l'importance des contraintes aérodynamiques et leur interactions avec la glotte et la position des articulateurs.

3. Construction des scenarii de coordination

Les scenarii concernent toutes les consonnes ou groupes de consonnes, et plus précisément les phases de transition de la voyelle qui précède à la consonne, et ensuite la seconde transition vers la voyelle de la syllabe suivante. Ils seront utilisés par la simulation aéro-acoustique et donneront donc l'évolution temporelle de la fonction d'aire dérivée du modèle articulatoire à une échelle de temps très fine (de l'ordre de la milliseconde). Ces scenarii sont destinés à piloter les paramètres de contrôle de ces transitions qui auront été mis en évidence dans les sous-tâches précédentes. En plus des travaux présentés plus haut ils seront élaborés sur la base de ceux développés au LPP par S. Maeda [31].

Livrables
Scenarii de contrôle utilisables par les simulations aéro-acoustiques.
Contributions, qui fait quoi ?
Gipsa-lab pour la sous-tâche 1, LPP pour la sous-tache 2, LPP et LORIA pour la sous-tâche 3.
Risques
Les risques sont de ne pas obtenir les indices acoustiques observés dans la parole naturelle ou de faire appel à des paramètres qui ne sont pas pertinents.
Alternatives
Une solution consiste à utiliser des scenarii qui produisent les bons indices acoustiques quitte à faire appel à des paramètres de contrôle moins bien fondés du point de vue physique.

3.2.4. TACHE 3: CONTROLE DE L'EVOLUTION TEMPORELLE DE LA GEOMETRIE DU CONDUIT VOCAL.

Responsable : LORIA

Objectifs
L'objectif consiste à fournir aux simulations aérodynamique et acoustique la fonction d'aire du conduit vocal, c'est-à-dire l'aire transverse du conduit vocal depuis la glotte jusqu'aux lèvres. Les travaux que nous proposons s'appuient sur l'état de l'art présenté au paragraphe 2.2.2 Nos travaux porteront sur : 1. L'élaboration de modèles articulatoires afin de contrôler la forme bi ou tridimensionnelle du conduit vocal. Ces modèles devront à la fois être précis géométriquement pour toutes les classes de sons et suffisamment concis pour être facilement utilisables. 2. Le passage à la fonction d'aire à partir d'un modèle bi ou tridimensionnel. Il faut noter que même dans le cas d'un modèle tridimensionnel le passage à la fonction d'aire n'est pas immédiat. 3. L'élaboration d'un modèle numérique de coarticulation.
Défis
Les défis sont de pouvoir décrire la géométrie qui impacte les propriétés acoustiques du conduit vocal précisément dans l'espace et dans le temps, et de pouvoir adapter la description géométrique à un nouveau locuteur sans avoir à reprendre tout le processus d'acquisition d'images médicales.
Indicateurs de succès
Le premier indicateur de succès est **géométrique** et consiste à mesurer la distance entre la position des articulateurs prédite par les modèle articulatoire et de coarticulation et les positions mesurées à partir de données de parole, soit des ciné IRM acquises dans le cadre de ce projet, soit de données d'articulographie acquises au LORIA.

Le second indicateur est **acoustique** et consiste à évaluer la qualité de la parole produite par rapport à des signaux de parole naturelle. L'évaluation pourrait être faite au niveau spectral mais il est plus pertinent de la faire au niveau des formants (les pics spectraux correspondant aux fréquences de résonance du conduit vocal) et de leur évolution dans le temps.

Programme détaillé

1. Construction de modèles articulatoires

L'objectif est de développer un modèle tridimensionnel à partir d'images IRM tridimensionnelles statiques et de cinéIRM médiosagittales. Les IRM tridimensionnelles ne peuvent pas être acquises en temps réel mais elles fournissent la géométrie complète du conduit vocal, et inversement les IRM dynamiques capturent les gestes réels de la parole, et permettent de déterminer les modes de déformation du conduit vocal lors de la production de la parole. Les modes de déformation tridimensionnels dynamiques seront obtenus par interpolation linéaire à partir des modes tridimensionnels statiques en identifiant la coupe médiosagittale dynamique à partir des coupes médiosagittales statiques. Nous prévoyons d'utiliser un peu plus d'une centaine d'IRM statiques afin de couvrir les voyelles et des articulations de consonnes et de clusters consonantiques dans plusieurs contextes vocaliques.

L'un des points clés est que le modèle soit capable de reproduire avec précision toutes les déformations de la langue, y compris le geste rétroflexe notamment, et les contacts entre la langue et le palais qui apparaissent lors d'une occlusive. Pour ce dernier point les collisions entre la langue et le palais seront prises en compte. Il faut noter que nous avons déjà une bonne expérience en ce domaine avec les modèles articulatoires bidimensionnels que nous avons développés [25] récemment.

2. Passage à la fonction d'aire et topologie du modèle

Le passage à la fonction d'aire nécessite de déterminer la ligne centrale du conduit vocal et d'identifier toutes les cavités secondaires (sinus piriformes, cavité nasale et port vélopharyngé, cavité sous-linguale, cavités latérales lors de la production du son /l/) qui influencent l'acoustique. La difficulté est d'obtenir une topologie (apparition ou disparition de cavités) et des points de branchement avec les cavités pharyngale et buccale qui évoluent continument au cours du temps.

3. Adaptation de modèles articulatoires à un locuteur

L'adaptation d'un modèle articulatoire est essentielle dans la perspective de pouvoir adapter la synthèse à un locuteur quelconque sans avoir besoin d'acquérir un grand nombre d'IRM. Si les modes de déformation généraux et les principales propriétés acoustiques sont communes à tous les locuteurs (par exemple l'ouverture de la mâchoire qui provoque un accroissement de la première fréquence de résonance) les propriétés plus fines dépendent de la taille des cavités buccale, pharyngale, nasale, de la forme du palais, et des modes de déformation de la langue et des autres articulateurs.

Nous acquerrons des données géométriques statiques et dynamiques pour une dizaine de locuteurs afin de pouvoir élaborer une approche de normalisation, soit à l'aide d'analyse de données appliquées aux différentes géométries du conduit vocal, soit en utilisant des repères anatomiques facilement discernables sur les images IRM.

Il s'agira de relier les caractéristiques acoustiques générales d'un locuteur aux caractéristiques géométriques.

L'adaptation pourra être réalisée à partir de trois types de données : (1) propriétés acoustiques seules, (2) images ou scan 3D du visage, (3) une IRM. Plusieurs types de données peuvent être utilisés conjointement.

4. Algorithme de coarticulation

La phonologie articulatoire et la dynamique des tâches fournissent un cadre théorique intéressant mais dont la mise en œuvre se heurte souvent au trop grand nombre de paramètres à régler ce qui conduit à des hypothèses simplificatrices trop fortes, par exemple sur la durée des gestes [36]. L'activation des articulateurs exploitera les frontières syllabiques, lexicales ou syntagmatiques et la présence d'articulateurs critiques afin de contraindre la coordination articulatoire. Nous utiliserons au mieux les frontières syllabiques afin de limiter le nombre de degrés de liberté et de paramètres à apprendre. Pour cela nous acquerrons des IRM statiques de voyelles et de consonnes ou groupes de consonnes dans différents contextes vocaliques (les mêmes que celles utilisées pour déterminer les modes de déformations du conduit vocal).

