Traçage et suivi du palais dans les images échographiques : un système de biofeedback visuel pour l'apprentissage d'une langue seconde

par

Hana BEN ASKER

MÉMOIRE PAR ARTICLES PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE AVEC MÉMOIRE EN GÉNIE DES TECHNOLOGIES DE L'INFORMATION
M. Sc. A.

MONTRÉAL, LE 19 DÉCEMBRE 2025

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CETTE THÈSE A ÉTÉ ÉVALUÉE

PAR UN JURY COMPOSÉ DE:

Mme. Catherine Laporte, directrice de mémoire
Département de génie électrique à l'École de technologie supérieure

Mme. Lucie Ménard, codirectrice
Département de linguistique à l'Université de Québec à Montréal

M. Walcir Cardoso, codirecteur
Département d'éducation à l'Université Concordia

M. David Labbé, président du jury
Département de génie logiciel et TI à l'École de technologie supérieure

Mme. Sylvie Ratté, membre du jury
Département de génie logiciel et TI à l'École de technologie supérieure

ELLE A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 11 DÉCEMBRE 2025

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Je tiens à exprimer ma sincère gratitude à ma directrice de recherche, Mme Catherine Laporte, pour sa confiance, son mentorat tout au long de ce cursus et ses encouragements à repousser mes limites. Sa bienveillance, son écoute et son empathie dépassent le cadre d'une direction de recherche.

Je remercie également mes codirecteurs, Mme Lucie Ménard et M. Walcir Cardoso, pour leur encadrement, leurs conseils et leurs précieux retours.

Je souhaite remercier les membres du jury, Mme Sylvie Ratté et M. David Labbé pour le temps qu'ils ont consacré à l'évaluation de ce mémoire.

Ma reconnaissance s'étend aussi à Mme Eija Aalto, avec qui j'ai eu le plaisir de collaborer. Elle a su m'initier au monde de la parole, toujours avec son esprit humoristique.

Je remercie chaleureusement mes collègues de laboratoire – Dominic Tremblay, Minoru Yoshida, Sahar Farazi, Gabrielle Roy et Soroush Javadi – pour leur soutien, leur aide précieuse et les bons moments partagés au laboratoire.

Ces travaux n'auraient pas été possibles sans le soutien financier de Mitacs, de l'ÉTS, du CRSNG et du programme OPSIDIAN. Je remercie également l'équipe OPSIDIAN pour la qualité de ses formations et pour les amitiés qui s'y sont nouées.

Enfin, un immense merci à mes parents, mes sœurs, ma famille et mes amis pour leur soutien constant.

Traçage et suivi du palais dans les images échographiques : un système de biofeedback visuel pour l'apprentissage d'une langue seconde

Hana BEN ASKER

RÉSUMÉ

L'imagerie échographique (US) émerge comme un outil précieux dans les sciences de la parole, offrant une fenêtre non invasive et en temps réel sur les mouvements de la langue durant l'articulation. Pour l'analyse articulatoire, et dans les systèmes de biofeedback visuel destinés à l'apprentissage d'une langue seconde (L2), un tracé du palais dur peut constituer une valeur ajoutée significative. Il sert non seulement de cible passive pour de nombreux sons consonantiques, mais aussi de cadre de référence stable pour normaliser les mesures articulatoires, comparer les productions et guider les apprenants vers des gestes plus ciblés. Toutefois, l'utilisation d'un tracé du palais en biofeedback est limitée par un défi majeur : son invisibilité durant la parole, due à l'interface air-tissu qui bloque les ultrasons. Les méthodes existantes, peu nombreuses, se limitent donc à la reconstruction statique du tracé, souvent acquise par le biais d'une tâche de déglutition.

Ce mémoire propose deux contributions complémentaires. Premièrement, nous établissons des pratiques exemplaires pour le traçage fiable du palais. Notre analyse comparative (51 vidéos, 17 participants, 3 tâches, 3 méthodes) démontre que la déglutition sèche combinée à la méthode du squelette d'échos cumulés (CES) produit le meilleur accord inter-juges (erreur de 2,87 mm) et valide la viabilité du CES automatique (erreur de 2,63 mm).

Deuxièmement, ce mémoire introduit une méthode de suivi automatique du palais. L'approche repose sur l'inférence du mouvement palatin à partir d'un repère constamment visible : le tendon du muscle génioglosse. Un système hybride, combinant un détecteur YOLOv8 et un filtre particulaire, assure un suivi robuste du tendon, permettant d'inférer la position du palais via un modèle de transformation rigide. Évaluée sur 71 vidéos (déglutitions et parole libre), la méthode atteint une erreur moyenne de 1,34 à 2,68 mm et reste fiable même avec moins de 5 % de visibilité palatine. L'hypothèse d'inférence anatomique est validée par une corrélation significative ($r = 0,64$, $p = 0,001$) entre l'exactitude du suivi du tendon et celle du palais.

Ces avancées posent les fondations de systèmes de biofeedback visuel améliorés pour l'orthophonie et l'apprentissage d'une L2. La première étude établit une pratique exemplaire (déglutition sèche et CES) pour obtenir un tracé de référence fiable et valide la méthode CES automatique, notant ses limites face aux artefacts causés par la présence de liquides. La seconde contribution s'appuie sur cette base pour proposer un suivi continu malgré l'invisibilité. Le potentiel de ce suivi est démontré par le prototype ReaPT, développé en marge de ce projet et salué par les utilisateurs. Les limites actuelles, comme l'hypothèse de transformation rigide affectée par les mouvements mandibulaires, ouvrent des pistes futures : modèles non rigides, ré-initialisation, et validation clinique.

Mots-clés: imagerie échographique de la langue, suivi du palais, biofeedback visuel, apprentissage d'une langue seconde, tendon du génioglosse, filtre particulaire

Tracing and Tracking the Palate in Ultrasound Images : A Visual Biofeedback System for Second Language Learning

Hana BEN ASKER

ABSTRACT

Ultrasound (US) imaging has emerged as a valuable tool in speech sciences, offering a non-invasive, real-time window into tongue movements during articulation. For articulatory analysis, and particularly in visual biofeedback systems for second language (L2) learning or clinical intervention, the hard palate trace can provide a significant added value. It serves not only as a passive target for many consonants but also as a stable frame of reference to normalize articulatory measurements, compare productions, and guide learners toward more targeted gestures. However, the use of the palate trace in biofeedback is limited by a major challenge : its invisibility during speech, caused by the air-tissue interface that blocks the ultrasound waves. Existing methods, which are few, are therefore limited to static reconstruction of the contour, often acquired through a swallowing task.

This thesis proposes two complementary contributions. First, we establish best practices for reliable palate tracing. Our comparative analysis (51 swallowing videos, 17 participants, 3 tasks, 3 methods) demonstrates that the dry swallow task combined with the Cumulative Echo Skeleton (CES) method yields the best inter-rater agreement (mean error : 2.87 mm) and validates the viability of the automatic CES approach (mean error : 2.63 mm).

Second, this thesis introduces an automatic palate tracking method. The approach relies on inferring palatal motion from a consistently visible anatomical landmark : the genioglossus tendon. A hybrid system, combining a YOLOv8 detector with a particle filter, ensures robust tendon tracking, allowing the palate's position to be inferred via a rigid transformation model. Evaluated on 71 videos (swallowing and free speech), the method achieves mean errors from 1.34 to 2.68 mm and remains reliable even when palate visibility drops below 5%. The hypothesis of anatomical inference is validated by a significant correlation ($r = 0.64$, $p = 0.001$) between tendon and palate tracking accuracy.

These advances lay the foundation for improved visual biofeedback systems for speech therapy and L2 learning. The first study establishes a best practice (dry swallow and CES) for obtaining a reliable reference trace and validates the automatic CES method, while noting its limitations with artifacts caused by the presence of liquids. The second contribution builds on this to propose continuous tracking despite invisibility. The potential of this tracking is demonstrated by the ReaPT prototype, developed as part of this project and praised by users. Current limitations, such as the rigid transformation hypothesis being affected by jaw movement, open clear future directions : non-rigid models, re-initialization, and clinical validation.

Keywords: tongue ultrasound imaging, palate tracking, visual biofeedback, second language acquisition, genioglossus tendon, particle filter

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

Acc@d	Accuracy at a distance threshold d
CES	Squelette d'échos cumulés
CNNs	Réseaux de neurones convolutifs
EMA	Électromagnétoarticulographie
EPG	Électropalatographie
ÉTS	École de Technologie Supérieure
FOV	Field Of View
IRM	Imagerie par résonance magnétique
L2	Langue seconde
MED	Mean Euclidean Distance
MSD	Mean Sum of Distances
PC	Palate-Corrected
ReaPT	Real-time Palate Tracker
RMSE	Erreur Quadratique Moyenne
ROI	Région d'intérêt
SLP	Speech-Language Pathology
SSD	Speech Sound Disorders
STD	Standard Deviations
TB	Tendon-Based
US	Ultrasound
YOLO	You Only Look Once

LISTE DES SYMBOLES ET UNITÉS DE MESURE

mm	millimeters

INTRODUCTION

0.1 Énoncé du problème

L'apprentissage d'une langue seconde (L2) pose souvent des difficultés, en particulier lorsque les apprenants doivent produire des sons absents de leur langue maternelle ou réalisés selon des schémas articulatoires différents. Ces difficultés peuvent persister malgré une exposition prolongée à la langue cible. Les gestes linguaux, essentiels à la production phonétique, se déroulent à l'intérieur de la cavité buccale et ne sont donc pas directement observables (Bliss, Abel & Gick, 2018; Bryfonski, 2023).

Par ailleurs, l'enseignement de la prononciation en L2 demeure souvent limité. Les cours reposent principalement sur des consignes auditives et des descriptions verbales, sans supports permettant de visualiser les gestes articulatoires. Cette absence de retour visuel complique l'acquisition de sons dont la production ne peut être déduite de l'écoute seule (Bliss *et al.*, 2018).

Le conduit vocal, agissant comme une « boîte noire », limite la conscience proprioceptive de l'apprenant. Même les apprenants expérimentés ont souvent une perception imparfaite des positions et mouvements précis de leur langue, ce qui entrave leur capacité à autoréguler leur production articulatoire (Ouni, 2014). Ce manque de visibilité n'affecte pas seulement l'apprenant ; il pose également un défi aux professionnels tentant de décrire ces gestes cachés. Par exemple, des études ont montré que les orthophonistes eux-mêmes, malgré leur formation spécialisée, peuvent développer des conceptualisations internes inexactes de l'articulation de sons complexes (Diekhoff & Lulich, 2022).

Depuis plusieurs décennies, différentes techniques d'imagerie articulatoire ont permis aux phonéticiens d'explorer la dynamique du conduit vocal, notamment l'imagerie par résonance magnétique (IRM), l'électromagnétoarticulographie (EMA) ou encore l'électropalatographie (EPG). Ces approches ont joué un rôle fondamental dans l'ouverture de la « boîte noire » du

conduit vocal et dans la compréhension fine des gestes articulatoires. Toutefois, en raison de leur coût, de leur caractère intrusif ou de leur faible accessibilité, leur utilisation demeure largement confinée aux laboratoires de recherche.

Dans ce contexte, l'imagerie échographique de la langue s'est imposée comme un outil prometteur de biofeedback visuel, offrant une fenêtre non invasive et en temps réel sur les mouvements de la langue pendant la parole (Cleland, Scobbie, Nakai & Wrench, 2015a). Le palais, bien que non essentiel, pourrait représenter une valeur ajoutée en fournissant un repère visuel de la cible à atteindre. Ce dernier sert non seulement de cible passive pour de nombreux sons consonantiques, mais aussi de cadre de référence essentiel pour normaliser les mesures articulatoires, comparer les productions intra- et inter-locuteurs, et guider les apprenants vers des gestes plus ciblés (Epstein & Stone, 2005).

Malheureusement, le palais lui-même est généralement invisible sur ces images. En effet, lorsque les ondes ultrasonores montent depuis la sonde placée sous la mâchoire et traversent la langue, elles rencontrent la poche d'air dans la cavité buccale. Cette interface entre la surface de la langue et l'air provoque une réflexion quasi totale des ondes, créant une ombre acoustique qui empêche les ondes d'atteindre et de visualiser le palais situé au-dessus. Cette invisibilité, combinée à la variabilité apparente de sa position due aux mouvements relatifs entre la sonde et de la tête, et les mouvements de la mâchoire, constitue un obstacle majeur à l'intégration du palais dans les systèmes de biofeedback. Les approches existantes se limitent à la reconstruction statique du palais, en exploitant des tâches comme la déglutition pour extraire un contour de référence. De plus, ces techniques restent très peu étudiées et l'on dispose de peu d'informations sur les conditions qui favorisent leur bon fonctionnement. Bien que ces méthodes soient utiles pour l'analyse post-hoc, elles ne permettent pas un suivi continu du palais au cours de la parole.

0.2 Objectifs

L'objectif principal de ce mémoire est de combler cette lacune en proposant une solution innovante pour le suivi automatique et continu du palais dans les séquences échographiques, même lorsqu'il est invisible. Plus précisément, ce travail vise à :

1. Établir des pratiques exemplaires pour le traçage manuel et automatique du palais, en évaluant la reproductibilité de différentes méthodes à travers diverses tâches de déglutition.
2. Valider une méthode automatique existante (le squelette d'échos cumulés).
3. Proposer et évaluer une méthode novatrice de suivi automatique du palais qui contourne le problème de son invisibilité en exploitant un repère anatomique stable et constamment visible : le tendon du muscle génioglosse.

0.3 Organisation du document

Ce mémoire s'organise en quatre chapitres structurés selon l'approche par articles. Le premier chapitre constitue une revue de littérature approfondie sur l'imagerie échographique de la langue et du palais et les techniques de suivi articulatoire. Ce chapitre introduit les principes fondamentaux de l'échographie, ses avantages et ses limitations, puis examine les méthodes existantes de suivi de la langue. Il explore également le rôle crucial du palais comme repère anatomique, en mettant en évidence les défis spécifiques liés à sa visualisation et aux approches actuelles pour sa délinéation.

Le deuxième chapitre présente notre article intitulé "Best Practices for Palate Tracing in Ultrasound Images", accepté et présenté à l'IEEE Engineering in Medicine and Biology Conference 2025, qui évalue de manière systématique les pratiques exemplaires pour le traçage du palais. Cette étude propose une validation de la méthode du squelette d'échos cumulés (CES) et démontre que la déglutition sèche constitue la condition optimale pour une délinéation fiable du palais.

Le troisième chapitre présente notre article intitulé "Palate Tracking in Ultrasound Images", soumis à la revue Journal of the Acoustical Society of America. Ce chapitre décrit une méthode innovante de suivi automatique du palais dans les séquences échographiques, même après de longues périodes d'invisibilité. La méthode proposée s'appuie sur l'inférence à partir du tendon du muscle génioglosse, un repère anatomique constamment visible, pour surmonter le défi majeur de l'invisibilité du palais dans les images échographiques.

Enfin, le quatrième chapitre synthétise les principales conclusions de notre travail et discute des perspectives futures. Il examine les implications théoriques et pratiques de notre approche, en particulier comment le suivi automatique continu du palais peut améliorer les systèmes de biofeedback visuel en fournissant aux apprenants une référence anatomique stable en temps réel. Ce chapitre propose également des pistes concrètes pour des recherches futures et des améliorations possibles des systèmes de biofeedback en temps réel.

CHAPITRE 1

REVUE DE LITTÉRATURE

L'étude de la production de la parole se heurte depuis longtemps à un défi fondamental : la nature largement inaccessible et invisible des articulateurs à l'intérieur de la cavité buccale. Ce que l'on appelle la boîte noire du conduit vocal a historiquement limité notre compréhension des gestes articulatoires précis qui sous-tendent la parole humaine. Pour les phonéticiens, les orthophonistes et les apprenants de langues, cette opacité a longtemps représenté un obstacle majeur, rendant difficiles la visualisation, la correction et l'apprentissage des mouvements complexes de la langue.

C'est dans ce contexte que l'imagerie ultrasonore de la langue est apparue comme une technologie pivot, offrant une fenêtre non invasive et en temps réel sur la cinématique articulatoire (Stone, 2005). Cette revue examine l'imagerie échographique pour étudier le conduit vocal, en se concentrant sur les techniques, les progrès et les difficultés du suivi de la langue et du palais, ainsi que ses applications dans l'apprentissage des langues et l'orthophonie.

1.1 Anatomie du conduit vocal

Pour interpréter correctement les images ultrasonores, une compréhension de l'anatomie fonctionnelle du conduit vocal est essentielle (Figure 1.1a).

L'échographie en vue sagittale capture une coupe médiane des articulateurs, révélant principalement la cinématique de la langue par rapport aux structures fixes de la cavité buccale. La langue, un hydrostat musculaire, est l'articulateur principal. Sa capacité à se déplacer dans son ensemble avec la mandibule et le plancher buccal et à modifier sa forme lui permet de moduler le conduit pharyngo-oral de deux manières fondamentales. Pour la production des voyelles, elle crée des configurations relativement ouvertes, où sa position (antérieure/postérieure) et sa hauteur (haute/basse) déterminent les résonances acoustiques (formants) qui différencient les sons comme /i/, /a/ et /u/ (Hixon, Weismer & Hoit, 2018). Pour les consonnes, la langue forme des

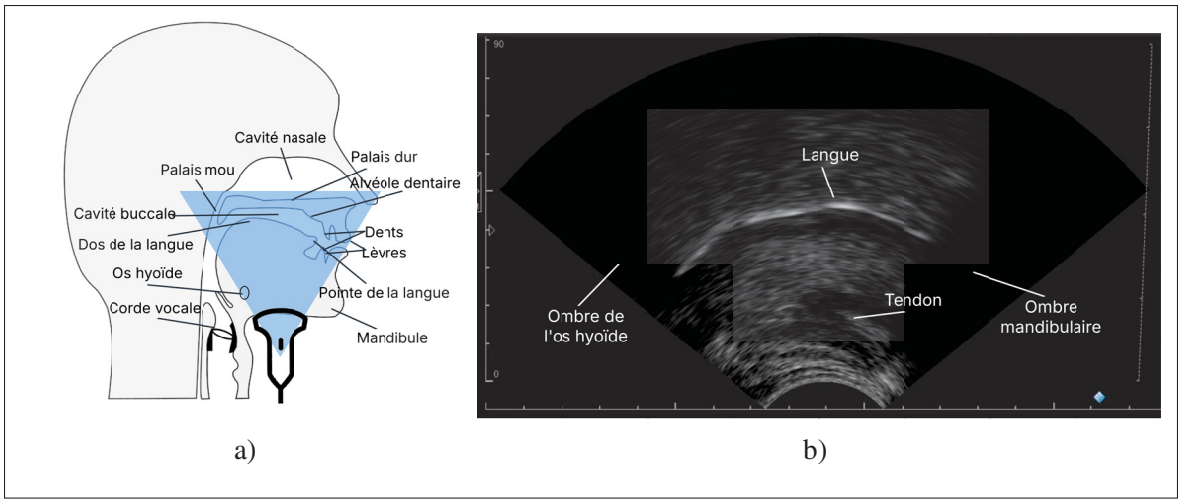


Figure 1.1 Création d'une image échographique du conduit vocal : (a) anatomie du conduit vocal ; (b) image échographique correspondante

constrictions ou des obstructions complètes du flux d'air en se rapprochant ou en entrant en contact avec d'autres articulateurs.

Le palais dur, la structure osseuse formant le toit de la bouche, constitue un point de contact rigide et essentiel pour de nombreuses consonnes. Les sons alvéolaires, tels que /t/, /d/ et /s/, sont produits lorsque l'apex ou la lame de la langue entre en contact ou se rapproche de la crête alvéolaire, juste derrière les dents supérieures. Les sons post-alvéolaires, comme /ʃ/ (ch) et /ʒ/ (j), impliquent une constriction légèrement plus en arrière. La forme et la voûte du palais dur influencent plutôt les stratégies articulatoires adoptées par les locuteurs. En particulier, (Brunner, Fuchs & Perrier, 2009) ont montré que les différences de morphologie palatine modifient la variabilité et l'organisation des mouvements de la langue.