Comme il n'est pas possible d'acquérir les consonnes et groupes de consonnes dans tous les contextes vocaliques, les formes manquantes seront calculées par interpolation à partir des voyelles connues [4]. L'optimisation des paramètres se fera en optimisant l'évolution temporelle de la forme du conduit vocal et les paramètres acoustiques de la parole synthétique.
Livrables
Modèle articulatoire Procédure d'adaptation du modèle articulatoire Algorithme de coarticulation
Choix technologiques
Les algorithmes seront développés en JAVA, les modèles seront sauvés sous une forme qui permettra de les utiliser indépendamment.
Contribution (qui fait quoi ?)
LORIA
Etique et réglementation
Cf. tâche 4 pour l'acquisition des IRM
Risques et alternatives
Les risques portent essentiellement sur l'adaptation du modèle articulatoire et l'algorithme de coarticulation. Pour l'adaptation il est possible de revenir à une stratégie simple en modifiant les tailles des cavités pharyngale et buccale. Pour la coarticulation, nous viserons d'abord la réalisation des indices acoustiques nécessaires à l'identification des sons, et notamment du lieu d'articulation des consonnes, avant de viser à la réalisation de détails plus fins.

3.2.5. TACHE 4: ACQUISITION DE DONNEES DE LA PRODUCTION DE LA PAROLE

Responsable : IADI

Ce projet n'est pas imaginable sans l'acquisition de données utilisées à deux niveaux : 1) celui de la source pour réaliser une synchronisation correcte de la source et du contrôle du conduit vocal, 2) celui de la déformation du conduit vocal pour obtenir une coordination correcte entre les articulateurs de la parole.
L'acquisition de mesures dynamiques à la glotte et dans le conduit vocal est un point clé puisque ces données serviront à concevoir et valider les modèles et simulations numériques. C'est en soi un sujet de recherche. Il n'est pas possible de réaliser simultanément des mesures à la glotte et dans le conduit vocal à cause de contraintes techniques liées aux technologies d'imagerie utilisées, par exemple le fait que l'Imagerie par Résonance Magnétique utilisée pour acquérir la géométrie du conduit vocal exclut l'utilisation d'un dispositif contenant des matériaux ferromagnétiques. Une seconde raison pratique est que cela imposerait aussi des contraintes trop fortes au sujet.

Nous prévoyons d'utiliser:

- l'IRM dynamique pour mesurer les mouvements des articulateurs pendant la production de la parole,
- l'IRM statique pour déterminer la géométrie tridimensionnelle du conduit vocal qui sera utilisée pour construire les modèles géométriques tridimensionnels,
- l'électrophotoglottographie pour estimer le degré d'ouverture de la glotte. Il s'agit d'une technique non invasive basée sur la détection de la lumière par un capteur photosensible collé sur le cou du sujet au niveau de la glotte,
- l'électroglottographie pour suivre le contact des plis vocaux en mesurant l'impédance entre des électrodes collées là encore sur le cou du sujet au niveau de la glotte,
- la pneumotachographie ou une technique similaire pour mesurer le flux d'air ou autres paramètres aérodynamiques de la production de la parole.

Les données seront acquises d'abord sur le français mais d'autres langues seront utilisées pour des études ponctuelles.

Objectifs
Méthodes d'acquisitions et corpus de données de production de la parole.
Défis
Produire des données de production de la parole le moins invasivement possible tout en garantissant une résolution spatiale et temporelle de haute qualité.
Indicateurs de succès
Un corpus de données de production de la parole permettant de contraindre le système de synthèse articulatoire.
Programme détaillé

1. Acquisition de données IRM statiques et dynamiques

1. Détermination du cahier des charges précis du corpus de données de production de la parole.
2. Développement des méthodes d'acquisition et de reconstruction IRM nécessaire à la production du corpus
3. Création d'un protocole de recherche clinique et des accords éthiques attenants pour l'acquisition des données.
4. Développement d'un outil d'intégration en post traitement de la segmentation et des données dynamiques permettant d'associer ce corpus au système de synthèse articulatoire.
5. Acquisition du corpus de données de production de parole.
6. Post traitement des données, vérification des spécifications en termes de résolution dynamique, publications.

Les acquisitions IRM se décomposent en trois volets distincts, tout d'abord sur 12 sujets sains deux types d'examens seront réalisés indépendamment :

Premièrement une série d'acquisitions 3D haute résolution pour des positions vocales statiques pendant une quinzaine de secondes. Ces données permettront de produire une bonne segmentation 3D pour les positions choisies (un développement informatique spécifique permettra de faciliter la segmentation semi-automatique réalisée par un expert). Ces IRM statiques seront utilisées pour déterminer la géométrie tridimensionnelle du conduit vocal qui sera ensuite utilisée pour construire les modèles géométriques tridimensionnels. Même si cette technique est souvent utilisée pour étudier la parole il reste à améliorer les protocoles d'acquisition pour obtenir des images de meilleure qualité et une meilleure couverture phonétique et articulatoire.

Deuxièmement sur les mêmes sujets une série de coupes 2D sur orientations choisies (voir Figure 2 « Il zappe pas mal ») et avec un ensemble de petites phrases permettant d'échantillonner la variabilité phonatoire de la langue française seront réalisées. Ces IRM dynamiques seront utilisées pour mesurer les mouvements des articulateurs pendant la production de la parole. L'IRM présente l'avantage déterminant de couvrir tout le conduit vocal depuis la glotte jusqu'aux lèvres. L'imagerie dynamique permettra d'acquérir la coupe médio-sagittale du conduit vocal. L'objectif est d'atteindre une résolution nettement meilleure que les travaux actuels à une cadence d'acquisition un peu supérieure à 25 Hz. Ces acquisitions utiliseront de plusieurs répétitions de la même phrase et un enregistrement du son produit (L'antenne IRM neurovasculaire et le microphone optique nécessaires ont déjà été obtenus par d'autres moyens de financement). Cet enregistrement permettra la définition d'une position relative au sein de la phrase et la reconstruction à haute résolution spatiale et temporelle grâce à l'algorithme GRICS l'algorithme GRICS [38, 58, 59]. Ces données 2D permettront de fournir des contraintes sur la dynamique de mouvement lors de la prononciation de phrases types permettant de calibrer les modèles de production vocale des tâches 1 et 2.

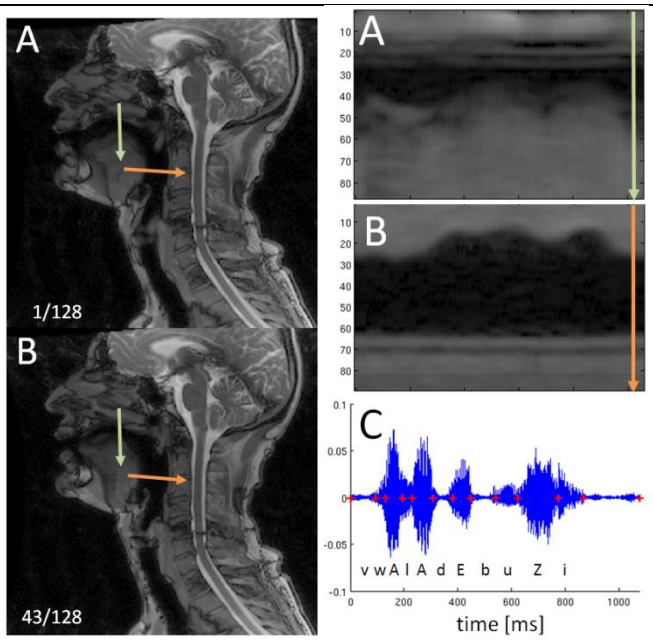


Figure 1 : Acquisition dynamique 2D IRM sagittale de la phrase « Voilà des bougies »

Le troisième volet présente la recherche la plus risquée, il va s'agir de mettre au point une acquisition temps réel à très haute vitesse pour essayer d'atteindre les 200 Hz (probablement uniquement sur une projection 1D, voir Figure 2 coupe frontale 1D) mais qui permettra de donner des contraintes supplémentaires sur les événements les plus rapides tels que les plosives ou les fricatives et destinées à la tâche 1. Pour ce faire une série d'acquisitions sur des objets de test ainsi que trois séances sur volontaires sont prévues, elles permettront de réaliser une preuve de concept de la méthode.

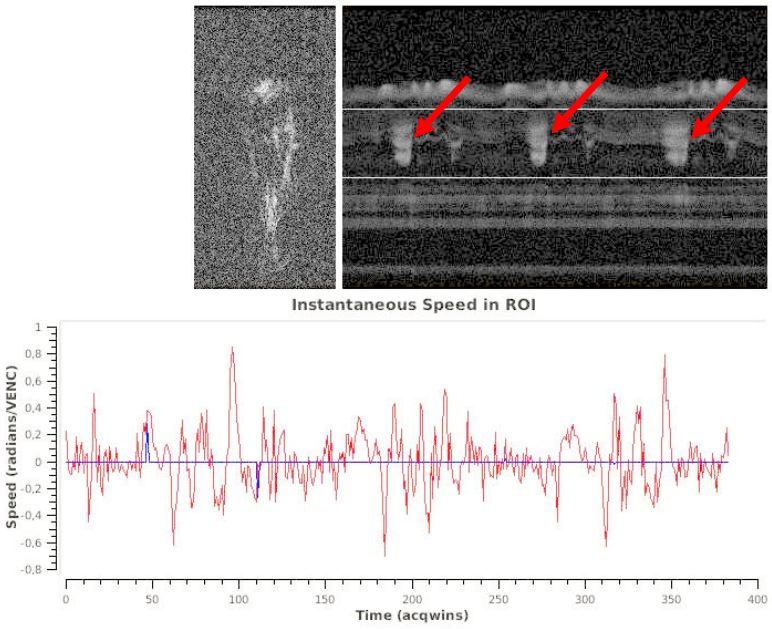


Figure 2 : Projection 1D d'une coupe frontale postérieur aux incisives, la fréquence d'acquisition des projection est d'environ 150 Hz. La région sélectionnée correspond à la cavité buccale et la variation de la vitesse dans cette région est représentée dans le graphique du bas lors de la triple répétition d'un mot. Les variations de vitesse provoquées par le mouvement de la langue sont visibles. Les trois flèches rouges indiquent le contact entre la langue et le palais lors de la production du /l/ de « il zappe pas mal ».

2. Acquisition de données aérodynamiques et sur les plis vocaux

Les données concerneront : les paramètres aérodynamiques (débit oral et nasal), les mouvements des plis vocaux, la pression intraorale et la pression sous-glottique. Les deux premiers types de données seront acquis à l'aide du photoglottographe et pneumotachographe développés récemment par le LPP [16] et dont l'avantage est de ne pas perturber la production de la parole. Par ailleurs, il sera aussi possible d'utiliser un photoglottographe traditionnel qui donne des images des plis vocaux mais dont le désavantage est d'être invasif. Ces dernières acquisitions se feront bien sûr dans un cadre médical à l'Hôpital Européen Georges Pompidou. La pression intra orale sera mesurée, soit en plaçant un petit tube entre les lèvres des sujets, soit en plaçant un tube derrière le voile du palais en passant par les narines. La première solution présente l'intérêt d'être parfaitement compatible avec l'utilisation du photoglottographe. Enfin, pour mesurer la pression sous-glottique nous envisageons le développement d'un nouveau plétismographe pour mesurer de manière indirecte la pression sous-glottique. Ce nouveau plétismographe sera développé à partir d'une combinaison étanche et sera donc aussi totalement non invasif. La sensibilité des nouveaux capteurs de pression permet d'envisager des mesures qui permettront de comparer les mesures avec celles obtenues par ponction trachéale directe. Ces données proviennent d'une base de données aérodynamique sur la parole dont nous disposons (projet antérieur de Didier Demolin). Cette base données sera formatée de manière à la rendre accessible à la communauté. Les données qui seront enregistrées porteront sur le français et des langues qui contrastent les consonnes aspirées et non aspirées. Le corpus sera établi pour observer un maximum d'interactions entre la glotte et le conduit vocal (aspiration, friction, interaction entre les deux modes).
Livrables
Publications décrivant une méthode d'acquisition et base de données de production de la parole sur une dizaine de sujets.
Choix technologiques
IRM dynamique 2D et IRM statique 3D Acquisition IRM temps réel. Photoglottographie, Electroglottographie, Pneumotachographie et Plétismographie.
Contribution (qui fait quoi ?)
Le laboratoire IADI développera le protocole d'acquisition de haute résolution spatiale et temporelle en IRM. Le laboratoire LPP se consacrera aux acquisitions de données aérodynamiques ou concernant la coordination entre les plis vocaux et les articulateurs du conduit vocal.
Etique et réglementation
Un protocole de recherche sur l'homme est soumis au Comité de Protection des Patients, la gestion du protocole et du recrutement sera sous-traité au centre d'investigation clinique INSERM CIC-IT 1433 du CHU de Nancy.
Risques et alternatives
Le temps d'acquisition nécessaire pour assurer la qualité des données limite l'applicabilité de la méthode à un grand nombre de sujets. Séquence clinique disponible en standard sur Imageur IRM avec plus faible résolution spatiale et temporelle.

3.2.6. TÂCHE 5: ARCHITECTURE GÉNÉRALE

Responsable : LORIA

Objectifs
Le premier objectif de ArtSpeech est la synthèse d'un signal de parole à partir de la connaissance des phonèmes. À côté de ce projet nous avons déjà développé une infrastructure complète de la synthèse à partir du texte du français qui sera utilisée pour générer une séquence de phonèmes à partir du texte.
Défis

Le défi est de nature essentiellement technique et consiste à interfacer les différents niveaux de représentation (langage naturel, sources et expressions, fonctions d'aire) utilisés pour la synthèse.
Indicateurs de succès
Possibilité de synthétiser de la parole à partir du texte et de contrôler les paramètres prosodiques simples.
Livrables
Système de synthèse à partir du texte d'une phrase.
Programme détaillé
Nous utiliserons le système de synthèse par concaténation développé au LORIA pour générer la suite de phonèmes à synthétiser. Comme la quasi-totalité des systèmes actuels, il n'utilise pas de module prosodique spécifique. Les informations prosodiques propres aux segments de parole du corpus sont stockées directement dans le corpus et contribuent avec les autres (acoustique, frontières de mots ou groupes de mots) au calcul du coût global. L'algorithme de concaténation recherche le meilleur parcours prosodique en fonction de la phrase à produire. **1. Adaptation des paramètres de source pour la synthèse articulatoire** Le travail portera sur l'adaptation des paramètres d'entrée de la synthèse articulatoire. Les paramètres de durée, de fréquence fondamentale et de d'intensité sont des sous-produits de la synthèse par concaténation. Il faudra en faire des paramètres à part entière, et surtout ajouter les paramètres de contrôle des plis vocaux. Dans un premier temps l'objectif est d'utiliser la même prosodie que celle produite par le système de synthèse par concaténation. Par la suite, ces paramètres pourront être complétés par le codage des expressions. Les informations sur les plis vocaux seront stockées au format XML pour assurer une interopérabilité simple. 2. **Intégration logicielle** Les informations de synthèse seront organisées en 3 niveaux : - Une suite de phonèmes complétés par la syllabification, et les frontières de mots et de syntagmes. Ce premier niveau de description est construit par le système de synthèse par concaténation qui applique des algorithmes d'analyse du langage naturel et l'étape de concaténation proprement dite à partir du corpus de parole enregistré. - Le fichier des paramètres de source dont la spécification du contenu fait l'objet de la sous-tâche décrite au-dessus. - Une suite de fonctions d'aire à chaque nœud temporel de synthèse. L'algorithme de coarticulation génère une suite de descriptions géométriques du conduit vocal sous la forme des cavités (sinus piriformes, cavité pharyngale, cavité nasale, cavités latérales, cavité sous-linguale), des liens de connexion qui existent entre elles et la nature de la transition avec le nœud suivant. Chaque fonction d'aire est donnée sous la forme d'une suite de couples (longueur et aire du tube). Une description du conduit vocal est donnée à tous les instants correspondant à une nouvelle spécification de la fonction d'aire, et la nature de la transition avec la description suivante. La nature de la transition dépend de la classe du son à produire (occlusive, fricative, voyelle…). Sa mise en œuvre correspond à l'un des scenarii qui font l'objet de la tâche 2. La simulation numérique de l'aérodynamique et de l'acoustique utilise les deux derniers niveaux, le premier étant utilisé par l'algorithme de coarticulation.
Choix technologiques
Chacun des modules sera disponible sous la forme d'une librairie C++, ou Java. Dans un premier temps l'objectif n'est pas d'atteindre le temps réel mais de pouvoir générer le signal acoustique à partir du texte.
Contribution (Qui fait quoi)
Le LORIA sera responsable de l'intégration.
Risques
Pas de risque particulier.