En plus de la langue et du palais dur, la vue sagittale permet de visualiser, dans certaines conditions, le palais mou (velum) et l'ombre acoustique projetée par l'os hyoïde. Une autre structure clé souvent visible est le tendon court du muscle génioglosse, qui sert de point d'insertion pour les faisceaux musculaires et se connecte à la mandibule.

1.2 Principes de l'Imagerie articulatoire par ultrasons

1.2.1 Fondamentaux de l'imagerie ultrasonore

L'imagerie ultrasonore repose sur la réflexion des ondes aux interfaces entre milieux d'impédances acoustiques différentes. Lorsqu'une onde ultrasonore traverse le corps et rencontre une frontière entre deux tissus aux propriétés acoustiques distinctes (comme entre le tissu mou et l'air, ou entre le muscle et l'os), une partie de son énergie est réfléchie vers le transducteur tandis que le reste se propage dans le second milieu. L'intensité de ces échos réfléchis dépend de l'ampleur de la différence d'impédance acoustique entre les deux milieux : plus cette différence est marquée, plus la réflexion est forte. Un transducteur émet ces ondes sonores et reçoit les échos qui rebondissent sur les structures internes. Ces échos sont ensuite convertis en signaux électriques et traités par un ordinateur afin de créer des images. Cette méthode fournit des séquences vidéo en temps réel des mouvements de la langue, qui sont faciles à acquérir et adaptées à un large éventail de sujets, y compris les enfants et les patients (Epstein & Stone, 2005). Cette méthode est non invasive et n'altère pas l'articulation des sujets, ce qui la rend préférable à d'autres techniques plus invasives (Ouni, 2014).

L'imagerie par ultrasons du conduit vocal fournit généralement différentes vues, la vue sagittale étant la plus fréquemment utilisée et l'une des plus informatives. Comme illustré dans la Figure 1.1, pour l'imagerie du conduit vocal, une sonde est placée sous le menton et orientée vers le haut. Le faisceau ultrasonore traverse les tissus mous du plancher buccal et de la langue (Figure 1.1a). Lorsqu'il atteint l'interface entre le tissu de la langue et l'air dans la cavité buccale, la différence d'impédance acoustique entre ces deux milieux provoque une forte réflexion des ondes. Ces échos sont captés par le transducteur et convertis en une image bidimensionnelle (Figure 1.1b) en temps réel, où l'interface air-tissu apparaît comme une ligne blanche et brillante correspondant à la surface supérieure de la langue (Epstein & Stone, 2005).

Cependant, cette technique présente des limitations physiques intrinsèques. Le faisceau ultrasonore est bloqué par les structures osseuses, créant des ombres acoustiques. L'os de

la mandibule (mâchoire inférieure) à l'avant et l'os hyoïde à l'arrière masquent respectivement la pointe antérieure et la racine postérieure de la langue. Par conséquent, le contour visible de la langue sur une image ultrasonore est généralement incomplet (Epstein & Stone, 2005; Stone, 2005).

La validité et la reproductibilité des données dépendent de manière critique de la stabilisation de la sonde. Tout mouvement du transducteur par rapport à la tête du locuteur peut modifier le plan d'imagerie et fausser les mesures. Des dispositifs de stabilisation, tels que des casques ou des supports personnalisés, sont donc nécessaires pour garantir la comparabilité des données au sein d'une même session et entre différentes sessions d'enregistrement (Stone, 2005). De plus, le palais est généralement invisible dans les images ultrasonores en raison de la présence d'air entre celui-ci et la langue, ce qui sera discuté plus en détail dans la section 1.4.2.

1.2.2 L'imagerie échographique comme outil de biofeedback visuel pour l'apprentissage des langues

La technologie ultrasonore émerge comme un outil de biofeedback visuel prometteur dans l'apprentissage d'une L2, permettant aux apprenants d'observer en temps réel la forme et le mouvement de leur langue pendant la production de la parole, offrant ainsi une représentation visuelle des gestes articulatoires autrement invisibles (Gick *et al.*, 2008). Cette approche non invasive a démontré son efficacité particulière pour l'acquisition de sons difficiles, comme les voyelles mandarines /u/ et /y/ par des locuteurs anglophones (Li, Ayala, Harel, Shiller & McAllister, 2019) ou les voyelles danoises par des locuteurs français (Kartushina, Hervais-Adelman, Frauenfelder & Golestani, 2015).

Contrairement au biofeedback auditif seul, l'échographie linguale fournit une information spatiale directe sur la position et le mouvement de la langue. Ce retour visuel s'avère particulièrement bénéfique pour les apprenants ayant une faible conscience phonologique ou une capacité perceptive limitée pour les sons cibles, ou pour les sons dont l'articulation ne peut être facilement déduite de l'écoute (Li *et al.*, 2019; Aalto, Ben Asker, Ménard, Cardoso & Laporte, 2025b).

Les apprenants bénéficient d'une amélioration significative de leur production articulatoire, et ce même avec des interventions relativement brèves (Song & Eckman, 2021; Aalto *et al.*, 2025b). L'efficacité du biofeedback échographique dépend toutefois du profil de l'apprenant et de l'exposition préalable à la technologie, suggérant la nécessité d'adapter les protocoles d'intervention aux caractéristiques individuelles (Li *et al.*, 2019; Aalto *et al.*, 2025b).

Lorsque le biofeedback ultrasonore est combiné avec des instructions auditives et des exercices traditionnels, il crée un environnement d'apprentissage multisensoriel qui renforce à la fois la représentation auditive et visuelle des gestes articulatoires (Aalto *et al.*, 2025b). Les études récentes indiquent également que les apprenants perçoivent cette méthode comme motivante et utile, renforçant l'engagement dans le processus d'apprentissage (Bryfonski, 2023).

1.3 L'analyse automatique des images échographiques du conduit vocal

L'analyse automatique des images échographiques constitue un domaine de recherche en pleine expansion, avec des applications croissantes en phonétique, en orthophonie et dans l'apprentissage des langues. La plupart des techniques d'analyse automatique développées à ce jour se concentrent sur l'extraction du contour de la langue, cette dernière étant l'articulateur principal dans la production de la parole. Cependant, pour une analyse articulatoire complète, l'étude d'autres structures du conduit vocal, notamment le palais dur, est tout aussi importante. Contrairement à la langue dont le contour est généralement visible dans les images échographiques, le palais présente des défis uniques d'analyse automatique en raison de son invisibilité acoustique intermittente causée par l'interface air-tissu. Les méthodes développées pour le suivi de la langue offrent néanmoins des fondations algorithmiques précieuses qui peuvent être adaptées ou étendues pour aborder les défis spécifiques liés au suivi du palais. Cette section examine les principales techniques de suivi de langue.

1.3.1 Techniques algorithmiques fondamentales

La première vague d'automatisation du suivi de la langue reposait sur des algorithmes qui tentaient de modéliser explicitement les propriétés d'un contour (continuité, régularité) et de l'image (gradients, intensité).

1.3.1.1 Modèles de contours actifs (Snakes)

Introduits dans ce domaine par Akgul, Kambhamettu & Stone (1999), les contours actifs, ou snakes, sont des splines déformables qui minimisent une fonction d'énergie. Cette énergie est une combinaison de deux forces : des forces internes qui maintiennent la régularité du contour (sa continuité et sa courbure) et des forces externes qui l'attirent vers les caractéristiques saillantes de l'image, comme les bords à fort gradient. Le système EdgeTrak (Li, Kambhamettu & Stone, 2005), une implémentation bien connue, a amélioré ce modèle en ajoutant des contraintes liées à l'intensité des pixels.

Cependant, les snakes souffrent de limites importantes : ils sont très sensibles à leur configuration initiale et peuvent facilement se coincer dans des minima locaux, échouant à trouver le véritable contour si l'initialisation est trop éloignée ou si l'image est trop bruitée.

1.3.1.2 Méthodes basées sur les filtres particulaires

Pour surmonter la sensibilité des snakes, Laporte & Ménard (2018) ont proposé une approche basée sur un filtre particulaire. Au lieu de suivre une seule hypothèse pour le contour de la langue, cette méthode en maintient plusieurs simultanément (les particules). Chaque particule représente un état possible de la langue (position, échelle, forme globale). À chaque nouvelle image, ces hypothèses sont propagées et leur vraisemblance est évaluée par rapport aux données visuelles. Les hypothèses peu probables sont éliminées et les plus probables sont multipliées. Ce mécanisme multi-hypothèses permet au système de se remettre beaucoup plus facilement des erreurs de suivi, car même si l'hypothèse la plus probable devient erronée, d'autres hypothèses

correctes peuvent subsister et prendre le relais. Cette approche s'est avérée particulièrement robuste aux mouvements rapides de la langue et à la dégradation de la qualité de l'image.

1.3.1.3 Stratégies pour la robustesse

D'autres techniques ont été développées pour améliorer la fiabilité des trackers sur de longues séquences. La réinitialisation automatique est une stratégie clé, où le système détecte les moments où la langue retourne à une position de repos (par exemple, en comparant la similarité de l'image actuelle avec la première image de la séquence) et utilise le contour initial pour réinitialiser le tracker, corrigeant ainsi les erreurs accumulées (Xu *et al.*, 2016; Karimi, Ménard & Laporte, 2019).

La cohérence temporelle, explorée par Tang, Bressmann & Hamarneh (2012), consiste à optimiser le contour non pas sur une seule image, mais sur une fenêtre de plusieurs images (passées, présentes et futures). Cela garantit que l'évolution du contour est lisse et physiquement plausible dans le temps, évitant les sauts brusques.

Enfin, les modèles de forme actifs et les modèles d'apparence actifs utilisent un modèle statistique de la variation de la forme de la langue, appris à partir d'un ensemble de données d'entraînement (Roussos, Katsamanis & Maragos, 2009). Ce modèle de forme contraint le suivi à des configurations plausibles et peut même être utilisé pour extrapoler les parties du contour qui sont temporairement invisibles.

1.3.2 Techniques d'apprentissage profond

1.3.2.1 Architectures pour la segmentation (CNNs, U-Net)

Le problème du suivi de contour peut être reformulé comme un problème de segmentation sémantique : classer chaque pixel de l'image comme appartenant ou non au contour de la langue. L'architecture U-Net (Ronneberger, Fischer & Brox, 2015), conçue à l'origine pour la segmentation d'images biomédicales, s'est avérée exceptionnellement efficace pour cette tâche.

Sa structure en encodeur-décodeur avec des connexions résiduelles (skip connections) lui permet de capturer à la fois le contexte global de l'image et les détails fins de localisation, ce qui est crucial pour délimiter précisément le contour de la langue (Chen, Tiede & Whalen, 2020).

1.3.2.2 Systèmes hybrides (DeepEdge)

Reconnaissant que les méthodes traditionnelles et l'apprentissage profond ont des forces complémentaires, des systèmes hybrides comme DeepEdge (Chen *et al.*, 2020) ont été développés. Dans cette approche, un réseau de neurones (U-Net) est d'abord utilisé pour sa robustesse au bruit afin d'identifier une région d'intérêt (ROI) approximative autour du contour de la langue. Ensuite, un algorithme de snake est appliqué à l'intérieur de cette ROI pour affiner le contour avec une plus grande précision. Cette stratégie combine la capacité de reconnaissance de haut niveau du modèle d'apprentissage profond avec la précision de détection des bords de l'algorithme classique.

1.3.2.3 Modèles d'estimation de pose (DeepLabCut)

Une approche alternative consiste à abandonner complètement l'idée de segmenter un contour continu. Des outils comme DeepLabCut (Mathis *et al.*, 2018), issus du domaine de l'analyse du comportement animal, sont entraînés à localiser et à suivre un ensemble de points-clés anatomiques discrets (par exemple, la pointe, le corps, la racine de la langue) (Wrench & Balch-Tomes, 2022). Le contour de la langue est ensuite simplement interpolé entre ces points-clés suivis. L'avantage majeur de cette méthode est qu'elle ne dépend pas de la présence d'un bord clair et continu dans l'image, ce qui la rend extrêmement robuste au bruit et aux occlusions partielles.

1.3.3 Évaluation de la performance et analyse d'erreurs

L'évaluation rigoureuse des algorithmes de suivi est importante pour valider leur utilité. La métrique la plus couramment utilisée est la Somme Moyenne des Distances (MSD) (Li *et al.*,

2005). Elle calcule la distance moyenne entre les points d'un contour généré automatiquement et les points les plus proches sur un contour de référence (vérité terrain), et vice versa. Cette métrique est particulièrement adaptée car elle peut comparer des contours qui n'ont pas le même nombre de points. L'Erreur Quadratique Moyenne (RMSE) est également utilisée (Csapó & Lulich, 2015), mais elle nécessite une correspondance point à point. Au-delà de la simple distance, des métriques basées sur la forme peuvent évaluer si des caractéristiques linguistiquement pertinentes, comme la courbure maximale, sont préservées.

La vérité terrain est généralement établie par des tracés manuels réalisés par des experts. Cependant, ces tracés manuels présentent eux-mêmes une variabilité non négligeable, avec des écarts-types pouvant aller de 0.95 mm à 2.11 mm dans une étude (Csapó & Lulich, 2015).

1.4 Le palais dans les images échographiques

Alors que le suivi de la langue a attiré l'essentiel de l'attention algorithmique, la délinéation du palais dur représente un défi tout aussi crucial, bien que différent. Si le contour de la langue est une carte dynamique du terrain articulatoire, le contour du palais en est le cadre de référence anatomique important.

1.4.1 Le palais comme référence anatomique

L'importance du suivi du palais est triple. Premièrement, il sert à désambiguïser la position de la surface de la langue. Dans certaines articulations, comme lors de la production d'une consonne occlusive (/t/, /k/), la langue touche le palais. Dans ce cas, le faisceau ultrasonore traverse la langue et se réfléchit sur l'os palatin. L'écho brillant qui en résulte correspond au palais, et non à la surface de la langue. Connaître la position du palais permet d'interpréter correctement de telles images (Epstein & Stone, 2005). Deuxièmement, le palais fournit un cadre de référence fixe dans l'espace crânien. Comme le palais est une structure rigide, son contour peut être utilisé pour aligner et recaler des images, que ce soit au cours d'une même session (pour corriger les mouvements de la sonde) ou entre différents locuteurs (pour normaliser les différences

anatomiques). Cela transforme des mesures relatives en mesures quasi absolues (Epstein & Stone, 2005). Troisièmement, la connaissance simultanée des contours de la langue et du palais permet de calculer des mesures phonétiquement significatives, telles que l'emplacement et le degré de la constriction la plus étroite dans le conduit vocal, des paramètres essentiels pour caractériser les voyelles et les consonnes fricatives. Enfin, en biofeedback visuel, le suivi automatique du palais fournit aux apprenants une cible anatomique stable et intuitive qui réduit considérablement la charge cognitive nécessaire à l'interprétation des images échographiques. En rendant visible la relation spatiale entre la langue et le palais en temps réel, il améliore la perception des gestes articulatoires et renforce l'efficacité pédagogique du biofeedback, particulièrement pour l'apprentissage des sons de L2 et la rééducation des troubles de l'articulation. Une étude pilote récente montrent que les apprenants bénéficient d'une meilleure conscience proprioceptive et d'une acquisition plus rapide des mouvements articulatoires lorsqu'ils peuvent visualiser leur langue par rapport à une référence palatine stable (Aalto, Ben Asker, Farazi, Ménard & Laporte, 2025a).

1.4.2 Défis de la visualisation du palais

Dans les images échographiques, capturer l'ensemble du palais constitue un défi en raison de l'air qui le sépare de la langue, ce qui entraîne la réflexion du faisceau ultrasonore à l'interface langue-air au lieu d'atteindre le palais (Epstein & Stone, 2005). Le palais reste principalement invisible sauf pendant la déglutition, et même dans ce cas, il n'est souvent pas visible dans son entièreté dans une seule image (Faucher, Karimi, Ménard & Laporte, 2019). De plus, des fragments du palais peuvent être aperçus lorsque la langue entre en contact avec lui, ce qui se produit lors de certains sons de la parole et permet au faisceau ultrasonore de traverser le tissu mou et de se réfléchir sur l'os palatin (Epstein & Stone, 2005).

1.4.3 Méthodologies pour la délinéation du palais

La solution pour contourner le problème de l'invisibilité du palais consiste à éliminer la couche d'air en demandant au locuteur d'effectuer une action où la langue entre en contact acoustique

avec le palais : une déglutition (Epstein & Stone, 2005). Pendant la déglutition d'un liquide ou de salive, la cavité entre la langue et le palais se remplit de liquide qui remplace l'air, permettant ainsi aux ondes ultrasonores de traverser cette interface et de se réfléchir sur l'os palatin. Par conséquent, la langue balaie le palais de l'avant vers l'arrière, rendant progressivement visible le contour palatin sur plusieurs images vidéo successives.

La méthode originale, décrite par Epstein & Stone (2005), consiste à visionner la vidéo d'une déglutition et à tracer manuellement les segments du palais au fur et à mesure de leur apparition, puis à les assembler pour former un contour complet.

Pour automatiser ce processus laborieux, Faucher *et al.* (2019) ont proposé une méthode basée sur une image squelette d'échos cumulés (CES). Cette technique consiste à superposer et à traiter toutes les images de la séquence de déglutition pour créer une seule image composite où les échos persistants, correspondant au palais, sont renforcés. Un algorithme peut ensuite extraire le contour de cette image composite. Les résultats ont montré une bonne concordance avec les tracés manuels, avec une erreur moyenne inférieure à 3 mm.

1.4.4 Le suivi du palais

1.4.4.1 Motivation

Le suivi automatisé du palais représente une contribution novatrice à l'imagerie ultrasonore linguale. Dans l'acquisition d'une L2, le suivi automatisé du palais pourrait améliorer l'entraînement par biofeedback échographique en fournissant aux apprenants une cible anatomique stable et intuitive, réduisant ainsi la charge cognitive nécessaire à l'interprétation de l'image. Des données récentes suggèrent que les tracés du palais déplacés et superposés manuellement pendant les séances de biofeedback échographique améliorent la perception spatiale des apprenants quant aux relations entre la langue et le palais et renforcent l'efficacité pédagogique du biofeedback échographique lingual (Aalto *et al.*, 2025a). Alors que le traçage fournit un contour statique du palais, le suivi permettrait une localisation continue, image par image, du palais, ce qui est

essentiel pour un enregistrement, une normalisation et une mesure précise des mouvements de la langue par rapport au palais dans le temps et au cours des tâches. Les méthodes de suivi automatisées réduiraient encore davantage le travail manuel, amélioreraient la reproductibilité et faciliteraient les études à grande échelle ou en temps réel, alors que le traçage manuel est laborieux, sujet à l'erreur humaine et limité à des moments spécifiques où le palais est visible.

1.4.4.2 Défis de suivi du palais

Le suivi continu du palais en imagerie échographiques présente deux défis fondamentaux : son invisibilité intermittente et les déplacements apparents causés par les mouvements relatifs entre la tête et la sonde ainsi que par les mouvements de la mâchoire.

Le premier obstacle majeur est d'ordre physique. L'interface d'air entre la langue et le palais crée une barrière acoustique qui empêche la transmission des ondes ultrasonores. Par conséquent, le palais reste invisible à l'image, sauf lors d'un contact direct entre la langue et le palais ou lorsque la cavité buccale est remplie d'un milieu de couplage acoustique (comme de l'eau) qui permet la propagation des ondes.

Pour garantir la fiabilité d'un système de suivi, il est donc crucial de définir les conditions optimales pour l'acquisition des contours du palais. La démarche pour y parvenir consiste à établir des pratiques de référence pour le traçage manuel en évaluant la cohérence et l'accord entre plusieurs annotateurs experts (accord inter-juges), puis à valider la précision et la fiabilité des méthodes automatiques, comme celle basée sur le squelette d'échos cumulés, en les comparant à ces annotations manuelles qui servent de référence.