3.2.7. CONSORTIUM

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 :

- la simulation aérodynamique et acoustique de la production de la parole et la modélisation de la source et du conduit vocal,
- l'imagerie par résonance magnétique et les autres techniques d'acquisition de données de parole.

Le consortium sera coordonné par **Yves Laprie (DR CNRS)** et il sera formé de:

Equipe MultSpeech du LORIA (Nancy)

Yves Laprie (DR CNRS), Vincent Colotte (MCf, Univ. Lorraine) spécialiste en synthèse de la parole et **Slim Ouni (MCf, Univ. Lorraine)** spécialiste en parole multimodale participeront à ce projet.

CV de Yves Laprie 54 ans Directeur de Recherche au CNRS (ISHS)

Ses domaines de recherche sont l'analyse de la parole, la modélisation articulatoire de la parole et l'inversion acoustique articulatoire. Il a en particulier développé plusieurs algorithmes de suivi de formants qui ont été intégrés dans le logiciel d'analyse de la parole WinSnoori qui offre une grande variété d'algorithmes d'analyse de la parole dont notamment une approche complète de la synthèse par copie pour le synthétiseur de Klatt. Ces algorithmes sont aussi utilisés dans le cadre du développement de logiciels pour l'apprentissage des langues. Dans le domaine de l'inversion acoustique articulatoire il s'est intéressé aux approches d'analyse par synthèse et a proposé des algorithmes garantissant à la fois une très bonne proximité avec les données acoustiques et des trajectoires articulatoires réalistes. Enfin il travaille sur la modélisation articulatoire afin de pourvoir produire de la parole par synthèse articulatoire. Les modèles articulatoires les plus récents ont permis de resynthétiser de la parole de bonne qualité à partir de films aux rayons X du conduit vocal.

Il a été responsable de l'équipe Parole du laboratoire LORIA entre 1997 et 2014 et a été impliqué dans de nombreux projets de recherche, notamment les projets européen FET ASPI (Audiovisual Speech Inversion 2005-2008) et ANR ARTIS (Inversion articulatoire de la parole audiovisuelle pour la parole Augmentée – 2009-2012) pour lesquels il a été coordinateur scientifique.

Equipe d'Acoustique du Gipsa-Lab (Grenoble) **Xavier Pelorson (DR CNRS), Annemie Van Hirtum** (CR1 CNRS) et **Xavier Laval** (IE G-INP) participeront au projet**.**

Xavier Pelorson est spécialiste de la modélisation physique de la production de la parole basée sur une approche aéroacoustique visant à identifier puis à décrire les mécanismes de production de son. Ses travaux théoriques ont porté sur la modélisation de la phonation, la propagation acoustique tri-dimensionnelle et ont été étendus plus récemment à l'étude des plosives et des fricatives en collaboration avec Annemie Van Hirtum, partenaire du projet. Parallèlement, un important travail expérimental est mené afin d'évaluer la pertinence et la précision des modèles théoriques. Celui-ci repose sur la conception puis la réalisation de maquettes mécaniques du conduit vocal, rigides ou déformables, sur lesquelles des conditions d'écoulement et de déformation comparables à celle observées in-vivo peuvent être reproduites, mesurées et contrôlées avec une grande précision. Le Gipsa-lab bénéficie d'un environnement expérimental unique au monde comprenant quatre salles expérimentales, un atelier mécanique et une imprimante 3D haute résolution.

Xavier Pelorson a été responsable de l'équipe Acoustique du Gipsa-lab de 2009 à 2014, Co-Editeur en chef de la revue Acta Acustica united with Acustica (1998-2003) et impliqué dans de nombreux projets Nationaux et Internationaux (ANR, EU-FET…) ainsi que responsable d'un partenariat industriel**.**

Laboratoire IADI (Imagerie Adaptative Diagnostique et Interventionnelle, unité INSERM U947) (Nancy)

Créée par le Pr. Jacques Felblinger l'unité INSERM U947 de l'Université de Lorraine possède une expertise dans la gestion de mouvement en Imagerie par Résonance Magnétique (IRM). La technologie GRICS (1)[1] destinée à corriger les mouvements du patient a été brevetée (4), de plus des applications en imagerie cardiaque (2) sont en cours de développement avec General Electric dans le cadre du projet européen BERTI (FP7). Le IADI a développé une expertise en dispositifs médicaux (DM), avec Schiller Médical. Le laboratoire a créé l'entreprise Heltis pour la compatibilité IRM des DM et entretient une étroite collaboration avec le CIC-IT de

[1] Les numéros se rapportent aux références bibliographiques du laboratoire IADI citées à la fin du document.

Nancy, qui organise les études cliniques. Ces locaux sont situés près de l'imageur 3T du CHU sur lequel seront réalisés les IRM de ArtSpeech. Le responsable ce projet pour IADI, sera **Pierre-André Vuissoz**.
CV de Pierre-André Vuissoz (Dr.) Directeur adjoint du laboratoire IADI depuis 2013. Ingénieur en physique théorique de l'Ecole Polytechnique Fédérale de Zurich, il a un doctorat en Résonance Magnétique Nucléaire électrochimique de L'Ecole Polytechnique Fédérale de Lausanne.
Depuis 2004, il est Ingénieur de Recherche au IADI est développe des séquences et algorithmes d'IRM adaptative. Monsieur Vuissoz a encadré 8 thèses, il est coauteur de 21 publications et co-inventeur de 3 brevets. Il a participé à des recherches multicentriques d'IRM urologique (3). Il sera assisté de **Dr. Freddy Odille**, ingénieur ENSEM, chargé de recherche INSERM au IADI, Honorary Researcher at UCL and KCL, London, UK, coauteur de 14 publications et coinventeur de 2 brevets. Le projet ArtSpeech fait suite à la collaboration déjà fructueuse des deux chercheurs avec Dr. Yves Laprie (5).