Le second défi concerne le mouvement. Bien que le palais soit une structure anatomiquement rigide, sa position dans le flux d'images ultrasonores est instable. Cette variabilité est due à plusieurs facteurs externes, notamment le déplacement du transducteur sous le menton, les mouvements de la tête du locuteur, ainsi que les ajustements de la mâchoire (mouvements mandibulaires) durant la parole.

Ces mouvements combinés compliquent le suivi continu du palais. Pour surmonter ce problème, une solution consiste à s'appuyer sur des repères anatomiques qui, eux, restent détectables de manière constante dans l'image. En modélisant la relation entre le déplacement de ces repères et le mouvement apparent du palais, il devient possible de développer une méthode automatisée pour suivre la position du palais de manière continue et fiable, même lorsqu'il est temporairement invisible.

1.5 Technologies alternatives pour les suivi du palais

Bien que l'imagerie échographique constitue une méthode privilégiée pour l'étude de la parole en raison de son caractère non invasif et de sa haute résolution temporelle, d'autres technologies offrent des capacités de suivi du palais qui méritent d'être examinées.

L'EMA permet le suivi en temps réel des mouvements articulatoires en trois dimensions grâce à des capteurs électromagnétiques fixés sur la langue et d'autres articulateurs. L'EMA peut être combinée avec des modèles 3D du palais obtenus par empreintes dentaires et balayage 3D pour améliorer la précision spatiale (Nota, Kitamura, Takemoto & Maekawa, 2024; Kochetov, 2020). L'avantage principal de cette approche réside dans sa capacité à suivre simultanément la langue et le palais avec une haute résolution temporelle, permettant ainsi des analyses précises des relations spatiales et temporelles entre ces structures. Cette synchronisation est utile pour l'étude des phénomènes de coarticulation et pour le développement de modèles articulatoires précis.

L'EPG utilise un faux palais artificiel équipé d'électrodes pour enregistrer les contacts entre la langue et le palais dur. Cette méthode excelle dans la visualisation des schémas de contact en temps réel, offrant une précision spatiale et temporelle pour analyser les articulations consonantiques (Hardcastle, Gibbon & Jones, 1991; Lee *et al.*, 2023). L'EPG fournit un retour visuel immédiat qui est particulièrement efficace en rééducation orthophonique, notamment pour les troubles de la parole associés à la fente palatine ou à la surdité (Patrick, Fricke, Rutter & Cleland, 2024; Verhoeven, Miller, Daems & Reyes-Aldasoro, 2019). Son principal avantage analytique

réside dans sa capacité à objectiver les contacts articulatoires qui seraient autrement invisibles à l'observation clinique standard, réduisant ainsi la subjectivité de l'évaluation perceptive.

L'IRM, particulièrement l'IRM en temps réel, offre une visualisation complète du conduit vocal, incluant le palais dur et mou, la langue et les structures pharyngées (Toutios, Byrd, Goldstein & Narayanan, 2019; Lin, 2021). Contrairement aux autres méthodes, l'IRM capture à la fois les structures fixes et mobiles dans leur contexte anatomique complet, permettant des analyses 3D détaillées des formes et des mouvements. Cette perspective holistique est inestimable pour comprendre comment la morphologie individuelle du palais influence les stratégies articulatoires (Brunner *et al.*, 2009), et pour modéliser précisément la dynamique du conduit vocal lors de la production de la parole.

Malgré leurs avantages, ces technologies présentent des limitations qui justifient le développement de méthodes de suivi du palais basées sur l'échographie. L'EMA nécessite l'attachement de capteurs dans la bouche, ce qui est intrusif et peut altérer l'articulation naturelle. L'EPG exige le port d'un faux palais artificiel, ce qui modifie la configuration articulatoire et empêche l'étude du palais dans sa configuration anatomique naturelle. L'IRM, quant à elle, est coûteuse, peu portable et présente une résolution temporelle limitée par rapport à l'échographie.

Une solution de suivi du palais en échographie offrirait plusieurs avantages. Premièrement, elle maintiendrait le caractère non invasif de l'échographie, préservant ainsi l'articulation naturelle sans altération par des dispositifs intrusifs. Deuxièmement, elle permettrait une plus grande accessibilité clinique et pédagogique, l'équipement échographique étant significativement moins coûteux et plus portable que les systèmes EMA ou IRM. Troisièmement, elle faciliterait l'intégration dans des systèmes de biofeedback en temps réel pour l'apprentissage des langues secondes et la rééducation orthophonique, sans nécessiter des procédures de calibration complexes ou des dispositifs supplémentaires.

CHAPITRE 2

BEST PRACTICES FOR TRACING THE PALATE IN ULTRASOUND IMAGES

Hana Ben Asker[1] , Eija M.A. Aalto[2] , Lucie Ménard[3] , Walcir Cardoso[4] , Catherine Laporte[2]

[1] Département de Génie des technologies de l'information, École de Technologie Supérieure,
[2] Département de Génie Électrique, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[3] Département de Linquistique, Université de Québec à Montréal,
405 Sainte-Catherine Est, Montréal, Québec, Canada H2L 2C4
[4] Département d'éducation, Université Concordia,
1455 Maisonneuve Ouest, Montréal, Québec H3G 1M8

Article accepté aux actes de la conférence internationale « EMBC », février 2025.

2.1 Abstract

Ultrasound (US) imaging is a powerful tool for visualizing the vocal tract and tongue motion during speech articulation. A key aspect of analyzing tongue motion is accurately tracing the palate's contour in US images, as it provides a fixed anatomical reference for measuring tongue position and deformation. However, accurately tracing the palate in US images remains challenging due to limited visibility caused by the air gap between the palate and the tongue. To address this limitation, we evaluate the reproducibility of manual palate tracing using methods that rely on varying levels of assistance from image-enhancement algorithms. One of these methods is the cumulative echo skeleton (CES), which enhances video frames to stack and reconstruct palate echoes. These methods are tested across different swallowing tasks.Results indicate that CES-based methods enhance rater agreement, primarily due to the cumulation of echoes. Among the tasks, dry swallow consistently yields higher agreement across methods. Additionally, the CES-based automatic method was benchmarked against manual annotations, showing promising accuracy in dry swallow with a mean sum of distances of 2.63 mm. These findings emphasize the important role of method and task selection in enhancing reproducibility and highlight the potential of automated approaches for palate tracing in US imaging.

Clinical relevance— In Speech-Language Pathology (SLP), the application of real-time US imaging can be used in diagnostics and treatment of Speech Sound Disorders (SSD). Furthermore, it facilitates observations of oral movements like swallowing and mastication. The incorporation of a static image of the palate—an anatomical structure that is both familiar and tangible within the oral cavity—provides a valuable reference point for interpreting the dynamic movements of the tongue. This visual aid enhances the understanding of real-time ultrasound images, benefiting both the clinician and the client.

2.2 Introduction

Ultrasound (US) imaging is widely recognized for its non-invasive visualization of soft tissues. In SLP, US is used to study typical and atypical speech sound acquisition (Cleland & Scobbie, 2021) by providing detailed insights into articulatory movements in the vocal tract. It also serves as an effective biofeedback tool in SSD (Sugden, Lloyd, Lam & Cleland, 2019), enabling individuals to correct articulation errors and improve speech production (Cleland *et al.*, 2015a). As illustrated in Figure 2.1, the transducer is placed under the chin to capture a midsagittal view of the vocal tract.

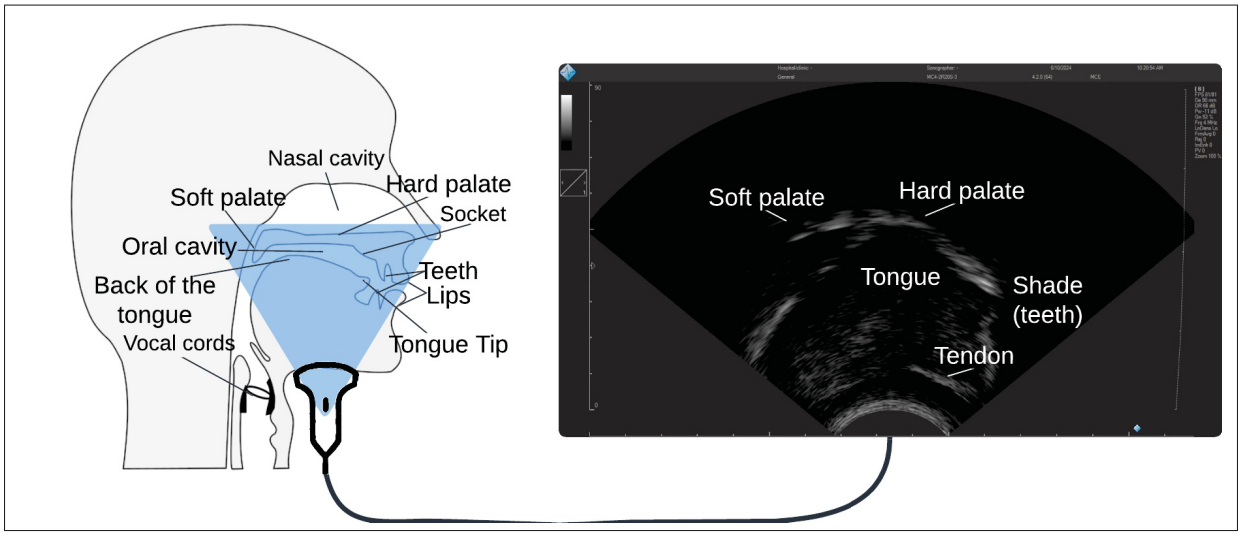


Figure 2.1 US imaging of the vocal tract

Beyond tongue visualization, delineating the palate (the bony structure forming the roof of the mouth) can provide important insights into articulation. This process enables the computation of phonetically significant measures, facilitates image registration within and between participants, and helps disambiguate the tongue surface from surrounding structures (Epstein & Stone, 2005). Additionally, the palate's shape influences articulatory behavior, impacting speech production and phonetic variations (Brunner *et al.*, 2009). By serving as a rigid reference, the palate enables consistent registration of tongue shape data over time, improving the precision of articulatory measurements. In biofeedback, the visualization of the palate provides participants with a clear reference point, helping them to relate their tongue placement to their palate, and thus aiding the acquisition of new articulatory movements. Biofeedback gives a visual dimension to speakers' tactile and proprioceptive awareness of their articulatory movements.

However, visualizing the palate using US presents challenges, as the air gap between the tongue and the palate causes the US waves to reflect back to the transducer, preventing the capture of the palatal contour (Epstein & Stone, 2005). As a result, the palate is not visible in US images (Figure 2.2a) and is only partially discernible when tongue-palate contact occurs (Figure 2.2b). Even in such cases, its contour is often difficult to reconstruct from a single frame (Faucher *et al.*, 2019; Epstein & Stone, 2005).

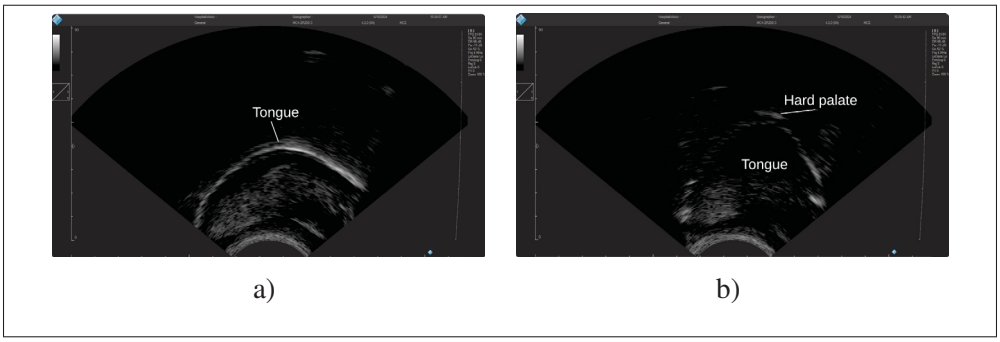


Figure 2.2 US images showing (a) the tongue with an invisible palate due to the air gap between the tongue and palate, and (b) partial visibility of the palate during /k/ sound

To enhance palate visibility in US images, researchers have employed techniques that involve introducing media into the oral cavity, such as yogurt or gelatin, which improve the visibility of the palate in US images (Mielke, Baker, Archangeli & Racy, 2005). Swallowing actions have also been used, where tongue-palate contact and acoustic coupling via saliva or other media enable US waves to be transmitted to and reflect off the palate (Epstein & Stone, 2005). Capturing videos during swallowing progressively reveals different sections of the palate as the tongue moves the bolus through the mouth. To optimize palate visualization, researchers recommend collecting various types of swallows, such as dry swallows, continuous swallows, and swallows involving boluses with varying consistencies, to identify the best possible palate images (Stone, 2005).

Traditionally, tracing the palate involves manual methods (Mielke *et al.*, 2005; Stone, 2005; Epstein & Stone, 2005). Operators connect visible palate segments across frames to reconstruct a complete contour, often combining traces from multiple images to achieve accuracy (Epstein & Stone, 2005). These manual approaches are labor-intensive and time-consuming, and little is known about their reliability and reproducibility.

Building on these manual techniques, the only fully automated approach for extracting the mid-sagittal palate contour from US swallowing videos was developed by (Faucher *et al.*, 2019). This technique uses a cumulative echo skeleton (CES) image to enhance consistently located features over time, connecting visible palate segments across frames into a continuous contour. Figure 2.3 illustrates the process of creating the CES image, showing how individual frames

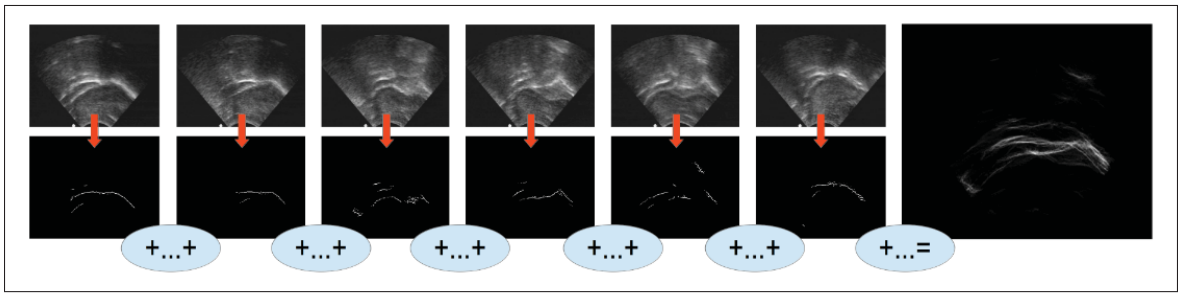


Figure 2.3 Creation of the cumulative echo skeleton (CES)
Tirée de Faucher et al., (2019)

contribute to the reconstruction of the palate. By reducing the reliance on manual tracing, this method paves the way for fully automatic tracing of the palate. However, the study did not investigate the method's performance across different swallowing tasks, such as water drinking or boluses with differing consistencies, which could impact palate visibility and delineation.

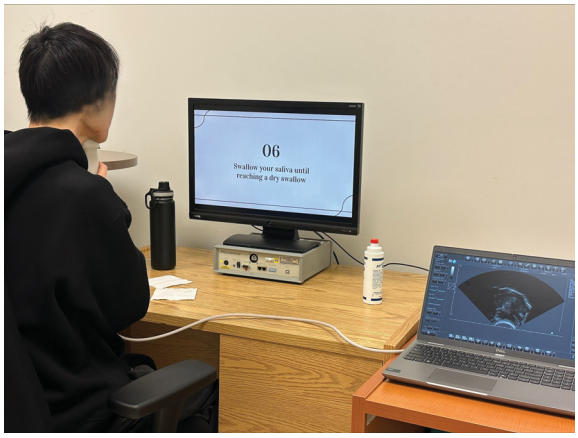
In this study, we seek to gain a better understanding of the conditions under which reliable palate tracings can be obtained from US videos of swallowing. Understanding these conditions is important for the design of research studies on articulation, as well as for the development of new articulatory biofeedback tools based on US. By evaluating the inter-rater agreement across three manual tracing methods and three swallowing tasks, this work identifies best practices for manual tracing. Moreover, it benchmarks the automatic CES-based tracing method against manual annotations, providing insights into its accuracy and reliability.

2.3 Methods

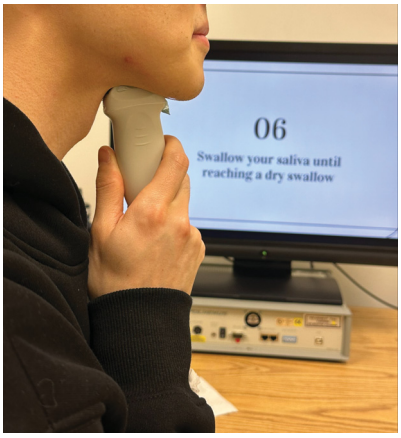
2.3.1 Data collection

Following approval by the Research Ethics Board at École de technologie supérieure, we recruited 17 volunteers with no known speech, language, or swallowing disorders for our study (7 male, 10 female ; age range : 21-51). The participants were imaged using a Telemed MicrUs EXT-1H US scanner equipped with a MC4-2R20S-3 microconvex probe (frequency range : 2–4 MHz, radius : 20 mm, scan depth : 90 mm). For each participant, we collected videos by positioning the US transducer under their jaw, which they held themselves (Figure 2.4b).

Participants performed three swallowing tasks : drinking water, drinking yogurt using a straw, and performing continuous swallows until reaching a dry swallow, at which point they they felt their tongue touch their palate. US videos were screen-recorded from Telemed's EchoWave II software using OBS Studio. From the full recordings, a total of 51 swallows (3 tasks per participant), were manually extracted from the video recordings for analysis.



a)



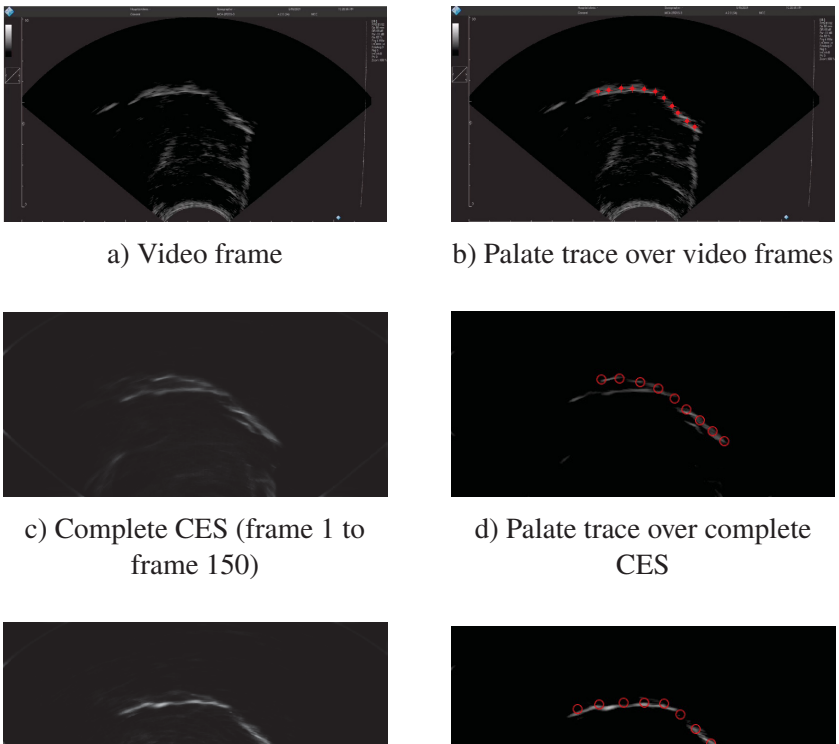
b)

Figure 2.4 (a) Data collection setup. (b) Participant positioning the US transducer under the jaw while following on-screen swallowing instructions

2.3.2 Manual palate tracing

The US video clips were annotated by three raters with moderate to extensive expertise in vocal tract US imaging. Each rater annotated the three tasks : dry swallow, water drinking, and yogurt, using the following three methods :

1. Video annotation : Raters were instructed to observe the entire video in a loop and trace the palate by drawing a single trace based on the observed palate throughout the video. This method is equivalent to what is shown in (Epstein & Stone, 2005). An example video frame is shown in Figure 2.5a, and the resulting palate trace is shown in Figure 2.5b.
2. CES annotation : For this method, raters were provided with the CES of the video, generated using the method described by (Faucher *et al.*, 2019). The CES (Figure 2.5c) combines all frames of the video, processed to enhance consistently located structures, such as the palate, while minimizing the enhancement of moving structures such as the tongue. Raters traced the palate based on the features highlighted in the CES, as shown in Figure 2.5d, over the entire range of video frames.



a) Video frame

b) Palate trace over video frames

c) Complete CES (frame 1 to frame 150)

d) Palate trace over complete CES

e) Selected CES (frame 61 to frame 110)

f) Palate trace over selected CES

Figure 2.5 Illustrations of manual palate tracing methods. Subfigures show examples of video frames, cumulative echo skeletons (CES), and selected CES, each with and without annotated palate traces

3. Selected CES annotation : In this method, raters first identified a restricted range of video frames from which to compute the CES, with the objective of maximizing palate visibility (Figure 2.5e). They then traced the palate based on this selected subset of frames, as illustrated in Figure 2.5f.

To assess intra-rater agreement, Rater 3 was randomly selected to repeat the annotation process three months after the initial session.

2.3.3 Automatic palate tracing

In addition to the manual traces, automatic palate traces were also obtained for each of the 51 swallowing video clips using Faucher et al.'s approach (Faucher *et al.*, 2019) with some adaptations to improve applicability. On each frame, the algorithm highlighted ridge-like structures that potentially corresponded to the palate. The screen-recorded videos contain more than just the US field of view (FOV), which causes spurious ridges to be identified.

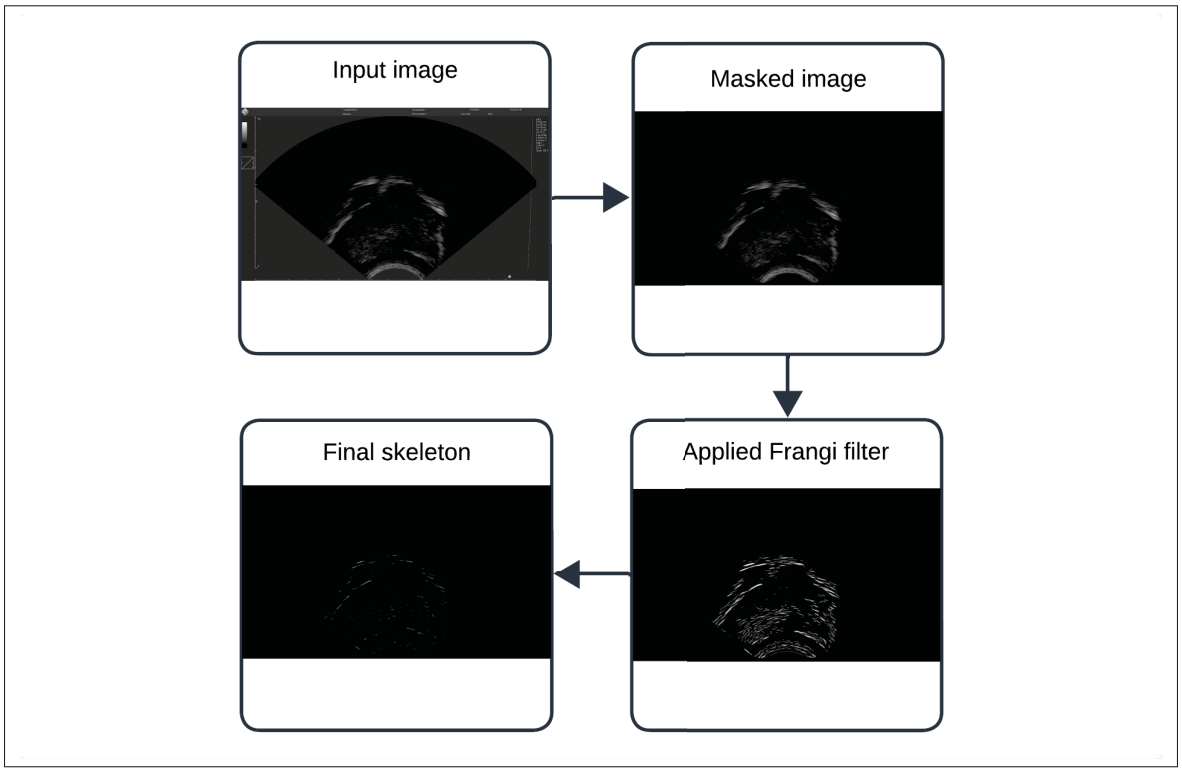


Figure 2.6 Individual frame processing

To address this, we dynamically segmented the US FOV for each frame by applying edge detection, dilation, and edge density techniques to isolate the largest dense region that predominantly corresponds to the US FOV. This approach takes advantage of the noisy nature of US imaging to identify the FOV. The ridge-like features were then enhanced using the phase symmetry filter (Kovesi *et al.*, 1999), as implemented in the prior work (Faucher *et al.*, 2019). To improve computational efficiency, as recommended by (Laporte, Takla, Tiede & Ménard, 2022), the

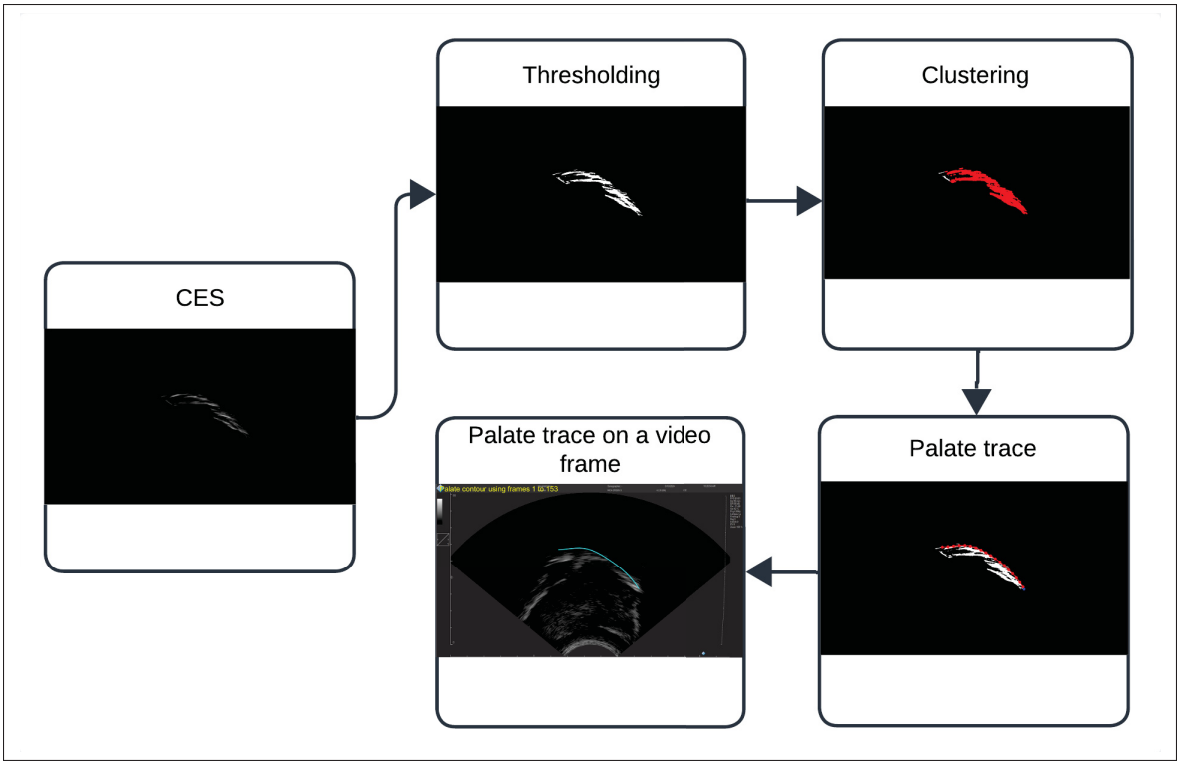


Figure 2.7 Palate trace extraction from CES

skeleton extraction step in the original implementation was replaced by Frangi filtering (Frangi, Niessen, Vincken & Viergever, 1998). This process is illustrated in Figure 2.6.

Processed frames are accumulated in a buffer over time to construct a CES, which emphasizes spatially consistent palate parts. This is followed by Otsu thresholding (Ostu, 1979) and DBSCAN clustering (Ester, Kriegel, Sander, Xu *et al.*, 1996). The widest cluster is chosen as the likely palate.

For each unique X-coordinate within the cluster, points along the Y-axis are refined by selecting the highest Y point. Figure 2.7 illustrates this process, including CES thresholding, clustering, palate tracing, and the resulting palate trace overlaid on a video frame. To ensure that the palate trace includes the alveolar ridge, the overall rightmost point among all cluster points, assumed to belong to the palate, is integrated into the trace (Figure 2.8).

Figure 2.8 Automatically traced palate (a) without and (b) with adding the rightmost point

This adjustment addresses cases where the accumulation process fails to emphasize that part. In such cases, the intensity in this region is too low, leading to its omission during CES thresholding or exclusion from the palate cluster during clustering.

All tests were run on a workstation equipped with an Intel Core Ultra 9-185H CPU (2.30 GHz, 32 GB RAM) under MATLAB R2023b. Over a 720 × 1280 video, the total runtime per sequence divided by the number of frames yields an average per-frame processing time of ≈ 1.3s.

2.4 Results & discussion

2.4.1 Rater annotation

In our study, for each participant–task–method combination, the three raters manually traced the palate, resulting in three curves. To quantify inter-rater agreement of the manual palate tracings, we used the mean sum of distances (MSD) metric. The MSD was computed as follows :

$$\mathrm{MSD}(u, v) = \frac{1}{m+n}\left(\sum_{i=1}^{n} \min_{j} \|\mathbf{v}_i - \mathbf{u}_j\| + \sum_{j=1}^{m} \min_{i} \|\mathbf{u}_j - \mathbf{v}_i\|\right) \tag{2.1}$$

where $u = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m\}$ and $v = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ represent the sets of points traced by each of two raters. The mean MSD over the 3 pairs of raters was then computed.

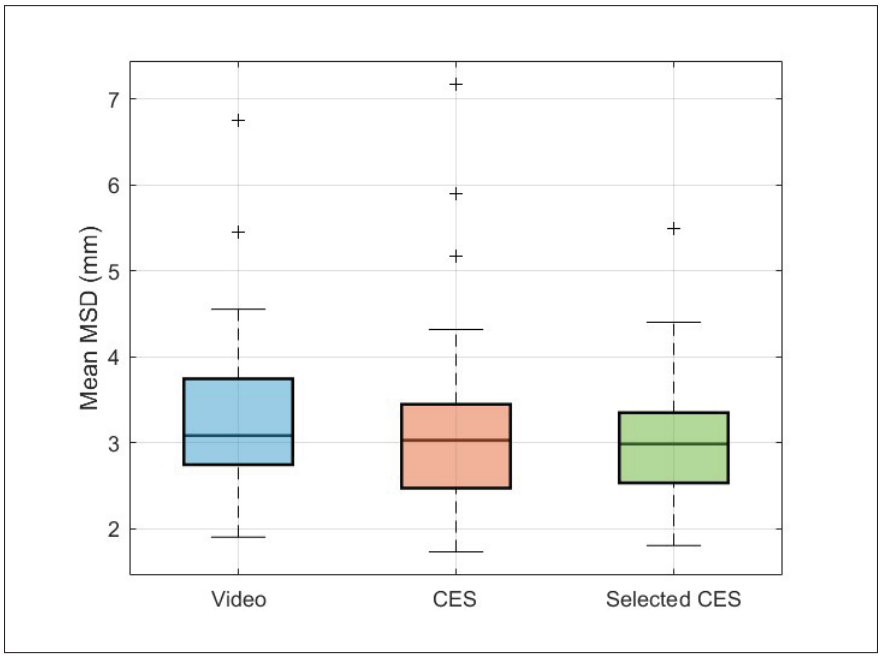


Figure 2.9 Mean MSD by method over the 51 swallowing videos

Figure 2.9 shows that the two CES-based methods demonstrate the strongest inter-rater agreement. While the Video method offers comparable median inter-rater agreement, it is comparatively less reliable. This can be attributed to the limitations of the Video method, where the complete visibility of the palate and the added uncertainty from probe movement may hinder consistent tracing. In contrast, the CES methods mitigate these challenges by accumulating echoes, which provide raters with a clearer and more stable representation of the palate, leading to better inter-rater agreement over a broader range of videos.

Figure 2.10 shows inter-rater agreement in palate tracings as a function of the different swallowing tasks. Here, dry swallow generally yields the lowest mean MSD, indicating better inter-rater agreement, while drinking water and yogurt lead to higher mean MSD values and more variation across participants.

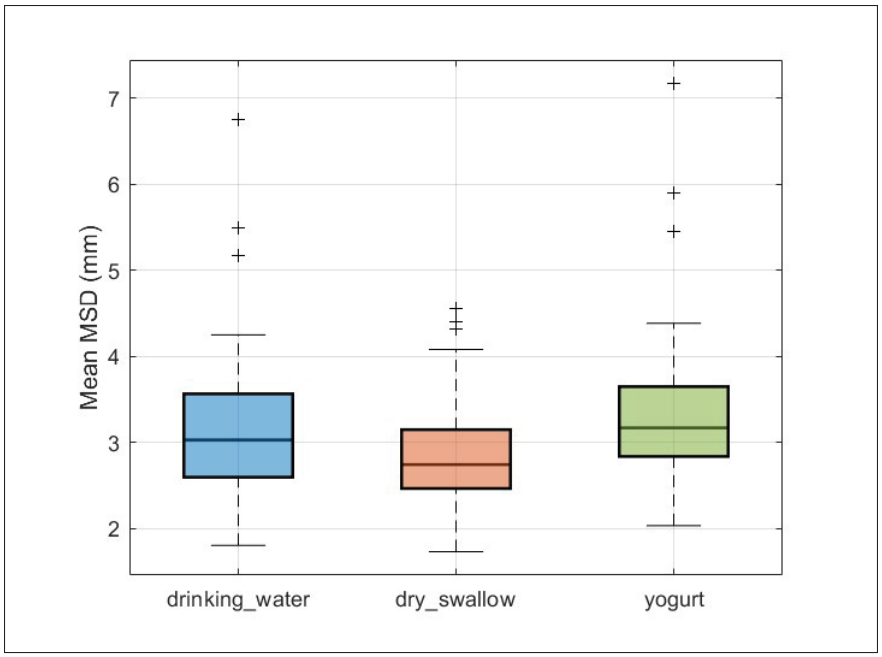


Figure 2.10 Mean MSD by task, over the 17 participants and 3 annotation methods

Figure 2.11 provides an overview of inter-rater reliability across different task and method combinations. The dry swallow task consistently leads to stronger inter-rater agreement when paired with the CES method. This can be attributed to the relative stability of the US probe during dry swallow, as the task also involves minimal jaw movement, allowing raters to achieve greater consensus on the palate trace. In contrast, tasks involving yogurt and water introduce more variability due to jaw movement, as participants open their mouths to intake the media and adjust to changes in bolus volume (empty mouth, filled mouth, then empty again). This variability affects the reproducibility of palate tracing in these tasks. In contrast, the CES method excels in video clips with probe or jaw movement where their CES relatively compensates for variability. The relative advantage of CES methods may diminish when probe stability is maintained, which can be achieved using a microphone stand (Chen, Whalen & Mok, 2024) or a stabilizing headset (Bryfonski, 2023).

To further explore the variability in inter-rater reliability, we computed the mean MSD by participant. Figure 2.12 illustrates these results. Low mean MSD values and narrow distributions,

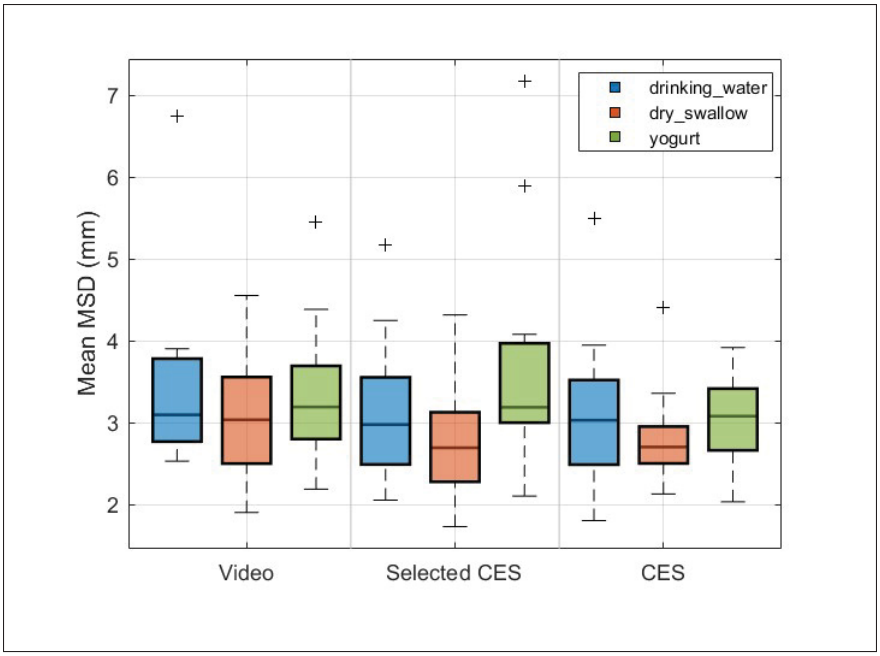


Figure 2.11 Mean MSD by task and method combination, over the 17 participants

indicating consistently strong inter-rater agreement, were obtained for some participants, whereas the opposite was true for others. This is due to factors such as image quality, probe movement, or task-specific difficulties. It is especially critical to consider the impact of poor imaging quality. Anatomical differences, suboptimal probe placement, or tissue composition may make key structures like the palate less distinct or distorted in US images, further complicating the task for raters. Figures 2.16a, 2.16b, 2.16d, and 2.16c illustrate the poor imaging quality and the resulting disagreement between raters in such cases.

Figures 2.13 and 2.15 examine the best and worst task-method combinations for each participant.

These results show that 11 out of 17 cases, the dry swallow task produced the best outcomes : 4 from CES, 4 from Selected CES, and 3 from Video (based on mean MSD).