Laboratoire LPP (Laboratoire de Phonétique et de Phonologie) (Paris). **Angélique Amelot (IR CNRS)** travaille sur le développement d'instrumentations destinées à enregistrer des données acoustiques aérodynamiques et articulatoires de la parole. Avec **Shinji Maeda** (DR CNRS émérite) et **Didier Demolin** (Prof. Université Sorbonne Nouvelle) elle travaillera elle travaillera sur l'acquisition de données acoustiques aérodynamiques et physiologiques concernant en particulier les plis vocaux.
CV de Didier Demolin: Professeur (PR1), Université Sorbonne nouvelle, Paris 3 (Phonétique expérimentale, physiologique et acoustique), laboratoire de Phonétique et Phonologie (LPP) UMR 7018; Prix de l'International Phonetic Association, San Francisco, 1999 ; Chaire Francqui, (2009-2010) ; Prix Freeman Université d'Amherst, Massachusetts (2010) ; membre de l'Académie Royale des Sciences d'outre-mer de Belgique et de l'Academia Europaea 'The European Academy'. Principaux axes de recherche: diversité, dynamique et complexité des systèmes sonores des langues humaines; aérodynamique et physiologie de la parole; communication animale et langage humain. Didier Demolin a été directeur du laboratoire de phonétique et phonologie expérimentale à l'Université libre de Bruxelles (1996 à 2002) et de l'équipe SLD au Gipsa-lab (2010-2014). Il a enseigné à l'Université de Provence (1996-1999), à l'Université de Sao Paulo (2003-2009).

Notre consortium a un niveau d'excellence internationale qui lui permettra d'atteindre les objectifs du projet **ArtSpeech**.

3.2.8. JUSTIFICATION SCIENTIFIQUE ET TECHNIQUE DES MOYENS DEMANDES PAR PARTENAIRE

LORIA

Un doctorant recruté dans le cadre de ce projet travaillera sur la modélisation articulatoire, l'adaptation des modèles articulatoires à un nouveau locuteur et sur le modèle de coarticulation, c'est-à-dire la tâche 2. Par ailleurs, nous emploierons quelques stagiaires (12 mois en tout) afin de préparer les données articulatoires.	108660€
3 portables et 3 postes de travail fixes puissants utilisés par les personnes travaillant sur la tâche 2 et la tâche 5.	15000€
Les frais de mission se répartissent en 7200 € destinés aux déplacements du doctorant qui passera une partie de son temps à Grenoble au Gipsa-lab et aux réunions d'avancement, 11400€ pour des conférences en Europe et 11700 € pour des conférences hors Europe.	30300€
Aide demandée en incluant les frais de gestion	160118€

Gipsa-lab

Un doctorant recruté dans le cadre de ce projet travaillera sur le développement et la validation de modèles physiques des sources de son en parole, de leur coordination et de la propagation et du rayonnement acoustique (tâche 1).	100980€
Equipement : 1 station de travail et capteurs de pression	15000€
Autres dépenses : frais de publication, logiciels (Labview), périphériques informatique (stockage), consommables pour impression 3D	7300€

Les frais de mission sont destinés à financer les déplacements du doctorant dans les différents laboratoires partenaires ainsi qu'à des conférences internationales.	9000€
Aide demandée en incluant les frais de gestion	137571€

IADI

L'équipe du laboratoire IADI se consacrera essentiellement à la réalisation de la partie IRM de la tâche 4 (Acquisition de données de la production de la parole) du projet ArtSpeech. Le IADI consacrera 38 homme/mois dont un post-doc (tâche 4.1) qui sera recruté à 100% sur les fonds du projet pour une période de 18 mois et un montant de 65250 €. Pierre-André Vuissoz et Freddy Odille se consacreront à 30 % chacun au projet pendant les 18 mois de la réalisation de la tache 4.	65250€
La venue d'un nouveau post-doc dans le laboratoire nécessite la mise à disposition d'une station de travail puissante pour le développement des algorithmes de reconstruction compensée en mouvement, qui permettront d'atteindre les spécifications nécessaires au projet en termes de résolution temporelle et spatiale.	5000€
La gestion du protocole d'acquisition IRM sur sujet sain avec accord CPP et assurance attenante proprement dite sera sous-traitée au CIC-IT de Nancy qui dispose des infrastructures et personnel nécessaires. Pour les justifications des IRM sur sujet sain inclus dans le devis voir programme détaillé de la tache 4.	17677€
En plus des déplacements sur les sites des différents partenaires du Projet (Paris, Grenoble) pour préparer l'acquisition et dépouiller les différentes bases d'images IRM, et la formation des chercheurs dans des workshops spécialisés, l'aide servira à financer la publication dans les journaux de référence du domaine et la participation à diverses conférences internationales de sociétés savantes telles que l'ISMRM, l'ESMRMB, … dans lesquelles ces acquisitions et leurs résultats seront présentées. (10000 €/ 3 personnes).	10000€
Aide demandée en incluant les frais de gestion	101844€

LPP

Un postdoctorant recruté dans le cadre de ce projet travaillera sur l'acquisition de données aérodynamiques (tâche 4.2 mais sans le développement du nouveau plétismographe qui sera pris en charge par les permanents) et la coordination entre les plis vocaux et les articulateurs de la parole (tâche 2.2).	69716€
Matériel d'acquisition de données aérodynamiques et poste de travail	10000€
Les frais de mission sont destinés à financer les déplacements du doctorant dans les différents laboratoires partenaires ainsi qu'à des conférences internationales.	17000€
Aide demandée en incluant les frais de gestion	100585€

3.3. Calendrier

Les chiffres représentent les efforts cumulés des membres permanents et non permanents (recrutés) dans le cadre du projet. Le taux de précarité est de 25 % = 36/(36 + 108) (total postdoc / (total permanents + non-permanents hors doctorants)).

	Tâches	LORIA	Gipsa-Lab	IADI	LPP	Chronogramme													
		Partenaires				Année 1				Année 2				Année 3				Année 4	
						I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II
T0	**Coordination**	5	2	2	2														

T 1	**Simulations aérodynamiques et acoustiques**																		
1.1	Modèle de source vocale		12																
1.2	Modèle de plosives		16																
1.3	Modèle de fricatives		18																
T2	**Source et scenarii de coordination**																		
2.1	Paramètres de contrôle		10																
2.2	Intégration des mesures				16														
2.3	Construction des scenarii	12																	
T 3	**Contrôle de l'évolution temporelle de la géométrie**																		
3.1	Modèles articulatoires	12																	
3.2	Passage à la fonction d'aire	10																	
3.3	Adaptation de modèles	10																	
3.4	Algorithme de coarticulation	19																	
T 4	**Acquisition de données de production de la parole**																		
4.1	IRM statiques et dynamiques			36															
4.2	Plis vocaux et aérodynamiques				24														
T 5	**Architecture générale**																		
5.1	Paramètres de source	4																	
5.2	Intégration	6																	

4. STRATEGIE DE VALORISATION, DE PROTECTION ET D'EXPLOITATION DES RESULTATS, IMPACT GLOBAL DE LA PROPOSITION

4.1. Impact scientifique

En premier lieu, les retombées scientifiques concernent les sciences de la parole. L'utilisation de la synthèse articulatoire en tant qu'outil d'exploration du lien entre gestes articulatoires ou phonatoires, et le signal acoustique avait été proposée dès la fin des années 70. En effet la phonétique a souvent consisté à interpréter les indices observés sur un spectrogramme de parole en termes de gestes articulatoires et de paramètres aérodynamiques. ArtSpeech ouvrira une nouvelle voie en offrant un lien direct entre les domaines articulatoire/phonatoire et acoustique. Il sera possible, par exemple, d'explorer l'origine des indices acoustiques qui sous-tendent l'identification des sons de la parole, et particulièrement des consonnes. Il s'agit d'un bouleversement profond de la manière d'aborder la phonétique. Du point de vue de la production de la parole l'une des grandes questions scientifiques a toujours été la nature des paramètres contrôlés par le locuteur. Faute d'un modèle physique proche de la réalité les paramètres utilisés actuellement ne sont sans doute pas les bons. Ce projet offrira donc de nouvelles pistes pour aborder le contrôle moteur de la parole.
Le second domaine concerné est la phonologie. On connaît mal l'étendue des traits phonétiques, et il est en général difficile de savoir si la cooccurrence de deux indices acoustiques découle d'une évolution du langage et de l'organisation d'un système phonologique ou de contraintes aérodynamiques et acoustiques qui imposent cette cooccurrence. Dans certains cas, il n'a pas été possible de répondre à cette question avant de trouver une langue où cette cooccurrence n'existe pas.