The tracings from the top 4 best combinations are shown in Figure 2.14 (CES images are shown instead of Selected CES images, as raters selected different frame ranges). In Figures 2.14a and

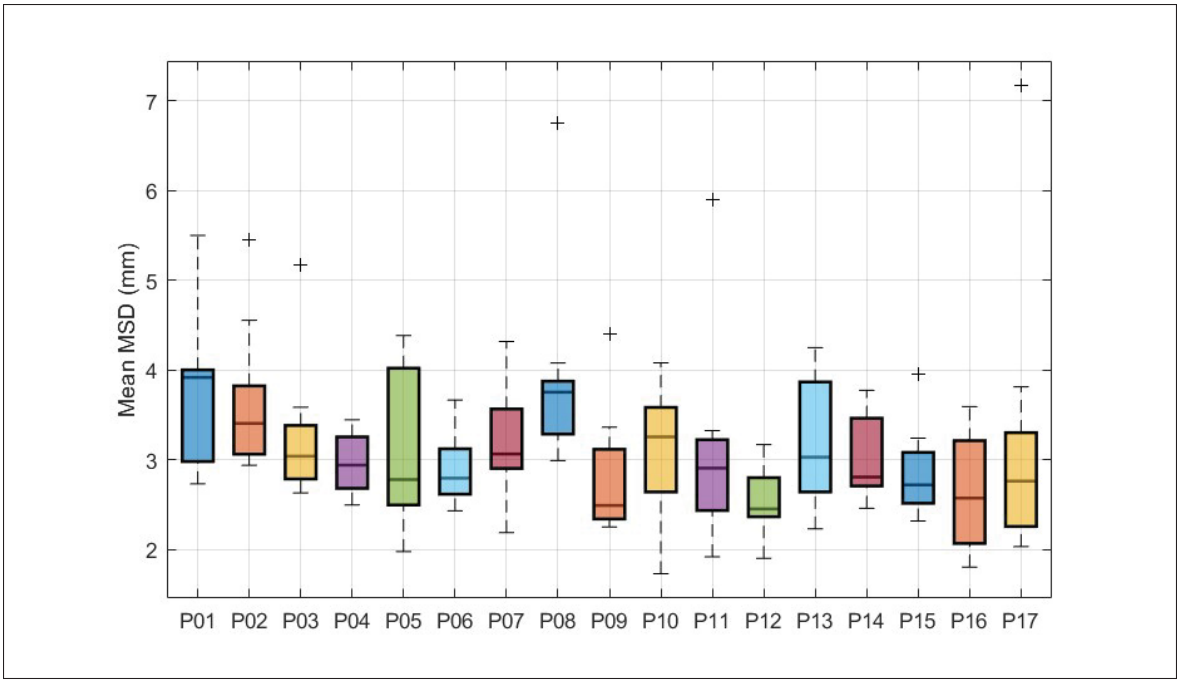


Figure 2.12 Mean MSD by participant, over the three tasks and three methods

2.14d, clear and consistent echoes are surrounding the palate trace. This clarity is attributed to the minimal probe movement during dry swallow, as the subject's jaw remains relatively stable. This is why selecting different frame ranges did not affect the level of agreement. In Figure 2.14b, despite strong echoes around the palate, the palate trace remains slightly visible, enabling raters to reach a consensus. Finally, in Figure 2.14c, the absence of probe movement during the dry swallow task ensures a stable visualization of the palate location across video frames, allowing for better rater agreement.

As shown in Figure 2.15, in 9 out of 17 cases, the worst inter-rater agreement is achieved with the yogurt task : 1 from CES, 3 from Selected CES, and 5 from Video (based on mean MSD). The tracings from the top 4 worst combinations are shown in Figure 2.16. Participant P17 with Selected CES and yogurt exhibits a low level of agreement due to multiple factors. First, the poor quality of the US recording impacts the construction of the CES. Additionally, probe movement during the task leads to inconsistencies, as raters who selected different frame ranges

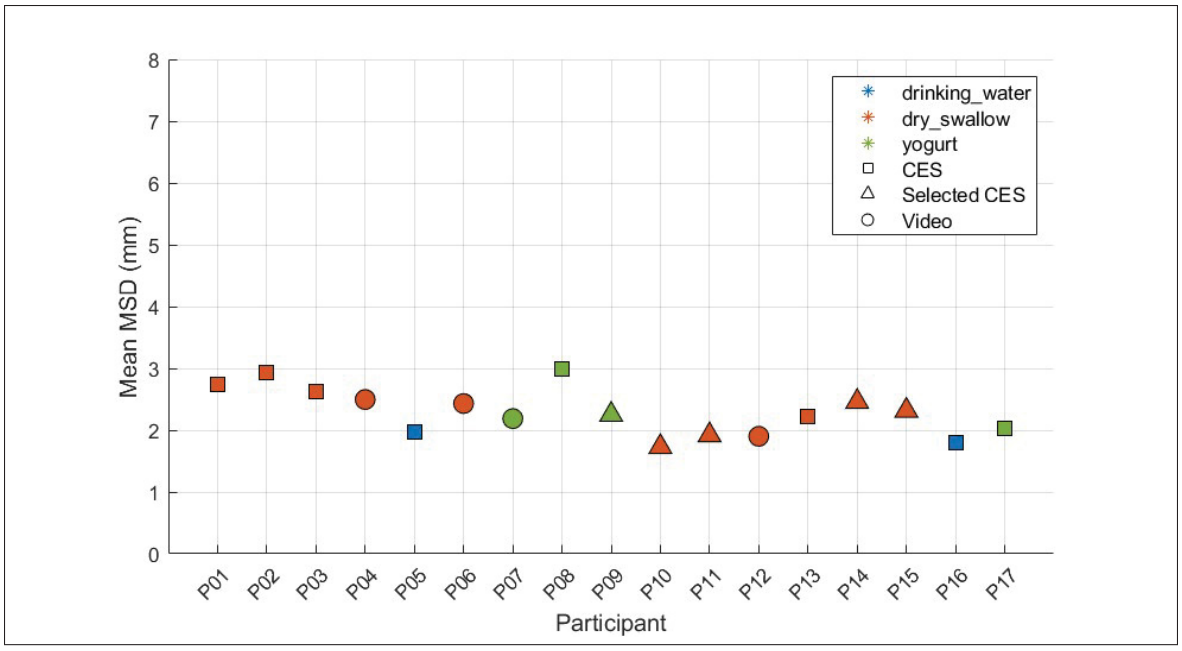
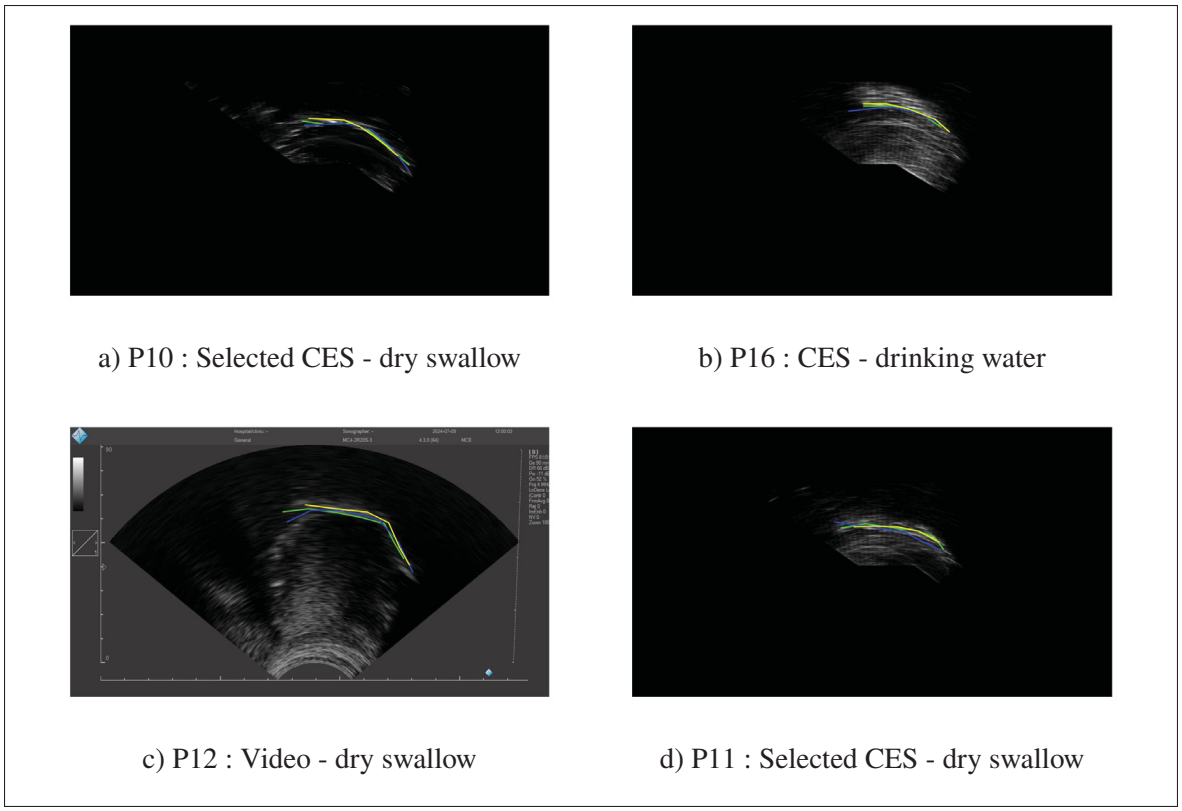


Figure 2.13 Best task-method combination per participant

(Rater 1 : frames 143 to 169, Rater 2 : frames 116 to 147) traced the palate in different locations within the CES. Figures 2.16a and 2.16b illustrate the CES reconstructed from the frame ranges selected by Rater 1 (green) and Rater 2 (blue), respectively. A similar scenario is observed with P11 in Figure 2.16c, where probe or jaw movement, poor recording quality, and differences in frame selection collectively reduced inter-rater agreement. In Figure 2.16d, even though raters annotated on the same full CES image with minimal probe movement, disagreement arises in tracing the palate. Figures 2.16e and 2.16f show two distinct frames from the same video of P08 drinking water, highlighting the effect of probe movement. While Figure 2.16e aligns with the tracings of Raters 2 and 3, Figure 2.16f corresponds with the tracing of Rater 1.

Figure 2.17 illustrates the intra-rater agreement of rater 3 after three months. Overall, the rater's measurements remain fairly consistent (with relatively low MSD and narrow box plots). Still, we see that the dry-swallow task tends to exhibit lower and tighter MSD distributions than the drinking water and yogurt tasks across all three methods.



a) P10 : Selected CES - dry swallow

b) P16 : CES - drinking water

c) P12 : Video - dry swallow

d) P11 : Selected CES - dry swallow

Figure 2.14 Best task-method combinations across participants

2.4.2 Automatic palate tracing

To evaluate the performance of the automatic palate tracing algorithm, we used the MSD to compare the algorithm's outputs with manual annotations performed by the raters across three swallowing tasks : dry swallow, yogurt consumption, and water drinking. We chose to compare with the raters' CES annotations, since CES consistently demonstrated the highest inter-rater agreement and is the basis of the automated method. Table 2.1 provides the mean MSD and the corresponding standard deviations (STD) for each rater and task combination. From these results, it is evident that dry swallows present fewer challenges for the algorithm, whereas drinking water shows higher MSD, likely due to strong echoes introduced by the water bolus (Figure 2.20a). Interestingly, manual raters are not affected by this.

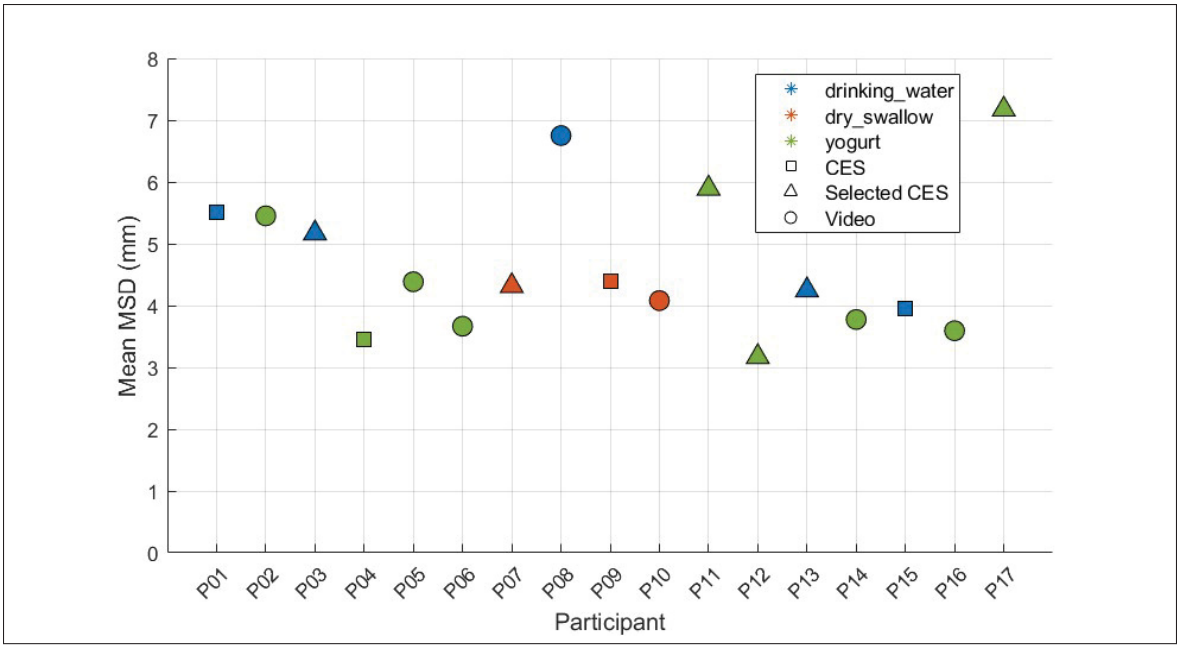


Figure 2.15 Worst task-method combination per participant

Tableau 2.1 MSD in millimeters between the automatic palate tracing and each rater's manual annotations, across three swallowing tasks (dry swallow, yogurt, and drinking water)

Task	MSD (mm)	Rater 1	Rater 2	Rater 3	Across raters
Dry swallow	Mean	**2.73**	**2.92**	**2.22**	**2.63**
	STD	**0.73**	**0.70**	**0.63**	**0.76**
Yogurt	Mean	4.17	3.87	3.27	3.77
	STD	1.89	2.01	1.73	1.88
Drinking water	Mean	4.30	4.52	3.74	4.19
	STD	3.19	2.88	2.87	2.49

To better understand how the automatic method behaves, we present example figures illustrating both successful and problematic cases for each swallowing task. Figure 2.18 shows good alignment between the automatic and manual traces, whereas Figure 2.20 highlights the more complex fluid artifacts encountered during yogurt and water trials. In dry swallows, the alignment is strong, reflecting the lower MSD values in Table 2.1. Although the mean MSD values are

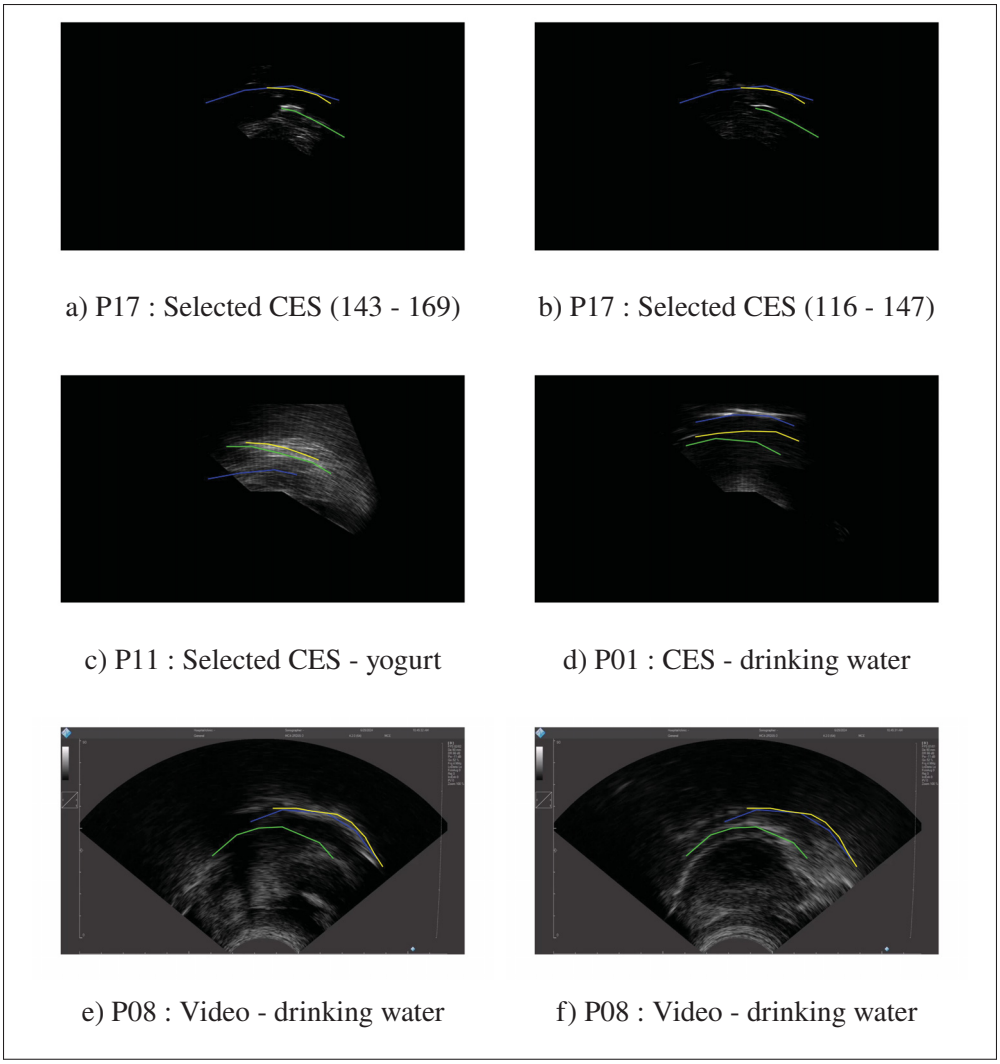


Figure 2.16 Worst task-method combinations across participants

higher for yogurt and water compared to dry swallows, the successful examples in Figure 2.18 highlight instances where the automatic trace matches the raters' annotations closely.

Despite these promising results, some cases revealed limitations of the current approach. Figure 2.19 presents a CES (2.19a) alongside the automatic and manual traces (2.19b) for a dry swallow. Here, the algorithm mistakenly identifies the tongue rather than the palate. Notably, one of the raters (blue) also traces the tongue. The largest-cluster selection strategy contributes to the

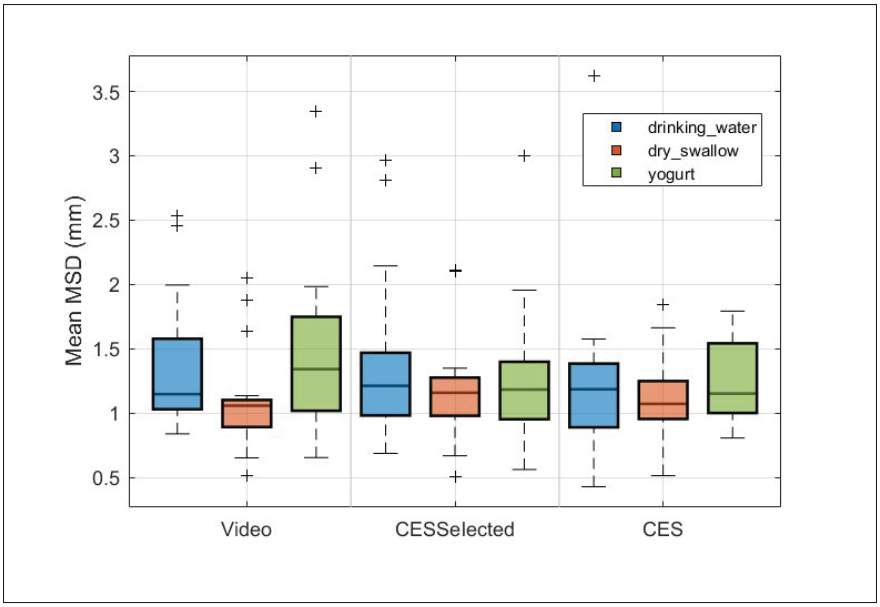


Figure 2.17 Mean MSD for intra-rater agreement over 3 months, grouped by method and task

mis-classification, as the tongue, in this example, produces a wider and stronger echo compared to the palate.

As shown in Figure 2.20, when a participant ingests yogurt or water, the bolus in the oral cavity produces strong echoes that can confuse the clustering step of the automatic algorithm, occasionally causing the palate trace to extend into non-palatal regions. Although these additional echoes do not lower inter-rater agreement, they nonetheless pose difficulties for the automatic method. Human raters can rely on prior knowledge of palatal shape and location to interpret ambiguous echoes; however, our algorithm currently does not incorporate such contextual information. Qualitatively, the raters and the algorithm have found mainly the same shape of the palate, even though the decision of the tracing placements varied. In addition, the raters did not discuss how to complete the tracing; yet, each tried to maintain agreement in their own tracing style.

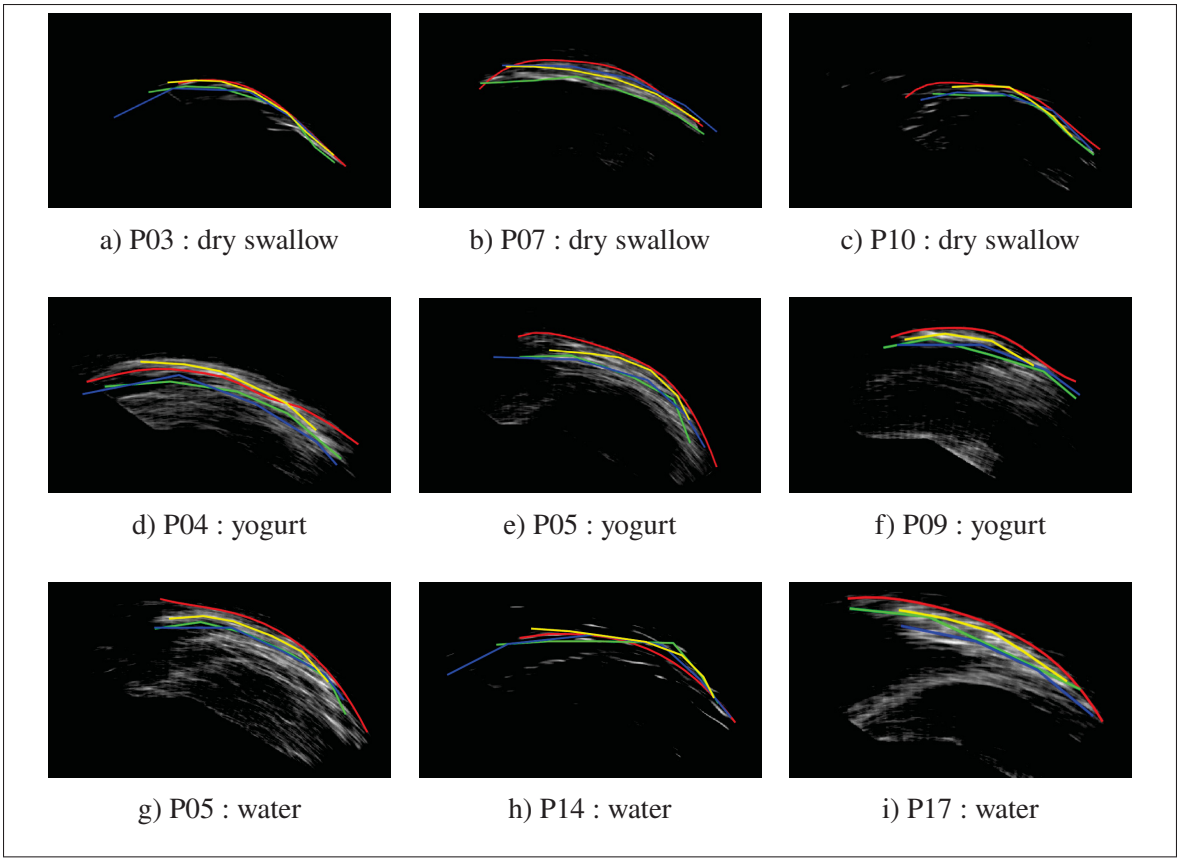


Figure 2.18 Examples of successful automatic palate tracing (red) overlaid on manual annotations from Rater 1, 2, and 3 (green, blue, and yellow, respectively)

2.5 Conclusions

By evaluating the reproducibility of three manual palate tracing methods across different swallowing tasks and benchmarking an automatic CES-based tracing method against manual annotations, we have identified critical factors influencing tracing reliability. A key finding is the effectiveness of CES methods in compensating for probe and jaw movements, enhancing agreement between raters. Additionally, using a dry swallow task as a means to observe the palate improves the reproducibility of manual tracings and the accuracy of the automatic method. This is encouraging, as dry swallows occur spontaneously and frequently during speech recordings, which could be beneficial for accurately and/or automatically keeping track of the palate over

a) b)

Figure 2.19 Example of palate misidentification in dry swallow. (a) CES. (b) Automatic trace (red) and manual traces from Rater 1 (green), Rater 2 (blue), and Rater 3 (yellow)

a) water b) yogurt

Figure 2.20 Example CES for water and yogurt swallowing tasks

time. Regardless of the tracing method or task employed, optimizing US image quality is crucial. It is also important to note that some subjects may never image very well due to factors such as submental adiposity or anatomical variations. The automatically traced palate proved its potential as an additional visual cue for SLPs during pronunciation training and serves as a stepping-stone toward real-time palate tracking. Future work will validate the method with speakers who have speech disorders and optimise the system for clinical use.

CHAPITRE 3

PALATE TRACKING IN ULTRASOUND IMAGES

Hana Ben Asker[1] , Eija M.A. Aalto[2] , Lucie Ménard[3] , Walcir Cardoso[4] , Catherine Laporte[2]

[1] Département de Génie des technologies de l'information, École de Technologie Supérieure,
[2] Département de Génie Électrique, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[3] Département de Linquistique, Université de Québec à Montréal,
405 Sainte-Catherine Est, Montréal, Québec, Canada H2L 2C4
[4] Département d'éducation, Université Concordia,
1455 Maisonneuve Ouest, Montréal, Québec H3G 1M8

Article soumis à la revue « Journal of the Acoustical Society of America », novembre 2025.

3.1 Abstract

Tracking the palate in ultrasound images provides an anatomical reference for speech analysis and clinical applications. To our knowledge, current methods only delineate the palate trace manually or automatically extract static traces, with no existing work on automatic dynamic tracking. In this article, we propose a novel solution to continuously track the palate trace even during acoustic invisibility. Our method leverages the observation that while the palate is intermittently visible due to air gaps preventing ultrasound transmission, landmarks like the tendon of the genioglossus remain consistently detectable. We model the correlation between tendon displacement and palatal motion as a rigid transformation. The system combines a YOLO tendon detector with particle filtering for robust temporal tracking, followed by palate position inference based on the proposed rigid transformation. Experiments on 71 videos (51 swallowing, 20 free speech) demonstrate that the method achieves mean MSD errors of 1.34-2.68 mm across different tasks, with a significant correlation ($r = 0.64$, $p = 0.001$) between tendon tracking accuracy and palate tracking performance. The system maintains reliable tracking even when palate visibility drops below 5% in free speech, validating tendon-based inference as an effective solution for continuous palate tracking in ultrasound-guided speech analysis.

3.2 Introduction

Understanding how articulators move is an important factor in advancing phonetic research, improving clinical treatment of speech sound disorders (SSD), and developing more effective approaches to second language (L2) pronunciation teaching. Among available imaging techniques, ultrasound has emerged as a particularly valuable tool for visualizing tongue articulation. Its advantages are considerable : it is portable, non-invasive, relatively affordable, avoids ionizing radiation, and provides the temporal resolution needed to capture the rapid and complex movements of the tongue in real time (Stone, 2005). These characteristics also make ultrasound suitable for field work and data collection.

Real-time ultrasound imaging has proven to be a valuable tool in both clinical and research settings (Gick *et al.*, 2008; Preston, Leece & Maas, 2017). In clinical contexts, it provides immediate visual biofeedback for treating SSD. This includes populations such as individuals with hearing impairment, where it has been used alongside electropalatography to achieve significant improvements in speech production (Bernhardt, Gick, Bacsfalvi & Ashdown, 2003), and children with persistent SSD that have been resistant to traditional therapy, including childhood apraxia of speech (Preston, Brick & Landi, 2013; Cleland, Scobbie & Wrench, 2015b; McCabe, Preston, Evans & Heard, 2023). Phonetics researchers utilize it as a standard instrument to quantitatively measure and model the tongue shapes for vowels and consonants across languages, following established methodological guidelines (Stone, 2005). The technology also enables the detailed study of coarticulation by allowing researchers to measure the influence of neighboring sounds through comparing the shape and position of tongue contours (Zharkova, Hewlett & Hardcastle, 2012). For second-language (L2) learners, ultrasound can offer visual biofeedback for articulatory gestures that are otherwise hidden (Bliss *et al.*, 2018).

While the tongue is the primary focus of articulatory analysis, its movements are best understood within the fixed anatomical frame of reference provided by the hard palate. The hard palate serves as the passive articulatory target for numerous consonants (from alveolar stops to palatal approximants). However, because of the large acoustic impedance mismatch between air and

tissue, the air-filled oral cavity acts as a nearly perfect reflector of ultrasound waves, preventing visualization of the palate during standard speech ultrasound imaging. From a quantitative perspective, having a palate trace is important for comprehensive articulatory analysis for several key reasons : first, it establishes a coordinate system that enables the calculation of articulatory metrics like the location and degree of vocal tract constriction ; second, it provides the foundation for normalized cross-speaker comparisons by aligning anatomical landmarks, facilitating large-scale phonetics studies (Mielke *et al.*, 2005) ; and third, it provides necessary context, as research has shown that individual differences in palate shape systematically influence speech production strategies (Brunner *et al.*, 2009).

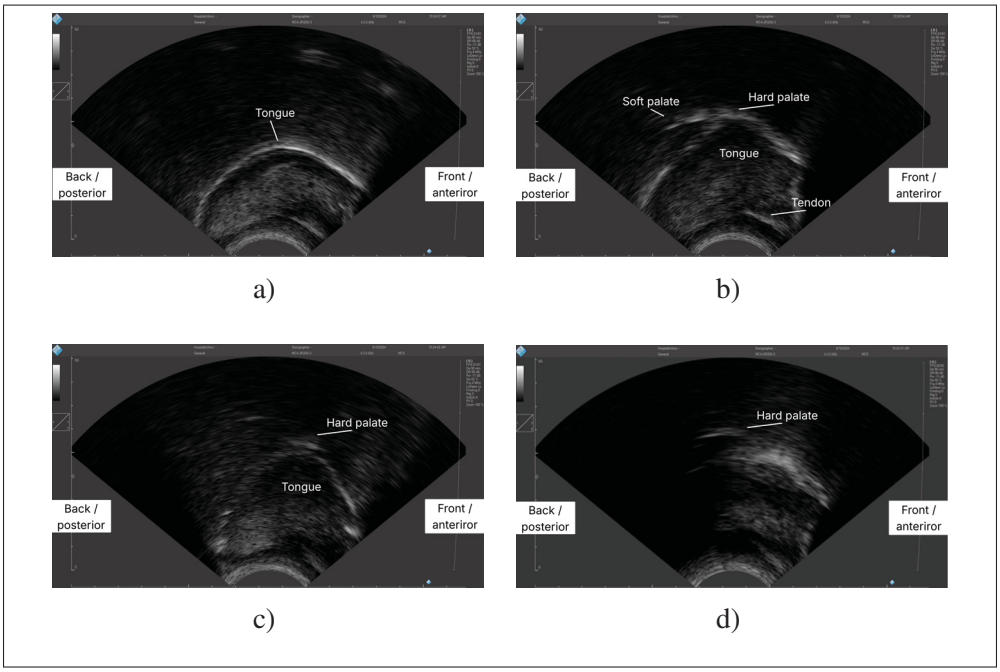


Figure 3.1 (a) Palate invisibility problem. (b) Completely visible palate. (c) Partially visible palate, /k/ pronunciation. (d) Bolus of water in the mouth

Obtaining a continuous palate reference throughout a recording, however, is complicated by two fundamental challenges : acoustic invisibility and apparent motion. First, the air gap between tongue and palate prevents ultrasound transmission, rendering the palate invisible (Figure 3.1a) except during direct linguopalatal contact (complete as in Figure 3.1b, or partial as in /k/ production, Figure 3.1c) or when acoustic coupling media fill the oral cavity (Figure 3.1d).

Second, although the hard palate is anatomically rigid, its position in the ultrasound image varies across frames due to relative motion between the transducer and the speaker's head. This apparent displacement is especially pronounced when the probe is hand-held, as is common in clinical, educational and field work settings, and is further compounded by the speaker's own head movements and mandibular adjustments during speech. While some researchers employ probe stabilization devices to reduce transducer motion, such setups constrain natural jaw movement to different degrees and may not be suitable for all experimental contexts or clinical applications. These two challenges, intermittent visibility and continuous apparent motion, necessitate an active tracking approach rather than a simple static trace.

Automated palate tracking represents a novel contribution to lingual ultrasound imaging. In L2 acquisition, automated palate tracking could enhance ultrasound biofeedback training by providing learners with a stable and intuitive anatomical target, reducing the cognitive load required to interpret the image. Recent evidence suggests that manually moved and overlaid palate traces during ultrasound biofeedback sessions improve learners' spatial awareness of tongue-palate relationships and enhance the pedagogical effectiveness of lingual ultrasound biofeedback (Aalto *et al.*, 2025a). While tracing provides a static palate contour, tracking would enable continuous, frame-by-frame localization of the palate, which is essential for accurate registration, normalization, and measurement of tongue movement relative to the palate across time and tasks. Automated tracking methods would further reduce manual labor, improve reproducibility, and facilitate large-scale or real-time studies, whereas manual tracing is labor-intensive, subject to human error, and limited to specific moments when the palate is visible.

This paper presents a novel solution to the palate tracking challenge. Our approach builds on the observation that, although the palate is only intermittently visible and shifts with probe or head movement, stable landmarks such as the tendon of the genioglossus remain consistently detectable and trackable. The genioglossus tendon appears in lingual ultrasound images as a short, bright, linear structure (Stone, 2005). By tracking the tendon and explicitly modeling the spatial relationship between tendon displacement and palatal motion, and assuming that this

relationship is relatively stable over time, we introduce an automated method for palate tracking in ultrasound images.

This paper begins by reviewing related work in Section 3.3. We then describe the proposed method in Section 3.4 and analyze our findings in Section 3.5. Finally, we offer concluding remarks in Section 3.6.

3.3 Related works

Automatic tracking in ultrasound imaging has reached notable maturity for tongue tracking, yet palate analysis remains constrained by the physics of ultrasound. Early semi-automatic tongue tracking methods, such as EdgeTrak, used local intensity cues but often failed when contours blurred (Li *et al.*, 2005; Csapó & Lulich, 2015). Later systems improved robustness via global constraints, as in TongueTrack (mean accuracy ≈ 3 mm) using a Markov Random Field model (Tang *et al.*, 2012), or particle filtering, as in SLURP (MSD = 1.69 ± 1.10 mm) (Laporte & Ménard, 2018). The need for manual initialization was removed by fully automatic detectors maintaining comparable accuracy (Karimi *et al.*, 2019). Meanwhile, statistical shape models, notably Active Appearance Models trained on multimodal data, provided strong priors for plausible tongue shapes even under noise (Roussos *et al.*, 2009).

Deep learning has since transformed the field, enhancing both accuracy and methodology. BowNet employs dilated convolutions for fast, robust, end-to-end segmentation (Mozaffari & Lee, 2020), while DeepEdge (1.44 mm MSD) merges deep detection with active-contour refinement (Chen *et al.*, 2020). Most recently, markerless pose estimation tools like DeepLabCut have shifted focus from dense contours to sparse kinematic keypoints, achieving human-level accuracy (0.93 mm MSD) (Wrench & Balch-Tomes, 2022).

In stark contrast to the tracking of the tongue, obtaining the palate contour is a problem of static reconstruction, necessitated by its acoustic invisibility during speech (Epstein & Stone, 2005). Foundational methods require a non-speech task, such as swallowing a liquid bolus, to create a temporary acoustic window to the palate ; a composite static trace is then manually assembled

from partial segments visible across multiple video frames (Stone, 2005; Epstein & Stone, 2005). The only fully automated method, the Cumulative Echo Skeleton (CES) technique, automates this reconstruction by temporally averaging echo information from a swallow sequence to isolate the immobile palate, achieving results within 3 mm of manual traces (Faucher *et al.*, 2019). Our recent evaluation of the CES method found mean errors of 2.63 mm for dry swallows, 3.77 mm for yogurt, and 4.19 mm for water (Ben Asker, Aalto, Ménard, Cardoso & Laporte, 2025). Alternative strategies are based on indirect inference from the tongue's movement envelope (Wrench, 2017). All of the mentioned methods have a critical limitation : they are designed to reconstruct the palate's static shape, not to track its apparent motion in real-time. This leaves a gap for the analysis of continuous speech, where a method is needed to maintain an accurate palate reference on a frame-by-frame basis.

In our recent work (Ben Asker *et al.*, 2025), we examined the conditions under which the palate can be most reliably traced in ultrasound sequences, both manually and using an enhanced version of the CES method (Laporte *et al.*, 2022). We found that dry swallows consistently yielded the most reproducible and accurate contours, which is encouraging given their common and spontaneous occurrence during speech recordings. This insight offers a practical foundation for the initialization and evaluation of automated palate trackers, while also providing evidence that reliable palate tracking may be achievable under natural speech conditions.

3.4 Methods

3.4.1 Data collection

Seventeen healthy adults (10 females, 7 males ; aged 21-51 years) with no self-reported speech, language, or swallowing disorders participated in the study. Data acquisition employed a Telemed MicrUs EXT-1H ultrasound system with a MC4-2R20S-3 microconvex transducer (2-4 MHz frequency range, 20 mm radius, 90 mm scanning depth). Participants handheld the transducer in a submental position throughout the imaging procedure. This method allows for natural jaw articulation, which is captured in the image plane alongside any probe displacement.

The experimental protocol encompassed three distinct swallowing conditions : water drinking, yogurt, and dry swallowing. Data collection yielded 51 swallowing videos distributed equally across the three tasks (n=17 per task). To capture natural dry swallows during spontaneous speech, we asked participants to name their favorite dish before the recording session. This dish was subsequently presented as a visual stimulus during data collection, with participants instructed to provide detailed descriptions including the dish's composition, preparation method, aroma, and taste, thereby inducing spontaneous dry swallows during speech. This procedure generated 17 food talk videos. The dataset was further augmented with spontaneous speech data from an ongoing L2 acquisition study conducted in our laboratory. These recordings were acquired using identical ultrasound equipment but with probe stabilization achieved via a microphone stand. This setup, which requires the participant to maintain submental contact with the transducer, constrains the natural range of jaw movement during speech. These recordings comprised speech samples on random topics from 9 participants.

The total collected dataset dataset comprised ultrasound video recordings from two distinct categories : 51 swallowing videos across three conditions (17 videos each for dry swallow, drinking water, and yogurt) and 26 free speech videos (17 food talk videos and 9 random topic videos).

3.4.2 Data annotation

Palate traces were annotated when the palate was completely visible in the ultrasound image. Among the 26 free speech videos, 15 of 17 food talk videos contained observable spontaneous swallows compared to only 5 of 9 random topic videos. The 6 free speech videos (2 food talk, 4 random topic) that contained no observable swallows were excluded from the palate tracking analysis, as no ground-truth palate trace could be established. The final palate analysis was therefore performed on 71 videos (the 51 swallowing videos and the 20 free speech videos containing swallows). Palate visibility varied dramatically between recording conditions, as shown in Table 3.1.