4.1.1. STRATEGIE DE VALORISATION ET DE PROTECTION DES RESULTATS

Les partenaires du projet acceptent de conclure un accord de consortium basé sur le modèle unicANR (http://www.anrt.asso.fr/fr/espace_europe/accords_consortium.jsp), afin de mettre en place les conditions et modalités de la réalisation du projet, mais aussi de fixer leurs droits et devoirs respectifs.

Cet accord précisera notamment que la propriété des résultats du projet revient aux parties qui ont contribué à l'obtention de ces résultats, à proportion de leurs contributions respectives. Les parties copropriétaires de résultats communs brevetables décideront si ces résultats feront l'objet d'un dépôt de brevet en leurs noms communs et désigneront l'une d'entre elles qui sera responsable des procédures d'enregistrement et de maintien en vigueur au nom des parties concernées. L'utilisation commerciale ou industrielle des résultats doit faire l'objet d'une compensation financière versée aux autres parties copropriétaires, selon des termes qui seront définis par la suite.

Les résultats attendus du projet regroupent les données sur la production de la parole, les dispositifs et algorithmes d'acquisitions des données, les algorithmes de simulation aéroacoustique, les modèles articulatoires et algorithmes de modélisation de la coarticulation, et enfin le système de synthèse articulatoire.

Compte tenu des efforts que représentent l'acquisition de données de production de la parole et leur rareté, que ça soit pour les données aérodynamiques, comme pour les données IRM, nous les rendrons disponibles à la fin du projet sous licence Creative Commons Non Commercial - Attribution - Share alike (CC-BY-NC-SA). L'avantage de la mention Attribution est d'obliger les utilisateurs à citer l'origine des données ce qui est important pour valoriser l'image du consortium et de l'ANR.

Les algorithmes d'acquisition IRM seront valorisés à travers des brevets compte tenu des utilisations potentielles dans le domaine des pathologies de la parole et plus généralement des organes qui se déforment rapidement au cours du temps.
Les autres algorithmes concernant la synthèse 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.

4.1.2. STRATÉGIE ET DOMAINES D'EXPLOITATION

Les retombées économiques concernent en premier lieu de nombreuses applications de la synthèse de la parole, et plus spécifiquement celles demandant une grande flexibilité du point de vue des expressions ou plus généralement des plis vocaux, et du point de vue du conduit vocal. La flexibilité est un avantage déterminant pour le développement de nouveaux agents conversationnels, et cela d'autant plus qu'on demande souvent aux agents d'exagérer les expressions, et de s'adapter à des types de locuteurs variés.

Un autre champ applicatif est l'apprentissage des langues étrangères et de l'acquisition du langage. Il s'agit à la fois d'un enjeu économique et sociétal majeur. Dans les deux cas, ce sont surtout les possibilités de jouer sur l'articulation et le contrôle aérodynamique afin d'illustrer la production de nouveaux contrastes phonétiques et de fournir un retour acoustique aux apprenants qui seront exploitées.
Les retombées dans le domaine médical sont doubles. Le premier point porte sur l'exploitation des techniques d'acquisition de données dynamiques, particulièrement pour l'IRM qui est de plus en plus souvent utilisée pour imager des organes qui se déforment au cours du temps. Les techniques et algorithmes développés pourront donner lieu à des brevets. Le second point concerne les pathologies de la production de la parole, et l'impact des interventions chirurgicales qu'elles portent sur les plis vocaux (ou leur voisinage immédiat), ou directement le conduit vocal. Pour ce dernier point la possibilité de jouer sur la géométrie du conduit vocal ou les caractéristiques des plis vocaux peut guider le choix du chirurgien.
Le domaine de l'expertise judiciaire a été évoqué par l'un des relecteurs de la proposition courte. S'il est indéniable que notre projet vise à copier les processus de production de la parole, à la fois du point de vue géométrique et du contrôle des plis vocaux, il reste que le contrôle moteur est aussi une source de variabilité importante dont l'impact est sans doute lui aussi déterminant. Cette application n'est donc pas envisagée pour l'instant.

5. REFERENCES BIBLIOGRAPHIQUES DES PARTENAIRES DU CONSORTIUM

Gipsa-lab

1. Blandin R., Arnela M., Laboissiere R., Pelorson X., Guasch O., Van Hirtum A., Laval X., 2015. Effects of higher order propagation modes in vocal tract like geometries. J. Acoust. Soc. Am., 137:832-843.
2. Ruty N., Pelorson X., Van Hirtum A., Lopez I., Hirschberg A., 2007. An in-vitro setup to test the relevance and the accuracy of low-order models of the vocal folds. J. Acoust. Soc. Am., 121:479-490.
3. Fujiso Y., Nozaki K., Van Hirtum A., 2015. Towards sibilant physical speech screening using oral tract volume reconstruction: some preliminary observations. Applied Acoustics,96:101-107.
4. Delebecque L., Pelorson X., Beautemps D., Laval X.(2013) Physical modeling of bilabial plosives production. Proceedings of POMA - ICA 2013 Montreal CA 2013-06-02
5. Fabre B., Gilbert J., Hirschberg A., Pelorson X. (2012) Aeroacoustics of Musical Instruments, Annual Review of Fluid Mechanics 44 2012 1-25

LPP

1. Amelot, A. (2009). Dispositifs d'imagerie pour l'observation de l'activité vélo-pharyngée. In A. Marchal & C. Cavé (eds.). Techniques d'imagerie médicale pour l'étude de la parole. Paris. Hermès.
2. Demolin, D. Hassid, S. Metens T. and Soquet A. (2002). Real Time MRI and articulatory coordination in speech. In Model-driven acquisition : Acquisition conduite par le modèle, *Comptes Rendus de l'Académie des Sciences*. *Biologie* 325, 547-556.
3. Demolin, D. (2007). Phonological Universals and the control and regulation of speech production. In: Solé, M-J., Ohala, M. & Beddor, P. (Eds.) *Experimental approaches to phonology*. Oxford: Oxford University Press, 75-92.
4. Demolin, D. & Metens, T. (2009). L'imagerie par resonance magnétique en temps réel pour l'étude de la parole. In A. Marchal & C. Cavé (eds.). *Techniques d'imagerie médicale pour l'étude de la parole*. Paris. Hermès. 257-271.
5. Honda, K. & Maeda, S. (2008). Brevet Photoglottographe, number WO2009010655 A1.

IADI

1. Odille F, Vuissoz P.-A., Marie PY, Felblinger J. (2008). Generalized reconstruction by inversion of coupled systems (GRICS) applied to free-breathing MRI. Magn Reson Med;60(1):146-157
2. Vuissoz PA, Odille F, Fernandez B, Lohezic M, Benhadid A, Mandry D, Felblinger J. (2012). Free-breathing imaging of the heart using 2D cine-GRICS (generalized reconstruction by inversion of coupled systems) with assessment of ventricular volumes and function. J Magn Reson Imaging;35(2):340-351
3. Claudon M, Durand E, Grenier N, Prigent A, Balvay D, Chaumet-Riffaud P, Chaumoitre K, Cuenod CA, Filipovic M, Galloy MA, Lemaitre L, Mandry D, Micard E, Pasquier C, Sebag GH, Soudant M, Vuissoz PA (2014), Guillemin F. *Chronic Urinary Obstruction: Evaluation of Dynamic Contrast-enhanced MR Urography for Measurement of Split Renal Function.* Radiology: 2014;273(3):801-812.
4. Odille F., Vuissoz P.-A., Felblinger J. (2009); Procédé de reconstruction d'un signal à partir de mesures experimentales perturbees et dispositif de mise en œuvre patent FR2923598. 2009-05-15.
5. Vuissoz P-A, Odille F, Laprie Y, Vincent E, Hossu G, Felblinger J.(2014) *Speech Cine SSFP with optical microphone synchronization and motion compensated reconstruction.*; ISMRM Workshop Tromsø, Norway.