Tableau 3.1 Palate Visibility Statistics by Video Type

Video Type	**Total Frames**	**Visibility (%)**
Free Speech (Food Talk)	37031	4.2
Free Speech (Random Topic)	14187	0.8
Swallow (Drinking water)	3132	40.4
Swallow (Dry swallow)	3574	70.1
Swallow (Yogurt)	3515	34.9

Swallowing videos demonstrated substantially more frequent palate visibility across all tasks, reflecting the full tongue-palate contact that occurs during swallowing actions. In contrast, free speech videos were markedly less frequently visible during continuous speech articulation. Spontaneous swallows (defined as continuous periods of full palate appearance) remained infrequent overall, occurring in just 0.28% of frames in food talk videos and 0.07% in random topic videos during continuous speech. This is because swallowing occupied a much smaller proportion of these videos (where speech was the primary type of content) compared to the swallowing videos. These free speech recordings, with their extremely sparse palate visibility, are therefore important for validating the core hypothesis of this work : that continuous palate tracking is achievable by inference from a consistently visible landmark even when the palate itself is invisible for the vast majority of the recording.

In 76 of the 77 collected videos, the tendon of the genioglossus was annotated as a point landmark. One food talk video was excluded as the tendon was not visible due to suboptimal imaging conditions. The 76 annotated videos included 5 free-speech recordings without visible palate traces, which were excluded from palate tracking but retained for tendon tracking evaluation. Frame-by-frame annotation was carried out for all 51 swallowing videos, while the 25 remaining free speech videos were annotated at 10-frame intervals.

3.4.3 The tendon-palate relationship

The movement mechanism is modeled using a two-step transformation :

$$\mathbf{P}(t) = \mathbf{R}(\theta, \mathbf{T}(t)) \cdot \big[\mathbf{P}(t-1) + \Delta\mathbf{T}(t)\big], \tag{3.1}$$

where $\mathbf{P}(t)$ denotes a vector of coordinates representing the tracked palate contour at frame t, obtained from the ultrasound image data. The term $\Delta\mathbf{T}(t)$ represents the displacement vector of the ground-truth tendon of the genioglossus between consecutive frames. The matrix $\mathbf{R}(\theta, \mathbf{T}(t))$ describes a rotation centered at the tendon position $\mathbf{T}(t)$ with angle $\theta = \Delta x \cdot \alpha$. Here, Δx is the horizontal component of the tendon displacement vector $\Delta\mathbf{T}(t)$, and α is a scaling factor selected through parameter tuning based on the palate tracking error (Equation 3.15) on the swallowing dataset. This formulation captures both the translational coupling and a rotational component arising from probe motion. Figure 3.2 illustrates this two-step mechanism.

3.4.4 Validation of the tendon-palate motion model

Jaw openness can influence the palate-tendon relationship, and we acknowledge that the assumption of a strictly fixed spatial relationship is strong. However, we hypothesize that this assumption may hold sufficiently true in certain circumstances to allow for useful palate tracking. To evaluate the stability of the modeled tendon-palate relationship, we used the swallowing videos, as they are the only sequences annotated frame-by-frame and where the palate is most frequently visible. For each swallowing video, we selected a reference frame in which a manually annotated palate trace was anchored to its corresponding manually traced tendon position. We then propagated this palate trace across the entire video sequence by following the trajectory of the ground-truth tendon point and the defined movement mechanism (Equation 3.1). Swallowing videos were used for validation, as they are the only sequences annotated frame-by-frame and the ones in which the palate is most frequently visible.

The results in Table 3.2 support the use of a fixed model for movement between palate and tendon structures. Mean MSD values remained consistently low across tasks, with slightly lower

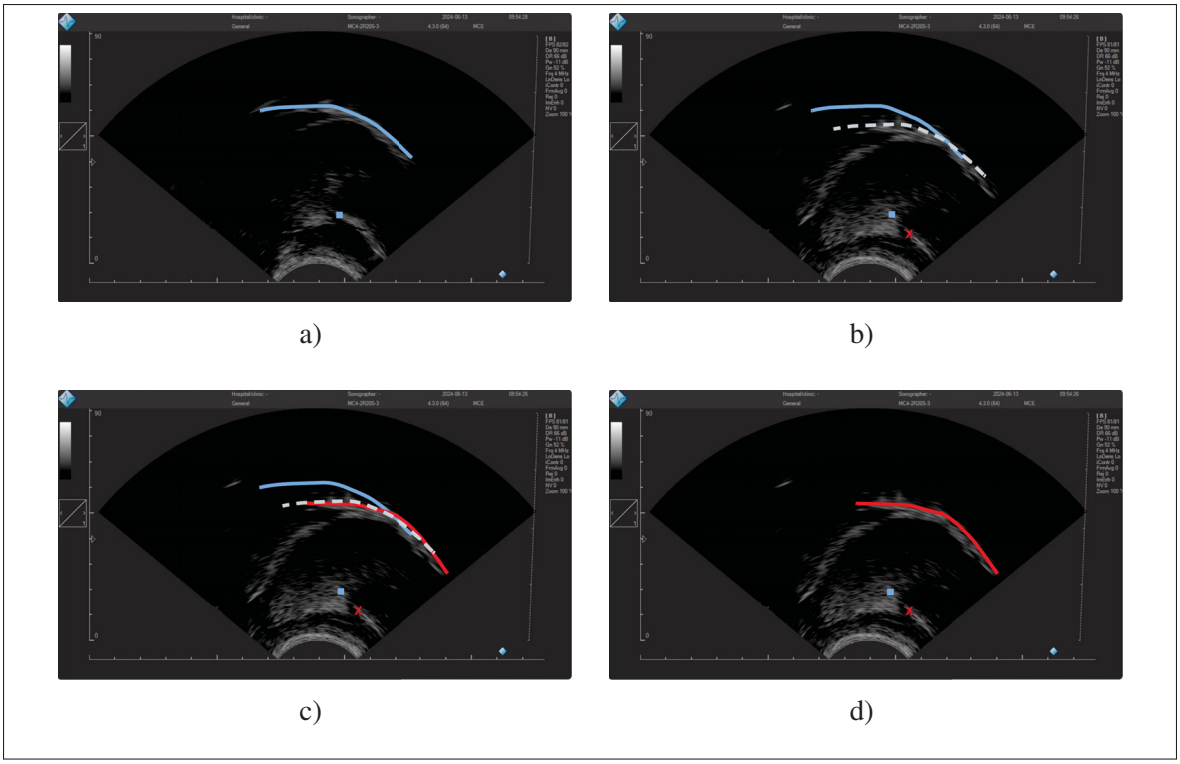


Figure 3.2 The two-step palate movement mechanism. (a) Initial position (blue, Frame 76). (b) Intermediate translation (gray) following tendon displacement (blue square to red X). (c, d) Final position (red) after rotation around the new tendon point (Frame 112)

Tableau 3.2 Mean ± SD MSD error (mm) between moved and manual palate traces and inter-rater variability (Ben Asker *et al.*, 2025) across tasks

Task	**This work**	**Inter-rater variability**
Drinking water	2.07 ± 0.83	3.18 ± 0.88
Dry swallow	1.75 ± 0.53	2.87 ± 0.65
Yogurt	2.13 ± 0.54	3.33 ± 0.92

averages when the palate was more frequently visible. This suggests that the more reliable the ground-truth palate traces, the more closely our tendon-based movement mechanism aligns with them. The low standard deviations across tasks indicate stable performance of the coupling mechanism, validating tendon movement as a reliable predictor of palate displacement within

the ultrasound field of view. To contextualize this performance, we can directly compare the model's error against the inter-rater variability of manual annotation, as reported in Table 3.2 and documented for the same dataset in our previous work (Ben Asker *et al.*, 2025). This demonstrates that even without explicitly accounting for independent jaw motion, the proposed two-step transformation model achieves an accuracy that meets or even exceeds the inter-rater consistency on the swallowing dataset.

3.4.5 Automated tracking of the tendon of the genioglossus

For robust tendon tracking, the proposed hybrid system combines a deep learning-based detector with a particle filter for temporal modeling, as illustrated in Fig. 3.3. The detection component employs a YOLOv8 model trained to identify the tendon in a given ultrasound image. For each frame k, the detector produces a measurement consisting of the bounding box's center coordinates, denoted as the position vector $\mathbf{z}_k$, and an associated confidence score, c_k.

The swallowing dataset was partitioned by participant into training (11 participants, 33 videos), validation (2 participants, 6 videos), and test (4 participants, 12 videos) sets. This speaker-independent partitioning assesses model generalization. The free speech (both random topic and food talk) videos were employed as additional, more complex test cases.

To address temporal discontinuities and noise inherent in ultrasound imaging, the system employs a particle filter with a constant velocity motion model augmented with a curvature bias and damping.

The proposed motion model is grounded in observed motion patterns of the genioglossus tendon during ultrasound imaging of speech and swallowing. In practice, the tendon exhibits two motion regimes : when the ultrasound probe or the head rocks, it follows smooth curved trajectories across the image plane, motivating the inclusion of curvature bias ($\mathbf{v}'^{\perp}_k$) that constrains motion along circular paths. Further, the tendon reflects intrinsic jaw motion, producing more variable vertical displacements. To capture this dual behavior, the model incorporates an orthogonal

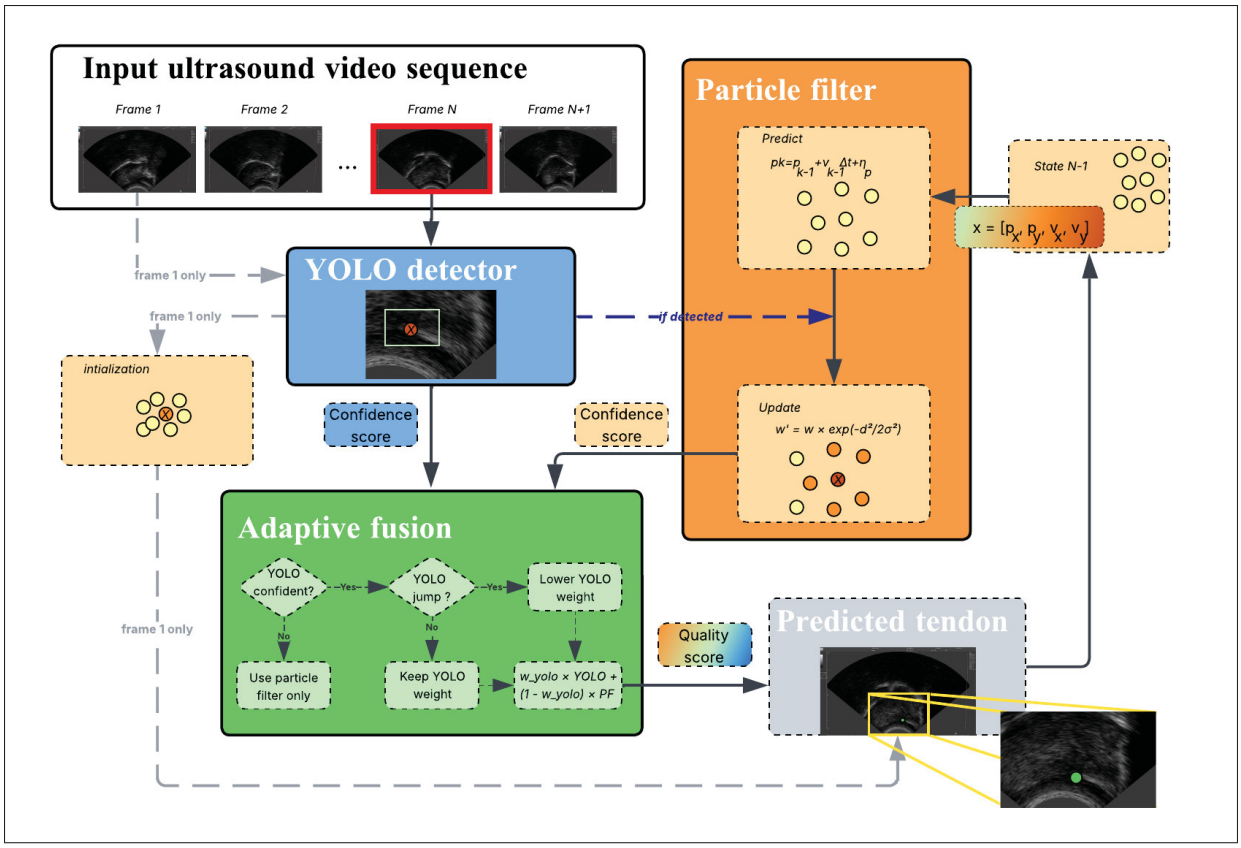


Figure 3.3 The architecture of the proposed tendon tracker. The system integrates a YOLOv8 detector for spatial localization with a particle filter for temporal tracking. An adaptive fusion module dynamically combines information from both components to produce a robust, continuous track of the tendon

force component to guide motion along curved paths, while adaptive process noise accounts for more variable displacements, such as those arising from jaw motion.

Each particle represents a hypothesis of the tendon's kinematic state. In this implementation, the state is a 4-dimensional vector :

$$\mathbf{x}_k = [\mathbf{p}_k, \mathbf{v}_k]^T = [p_x, p_y, v_x, v_y]_k^T \tag{3.2}$$

where $\mathbf{p} = (p_x, p_y)$ represents the normalized image coordinates, and $\mathbf{v} = (v_x, v_y)$ represents the corresponding velocity vector.

The prediction step updates each particle using a constant velocity model that incorporates tissue resistance (damping), a curvature bias, and adaptive process noise.

First, velocity damping is applied to model tissue resistance, which prevents unrealistic momentum and improves stability during detection gaps. The velocity from the previous state, $\mathbf{v}_{k-1}$, is dampened to produce an intermediate velocity $\mathbf{v}'_k$:

$$\mathbf{v}'_k = \gamma \cdot \mathbf{v}_{k-1} \tag{3.3}$$

where $\gamma \in (0, 1)$ is the velocity damping factor.

Next, a curvature bias is introduced as a small force perpendicular to the damped velocity, guiding the particles along an arc-like path. The perpendicular vector to $\mathbf{v}'_k = (v'_x, v'_y)$ is given by $\mathbf{v}'^{\perp}_k = (-v'_y, v'_x)$, yielding a second intermediate velocity, $\mathbf{v}''_k$:

$$\mathbf{v}''_k = \mathbf{v}'_k + \kappa \cdot \mathbf{v}'^{\perp}_k \tag{3.4}$$

with curvature strength $\kappa > 0$.

Finally, Gaussian process noise is added to both velocity and position to model system uncertainty :

$$\mathbf{v}_k = \mathbf{v}''_k + \boldsymbol{\eta}_v \tag{3.5}$$

$$\mathbf{p}_k = \mathbf{p}_{k-1} + \mathbf{v}_k \cdot \Delta t + \boldsymbol{\eta}_p \tag{3.6}$$

where $\boldsymbol{\eta}_v \sim \mathcal{N}(\mathbf{0}, \sigma_v^2 \mathbf{I})$ and $\boldsymbol{\eta}_p \sim \mathcal{N}(\mathbf{0}, \sigma_p^2 \mathbf{I})$. The noise standard deviations are adapted based on recent tracking reliability measures reported by YOLO :

$$\sigma_v = \sigma_{v,0} \cdot (2 - \bar{c}) \tag{3.7}$$

$$\sigma_p = \sigma_{p,0} \cdot (2 - \bar{c}) \tag{3.8}$$

Here, $\bar{c}$ is the moving average of the YOLO confidence scores (c_k) from recent frames. $\sigma_{v,0}$ and $\sigma_{p,0}$ are base noise levels.

When YOLO detects the tendon position $\mathbf{z}_k$, particle weights are updated using a Gaussian likelihood function. The measurement uncertainty is inversely proportional to the detection confidence :

$$\sigma_{\text{meas}} = \frac{\sigma_0}{c_k + \epsilon}, \tag{3.9}$$

$$\mathcal{L}_i = \exp\left(-\frac{\left\|\mathbf{p}_k^{(i)} - \mathbf{z}_k\right\|^2}{2\sigma_{\text{meas}}^2}\right), \tag{3.10}$$

$$w_k^{(i)} = w_{k-1}^{(i)} \cdot \mathcal{L}_i. \tag{3.11}$$

where $\|\cdot\|$ denotes the Euclidean norm, σ_0 is a base measurement noise constant, and $\epsilon > 0$ ensures numerical stability.

The effective number of particles, $N_{\text{eff}} = \left(\sum_{i=1}^{N} (w_k^{(i)})^2\right)^{-1}$, monitors particle diversity. Systematic resampling is triggered when $N_{\text{eff}} < \rho N$, where $\rho \in (0, 1)$ is the resampling threshold.

The filter's state estimate $\hat{\mathbf{p}}_k$ is the weighted mean of particle positions. This approach provides a robust expected value estimate of the tendon's position.

Before fusion, a detection $\mathbf{z}_k$ is considered confident only if its confidence c_k exceeds a minimum threshold, $c_{\min}$. If the detection is valid, the final fused position $\hat{\mathbf{p}}_{\text{final},k}$ is computed :

$$\hat{\mathbf{p}}_{\text{final},k} = w_{\text{YOLO}} \cdot \mathbf{z}_k + (1 - w_{\text{YOLO}}) \cdot \hat{\mathbf{p}}_k \tag{3.12}$$

The fusion weight w_{YOLO} is dynamically adjusted based on temporal consistency. The displacement between the current YOLO position $\mathbf{z}_k$ and the final fused position from the

previous frame $\hat{\mathbf{p}}_{\text{final},k-1}$ is computed :

$$d_{\text{jump}} = \|\mathbf{z}_k - \hat{\mathbf{p}}_{\text{final},k-1}\| \tag{3.13}$$

The YOLO weight is reduced for large jumps to ensure temporal smoothness :

$$w_{\text{YOLO}} = \begin{cases} \alpha_{\text{high}} & \text{if } d_{\text{jump}} < \tau_{\text{low}} \\ \alpha_{\text{med}} & \text{if } \tau_{\text{low}} \leq d_{\text{jump}} < \tau_{\text{high}} \\ \alpha_{\text{low}} & \text{if } d_{\text{jump}} \geq \tau_{\text{high}} \end{cases} \tag{3.14}$$

where $0 < \alpha_{\text{low}} < \alpha_{\text{med}} < \alpha_{\text{high}} \leq 1$ are weighting levels, and $\tau_{\text{low}}, \tau_{\text{high}}$ are displacement thresholds.

Key parameters include the particle count N, velocity damping γ, curvature strength κ, base process noise levels $\sigma_{v,0}$ and $\sigma_{p,0}$, measurement noise scaling σ_0, and resampling threshold ρ. This tuning was performed using a grid search aimed at minimizing the Mean Euclidean Distance (MED) (Eq. 3.16) between the tracker output and the ground-truth annotations. All parameters were optimized on the held-out validation dataset consisting of 6 videos (1,221 frames) from two participants who were also excluded from YOLO training dataset.

3.4.6 Automated tracking of the palate

The automatic tracking of the genioglossus tendon constitutes a fundamental component for fully automated palate tracking (Figure 3.4). To initialize the palate tracking algorithm, a reference frame is first selected in which the palate is completely visible. The palate contour in this reference frame is then retrieved and its position is anchored relative to the detected tendon position within the same frame.

The palate contour in the reference frame can be obtained through one of two approaches : either manually traced by the user prior to initiating the tracking procedure, or automatically generated

using the CES method (Faucher *et al.*, 2019). In these experiments, we used manually delineated palate traces on the reference frames.

Following the anchoring of the detected tendon and the palate contour, the tracking procedure is initiated from the first frame of the sequence. The tracking algorithm proceeds by applying the movement mechanism described in Section 3.4.3.