LORIA

1. Laprie Y, M. Loosvelt, S. Maeda, R. Sock, F. Hirsch. – « Articulatory copy synthesis from cine X-ray films ». – In: InterSpeech - 14th Annual Conference of the International Speech Communication Association - 2013. – Lyon, France, 2013.
2. Toutios A., Ouni S. & Laprie Y. (2011) – "Estimating the control parameters of an articulatory model from electromagnetic articulograph data". – The Journal of the Acoustical Society of America 129(5), pp. 3245–3257.
3. Laprie Y., R.Sock, B. VaxelairE, B. Elie. – « Comment faire parler les images aux rayons X du conduit vocal ? ». – In : Actes du Congrès Mondial de la Linguistique Française. – Berlin, 2014
4. Laprie Y. & Busset J. (2011), "Construction and Evaluation of an Articulatory Model of the Vocal Tract". In Proceedings of Eusipco, 466-470.
5. S. Ouni S., Colotte V., Musti U., A. Toutios A., Wrobel-Dautcourt B., Berger M.O., Lavecchia C. (2013). Acoustic-visual synthesis technique using bimodal unit-selection, EURASIP Journal on Audio, Speech, and Music Processing

6. RÉFÉRENCES BIBLIOGRAPHIQUES

[1] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux. Three-dimensional linear articulatory modeling of tongue, lips and face based on MRI and video images. *Journal of Phonetics*, 30(3):533–553, 2002.
[2] Youkyung Bae, David P. Kuehn, Charles A. Conway, and Bradley P. Sutton. Real-Time Magnetic Resonance Imaging of Velopharyngeal Activities With Simultaneous Speech Recordings. *CLEFT PALATE-CRANIOFACIAL JOURNAL*, 48(6):695–707, November 2011.
[3] A.P. Benguerel and H.A. Cowan. Coarticulation of upper lip protrusion in french. *Phonetica*, 30:41–55, 1999.
[4] P. Birkholz. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLOS one*, 8(4), 2013.
[5] P. Birkholz and D. Jackel. A three-dimensional model of the vocal tract for speech synthesis. In *15th International Congress of Phonetic Sciences - ICPhS'2003, Barcelona, Spain*, pages 2597–2600, Aug 2003.
[6] R. Blandin, M. Arnela, R. Laboissière, X. Pelorson, O. Guasch, A. Van Hirtum, and X. Laval. Effects of higher order propagation modes in vocal tract like geometries. *Journal of the Acoustical Society of America*, 137(2):832–843, 2015.
[7] M. Echternach, M. Markl, and B. Richter. Dynamic real-time magnetic resonance imaging for the analysis of voice physiology. *Curr Opin Otolaryngol Head Neck Surg*, 20(6):450–7, December 2012.
[8] B.D. Erath, S.D. Peterson, M. Za nartu, G.R. Wodicka, and M.W. Plesniak. A theoretical model of the pressure field arising from asymmetric intraglottal flows applied to a two-mass model of the vocal folds. *Journal of the Acoustical Society of America*, 130(1):389–403, 2011.
[9] Sandra L. Ettema, David P. Kuehn, Adrienne L. Perlman, and Noam Alperin. Magnetic resonance imaging of the levator veli palatini muscle during speech. *The Cleft palate-craniofacial journal*, 39(2):130–144, 2002.
[10] Bell-Berti F. and Harris K. S. Temporal patterns of coarticulation: lip rouding. *Journal of the Acoustical Society of America*, 71:449–459, 1982.
[11] G. Fant. *The F-Patterns of compound tube resonators and hors*. The Hague: Mouton & Co., 1970.
[12] E. Farnetani. Labial coarticulation. In W. J. Hardcastle and N. Hewlett, editors, *In Coarticulation: Theory, data and techniques*, chapter 8. Cambridge university press, Cambridge, 1999.
[13] Maojing Fu, Bo Zhao, Christopher Carignan, Ryan K. Shosted, Jamie L. Perry, David P. Kuehn, Zhi-Pei Liang, and Bradley P. Sutton. High-resolution dynamic speech imaging with joint low-rank and sparsity constraints. *Magnetic Resonance in Medicine*, 2014.
[14] W.L. Henke. Preliminaries to speech synthesis based on an articulatory model. In *IEEE Conference on Speech Communication and Processing*, pages 170–177, Air Force Cambridge Res. Lab., 1967.
[15] K. Honda and S. Maeda. Glottal-opening and airflow pattern during production of voiceless fricatives: A new non-invasive instrumentation. *Journal of the Acoustical Society of America*, 123(5):3788, 2008.
[16] K. Honda and S. Maeda. Procédé et équipement non-invasif de photoélectroglottographie, patent number WO2009010655 A1, 2008.
[17] M.S. Howe and R.S. McGowan. Aeroacoustics of [s]. *Proc. of the Royal Society A*, 461:1005–1028, 2005.
[18] K. Ishizaka and J. L. Flanagan. Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell Syst. Technol. J.*, 51:1233–1268, 1972.
[19] K. Iskarous, L.M. Goldstein, D.H. Whalen, M.K. Tiede, and P.E. Rubin. CASY: The Haskins configurable articulatory synthesizer. In *15th International Congress of Phonetic Sciences 2003 - ICPhS'2003, Barcelone, Espagne*, pages 185–188, Barcelona, Aug 2003.
[20] J. Jansson, A. Holmberg, R.V. de Abreu, C. Degirmenci, J. Hoffman, M.K. arlsson, and M. Abom. Adaptive stabilized finite element framework for simulation of vocal fold turbulent fluid-structure interaction. In *Proceedings of 21st International Congress on Acoustics, ICA 2013 - 165th Meeting of the Acoustical Society of America*, Montreal, 2013.
[21] Y. C. Kim, M. I. Proctor, S. S. Narayanan, and K. S. Nayak. Improved imaging of lingual articulation using real-time multislice MRI. *J Magn Reson Imaging*, 35(4):943–8, April 2012.
[22] M. Krane. Aeroacoustic production of low-frequency unvoiced speech sounds. *Journal of the Acoustical Society of America*, 118(1):410–427, 2005.
[23] Y. Laprie and J. Busset. Construction and evaluation of an articulatory model of the vocal tract. In *19th European Signal Processing Conference - EUSIPCO-2011*, Barcelona, Spain, August 2011.
[24] Y. Laprie, M. Loosvelt, S. Maeda, E. Sock, and F. Hirsch. Articulatory copy synthesis from cine x-ray films. In *Interspeech 2013 (14th Annual Conference of the International Speech Communication Association)*, Lyon, France, August 2013.
[25] Y. Laprie, B. Vaxelaire, and M. Cadot. Geometric articulatory model adapted to the production of consonants. In *10th International Seminar on Speech Production (ISSP)*, Köln, Allemagne, May 2014.
[26] N.J.C. Lous, G.C.J. Hofmans, R.N.J. Veldhuis, and A. Hirschberg. A symmetrical two-mass vocal-fold model coupled to vocal tract and trachea, with application to prosthesis design. *Acustica*, 42:1135–1150, 1998.

[27] J. C. Lucero and L. L. Koenig. Simulations of temporal patterns of oral airflow in men and women using a two-mass model of the vocal folds under dynamic control. *Journal of the Acoustical Society of America*, 117(3):1362–1372, 2005.
[28] Michael Lustig, David Donoho, and John M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *MAGNETIC RESONANCE IN MEDICINE*, 58(6):1182–1195, December 2007.
[29] S. Maeda. A digital simulation of the vocal tract system. *Speech Communication*, 1:199–229, 1982.
[30] S. Maeda. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W.J. Hardcastle and A. Marchal, editors, *Speech production and speech modelling*, pages 131–149. Kluwer Academic Publisher, Amsterdam, 1990.
[31] S. Maeda. Phoneme as concatenable units: VCV synthesis using a vocal tract synthesizer. In A. P. Simpson ans M. Pötzold, editor, *Sound Patterns of Connected Speech: Description, Models and Explanation, Proceedings of the symposium held at Kiel University, Arbeitsberichte des Institut für Phonetik und digitale Spachverarbeitung der Universitaet Kiel:31*, pages 145–164, June 1996.
[32] Shinji Maeda and Yves Laprie. Vowel and prosodic factor dependent variations of vocal-tract length. In *InterSpeech - 14th Annual Conference of the International Speech Communication Association - 2013*, Lyon, France, August 2013.
[33] R.S. McGowan, L. Koenig, and A. Löfqvist. Vocal tract aerodynamics ins /aca/ utterances: Simulations. *SPECOM*, 16:67–88, 1994.
[34] P. Mermelstein. Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53:1070–1082, 1973.
[35] H. Nam V. Mitra, M. Tiede, E. Saltzman, L. Goldstein, C. Epsy-Wilson, and M. Hasegawa-Johnson. A procedure for estimating gestural scores from natural speech. In *11th Annual Conference of the International Speech Communication Association - INTERSPEECH 2010*, Makuhari, Chiba, Japan, 2010.
[36] H. Nam, V. Mitra, M. Hasegawa-Johnson, C. Epsy-Wilson, E. Saltzman, and L. Goldstein. A procedure for estimating gestural scores from speech acoustics. *Journal of the Acoustical Society of America*, 132(6):3080–3989, 2012.
[37] Aaron Niebergall, Shuo Zhang, Esther Kunay, Götz Keydana, Michael Job, Martin Uecker, and Jens Frahm. Real-time MRI of speaking at a resolution of 33 ms: Undersampled radial FLASH with nonlinear inverse reconstruction. *Magnetic Resonance in Medicine*, 69(2):477–485, February 2013.
[38] F. Odille, P. A. Vuissoz, P. Y. Marie, and J. Felblinger. Generalized reconstruction by inversion of coupled systems (GRICS) applied to free-breathing MRI. *Magn Reson Med*, 60(1):146–57, July 2008.
[39] S.E. Öhman. Coarticulation inVCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, 39(1):151–168, 1966.
[40] S.E.G. Öhman. Numerical model of coarticulation. *J. Acoust. Soc. Am.*, 41:310–320, 1967.
[41] Browman C. P. and Goldstein L. Articulatory gestures as phonological units. *Phonology*, 6:201–251, 1989.
[42] X. Pelorson, A. Hirschberg, A.P.J. Wijnands, H.M., and A. Bailliet. Theoretical and experimental study of quasisteady-flow separation within the glottis during phonation. application to a modified two-mass model. *Journal of the Acoustical Society of America*, 96(6):3416–3431, 1994.
[43] X. Pelorson, X. Vescovi, C. Castelli, E. Hirschberg, A. Wijnands, A.P.J. Bailliet, and H.M.A. Hirschberg. Description of the flow through in-vitro models of the glottis during phonation. application to voiced sounds synthesis. *Acta Acustica*, 82:358–361, 1996.
[44] Michael Proctor, Erik Bresch, Dani Byrd, Krishna Nayak, and Shrikanth Narayanan. Paralinguistic mechanisms of production in human "beatboxing": A real-time magnetic resonance imaging study. *The Journal of the Acoustical Society of America*, 133(2):1043–1054, 2013.
[45] K. P. Pruessmann, M. Weiger, P. Bornert, and P. Boesiger. Advances in sensitivity encoding with arbitrary k-space trajectories. *Magn Reson Med*, 46(4):638–51, 2001.
[46] Sandra M. Rua Ventura, Diamantino Rui S. Freitas, Isabel Maria A. P. Ramos, and Joao Manuel R. S. Tavares. Morphologic Differences in the Vocal Tract Resonance Cavities of Voice Professionals: An MRI-Based Study. *JOURNAL OF VOICE*, 27(2):132–140, March 2013.
[47] N. Ruty, X. Pelorson, A. Van Hirtum, I. Lopez-Arteaga, and A. Hirschberg. An in vitro setup to test the relevance and the accuracy of low-order vocal folds models. *Journal of the Acoustical Society of America*, 121(1):479–490, 2007.
[48] R. C. Scherer, D. Shinwari, K.J. DeWitt, C. Zhang, B.R. Kucinschi, and A.A. Afjeh. Intraglottal pressure profiles for a symmetric and oblique glottis with a uniform duct. *Journal of the Acoustical Society of America*, 112(4):1253–1256, 2001.
[49] Andrew D Scott, Marzena Wylezinska, Malcolm J Birch, and Marc E Miquel. Speech MRI: Morphology and function. *Physica Medica*, 30(6):604–618, 2014.
[50] C. Scully. Linguistic units and units of speech production. *Speech communication*, 6(2):77–142, 1987.
[51] I. Steinecke and H. Herzel. Bifurcations in an asymmetric vocal-fold model. *Journal of the Acoustical Society of America*, 97(3):1874–1884, 1995.

[52] M. Stone, E. P. Davis, A. S. Douglas, M. NessAiver, R. Gullapalli, W. S. Levine, and A. Lundberg. Modeling the motion of the internal tongue from tagged cine-MRI images. *J Acoust Soc Am*, 109(6):2974–82, June 2001.
[53] B.H. Story. Phrase-level speech simulation with an airway modulation model of speech production. *Computer, Speech and Language*, pages 989–1010, 2013.
[54] B.H. Story and I. R. Titze. Voice simulation with a body-cover model of the vocal folds. *Journal of the Acoustical Society of America*, 97(2):1249–1260, 1995.
[55] Johan Sundberg. Articulatory configuration and pitch in a classically trained soprano singer. *Journal of Voice*, 23(5):546–551, 2009.
[56] M. Uecker, S. Zhang, D. Voit, A. Karaus, K. D. Merboldt, and J. Frahm. Real-time MRI at a resolution of 20 ms. *NMR Biomed*, 23(8):986–94, October 2010.
[57] Sandra M. Rua Ventura, Diamantino Rui S. Freitas, and Joao Manuel R. S. Tavares. Toward Dynamic Magnetic Resonance Imaging of the Vocal Tract During Speech Production. *Journal of Voice*, 25(4):511–518, July 2011.
[58] P.-A. Vuissoz, F. Odille, B. Fernandez, M. Lohezic, A. Benhadid, D. Mandry, and J. Felblinger. Free-breathing imaging of the heart using 2d cine-grics (generalized reconstruction by inversion of coupled systems) with assessment of ventricular volumes and function. *Journal of Magnetic Resonance Imaging*, 35(2):340–351, 2012.
[59] P.A. Vuissoz, F. Odille, Y. Laprie, E. Vincent, G. Hossu, and J. Felblinger. Speech Cine SSFP with optical microphone synchronization and motion compensated reconstruction. In *ISMRM Workshop on Motion Correction in MRI*, Tromso, Norvège, May 2014.