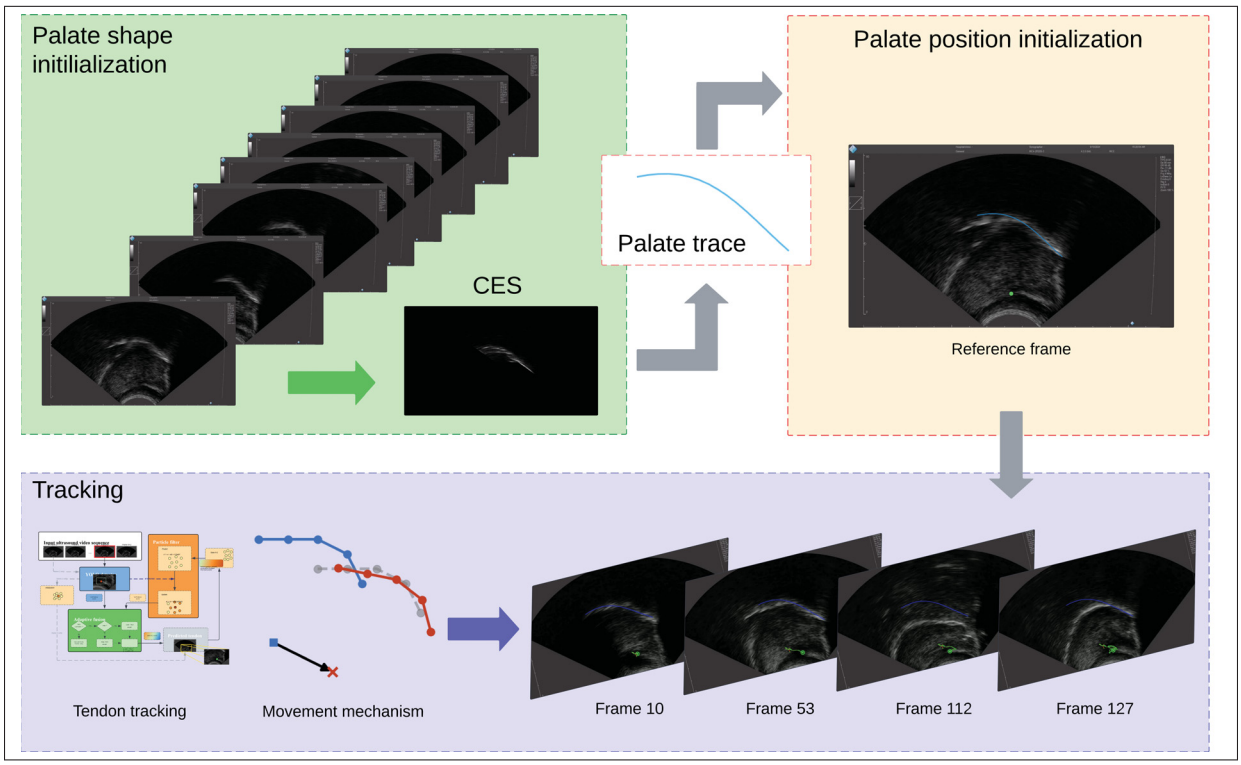


Figure 3.4 Pipeline for automatic palate tracking

3.4.7 Palate-guided drift mitigation

To explore potential improvements in tracking robustness, we investigated a palate-guided correction mechanism. The rationale for this approach is that the tendon tracker can experience temporary drift due to reduced echogenicity or neighboring anatomical structures that can be more echogenic than the tendon. When the palate becomes visible in the image, its echo provides a reliable anatomical reference that can be used to correct tracking errors.

To identify frames where the palate is sufficiently visible for correction, we trained a binary classification model using the ResNet-50 architecture. The training dataset consisted of a balanced set of ultrasound frames labeled as containing either visible or invisible palate contours. The model was trained to classify each input frame based on palate visibility, achieving 86.1% accuracy on the held-out test set. This classifier operates on each frame independently to determine whether palate-based correction should be attempted.

In frames classified as containing a visible palate, we perform automatic palate contour extraction using an echo skeleton approach. We identify the palate trace by extracting high-intensity echo patterns from the frame, followed by thresholding, clustering to select the palatal region, and spline fitting to obtain a smooth contour. The reference palate trace is then aligned to this newly extracted contour through rigid transformation. Alignment is achieved through the Kabsch algorithm, which uses singular value decomposition to compute the optimal rotation and translation minimizing the mean squared distance between corresponding points on both traces. This alignment preserves the shape of the reference trace while repositioning it to match the detected palate location in the current frame. We do not adopt the extracted contour directly ; rather, we use it only as a target for alignment to avoid propagating noise or geometric distortions from the single-frame extraction. The tendon position is then corrected to maintain consistency with the updated palate position : specifically, we infer the tendon location such that applying Equation 3.1 would reproduce the aligned palate trace. The tendon tracker is subsequently updated with this corrected tendon position.

When the palate is not visible, we rely solely on the tendon tracker (as described in Section 3.4.5), and the palate is moved according to the tendon's motion (as described in Section 3.4.3).

3.4.8 Evaluation metrics

To evaluate the accuracy of the proposed movement mechanism, we use the Mean Sum of Distances (MSD) Li *et al.* (2005). This metric quantifies the geometric discrepancy between two palate traces. Let $U = \{u_1, u_2, \ldots, u_n\}$ and $V = \{v_1, v_2, \ldots, v_m\}$ denote two sets of palate traces,

where each u_i and v_j represents a 2D coordinate point (x, y) in the ultrasound image space along the respective palate contours. The MSD is defined as the normalized sum of the distances from each point in one trace to its closest counterpart in the other, computed in both directions :

$$\text{MSD}(u, v) = \frac{1}{m+n}\left(\sum_{i=1}^{n} \min_j \|\mathbf{v}_i - \mathbf{u}_j\| + \sum_{j=1}^{m} \min_i \|\mathbf{u}_j - \mathbf{v}_i\|\right) \tag{3.15}$$

While the MSD is a robust measure for the accuracy of palate traces, evaluating the tendon tracker requires metrics that quantify the frame-by-frame error between the predicted position and a ground-truth reference.

Let $P = \{p_1, p_2, \ldots, p_N\}$ be the sequence of predicted tendon positions and $G = \{g_1, g_2, \ldots, g_N\}$ be the corresponding ground-truth positions for N validated frames, where p_i and g_i are 2D coordinates.

The primary metric is the Mean Euclidean Distance (MED). This represents the average straight-line distance between the predicted and true positions across all frames in a given video sequence. It is calculated as :

$$\text{MED} = \frac{1}{N}\sum_{i=1}^{N} \|p_i - g_i\|_2 \tag{3.16}$$

To assess clinical viability, we also compute the Accuracy at a distance threshold d (Acc@d). This metric calculates the percentage of frames where the tracking error is within an acceptable distance d. It is defined as :

$$\text{Acc@d} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(\|p_i - g_i\|_2 \leq d\right) \tag{3.17}$$

where $\mathbb{I}(\cdot)$ is the indicator function (1 if the condition is true, 0 otherwise).

3.5 Results

3.5.1 Performance of the automated tendon tracker

In this section, we present the tendon tracking results obtained across two distinct datasets : swallowing tasks and free speech videos. The evaluation encompasses 37 videos from 26 participants, totaling 10,825 annotated frames. These included 12 swallowing videos from 4 participants, 9 random topic free speech videos from 9 additional participants, and 16 food-talk videos from 16 participants, 3 of whom also took part in the swallowing tasks. The proposed

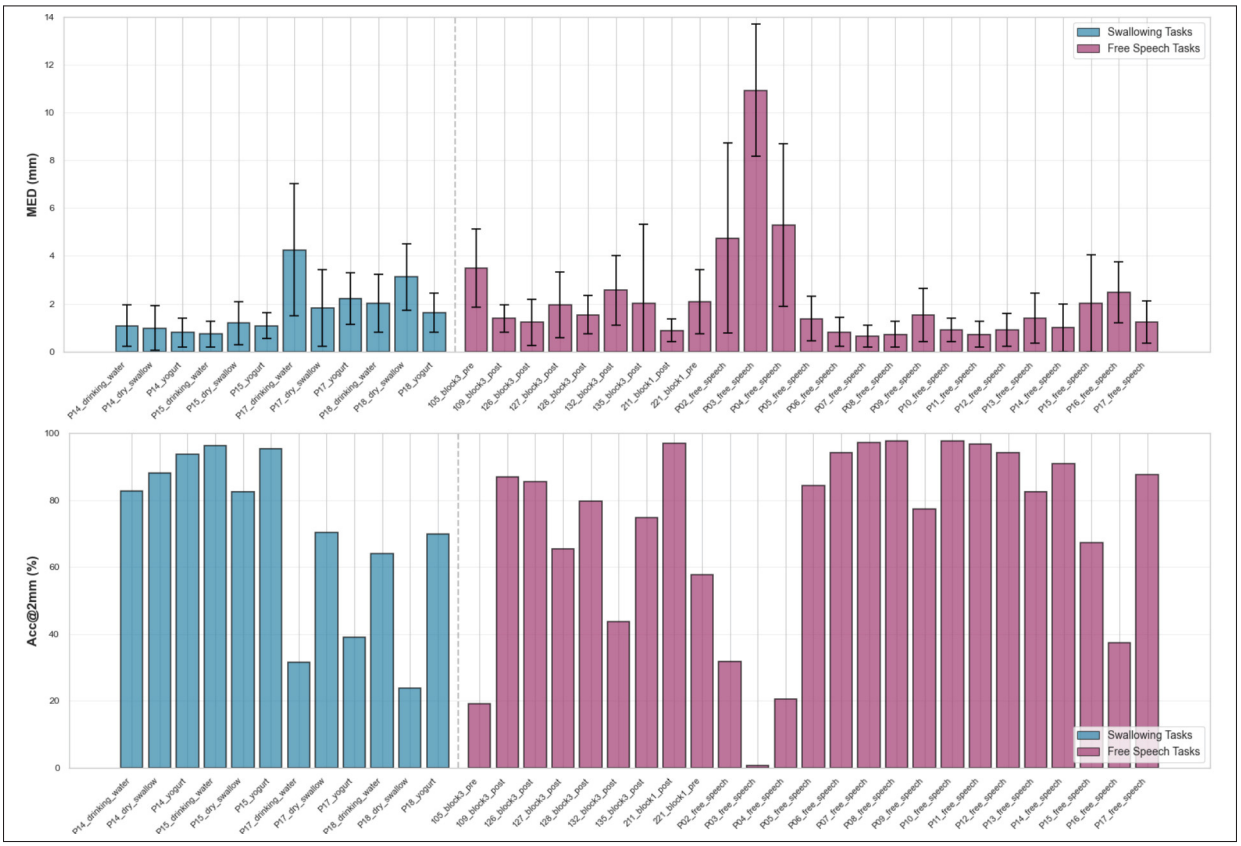


Figure 3.5 Performance of the tendon tracking algorithm on swallowing task and free speech videos

tendon tracking algorithm demonstrates variable performance across different participants and tasks, as summarized in Figure 3.5. For swallowing, the MED ranges from 0.79 ± 0.60 mm

to 4.26 ± 2.76 mm. While 7 out of the 12 sequences achieve an Acc@2mm above 70%, performance is more consistent at a larger error tolerance, with 10 out of 12 sequences achieving an Acc@5mm $>$ 90%. The free speech tasks exhibit greater variability, with MED values spanning from 0.64 ± 0.47 mm to 10.94 ± 2.76 mm. Notably, 17 out of 25 free speech sequences achieve an MED below 2 mm, indicating robust tracking under favorable conditions. However, some sequences (e.g., 105_block3_pre and P03_free_speech) demonstrate substantially degraded performance, with Acc@2mm below 20%, suggesting significant challenges in tendon visibility for specific participants and imaging configurations.

Overall, the tracking algorithm demonstrates robust performance when the genioglossus tendon presents as a distinct hyperechoic structure and when the ultrasound probe is optimally positioned and centered. However, tracking failure occurs under suboptimal imaging conditions where the tendon exhibits reduced echogenicity and becomes difficult to differentiate from surrounding tissue. These challenging cases, in which even expert human annotators reported uncertainty during manual delineation, typically arise when the ultrasound probe undergoes lateral displacement due to gel-mediated slipping.

Figure 3.6 shows representative cases where the genioglossus tendon was clearly visible as a sharply defined hyperechoic structure, enabling accurate tracking. For example, in Figures 3.6b and 3.6d, the algorithm achieved errors of 1.6 mm and 0.07 mm, respectively, under favorable imaging conditions.

Failure cases, shown in Figure 3.7, reveal the main sources of error. In P03_free_speech frame 827 (Figure 3.7b), poor tendon contrast and distracting anatomical structures produced a 16.52 mm error, with predictions deviating markedly from the ground truth. P18_dry_swallow frame 86 (Figure 3.7d) showed a 4.31 mm error caused by reduced tendon visibility and subsequent tracking drift. Such cases illustrate how failures often stem from diminished echogenicity, competing structures, or suboptimal probe placement.

Algorithmic behavior further explains these patterns. In challenging cases such as P17_drinking_water (MED = 4.26±2.76 mm, Acc@2mm = 31.6%) (Figure 3.5), the adaptive fusion mechanism

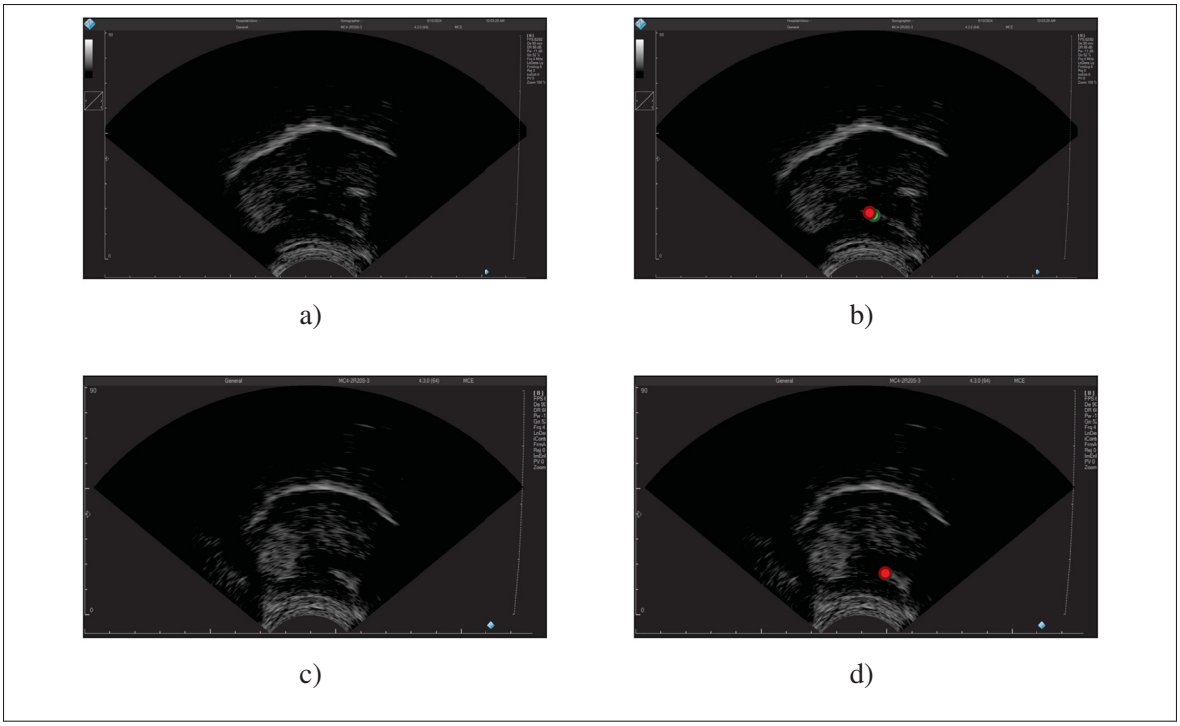


Figure 3.6 Example cases where the proposed tendon tracking model succeeds. Ground truth : Green. Prediction : Red. (a) P03_free_speech frame 2453. (b) MED = 1.6 mm. (c) 211_block1_post frame 659. (d) MED = 0.07 mm

experienced cascading failures. When YOLO detections degraded due to poor visibility, the particle filter relied increasingly on its motion model. Divergence between the model and the true trajectory then caused particle weights to concentrate on incorrect hypotheses, leading to unrecoverable drift during resampling.

Finally, the pronounced inter-participant variability underscores two critical factors. First, anatomical differences, such as tissue thickness, tendon echogenicity, and swallowing behavior, strongly influence performance. Second, imaging quality, including probe positioning, directly impacts tracking outcomes. Notably, even within a single participant (e.g., P03), performance ranged from excellent (1.6 mm) to poor (16.52 mm) depending on instantaneous imaging conditions.

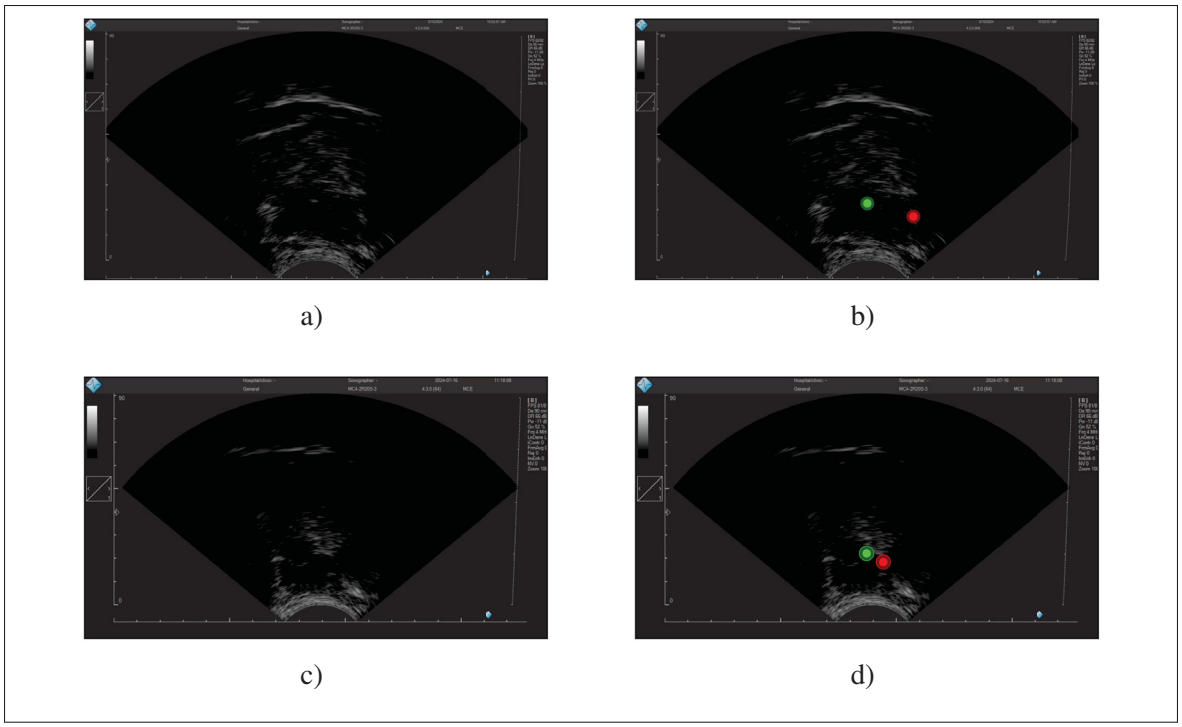


Figure 3.7 Example cases where the proposed tendon tracking model fails. Ground truth : Green. Prediction : Red. (a) P03_free_speech frame 827. (b) MED = 16.52 mm. (c) P18_dry_swallow frame 86. (d) MED = 4.31 mm

3.5.2 Performance of the automated palate tracker

The fully automated palate tracking pipeline (Figure 3.4) demonstrates variable but promising performance across different swallowing and speech tasks (Figure 3.8). Results in this section are based on tendon-only tracking without palate-guided correction. The system achieves mean MSD errors ranging from 1.24 mm in drinking water tasks to 2.77 mm in free speech with random topics.

The task-dependent variations shown in Figure 3.8 reveal the following pattern : swallow tasks, having the highest palate visibility, achieve the lower tracking error (1.24-1.81 mm mean MSD), while free speech tasks maintain reasonable accuracy (2.24-2.77 mm mean MSD) with lower palate visibility below. This suggests that the coupling mechanism between tendon and palate movement is particularly robust during swallowing actions, where the motion patterns are closer

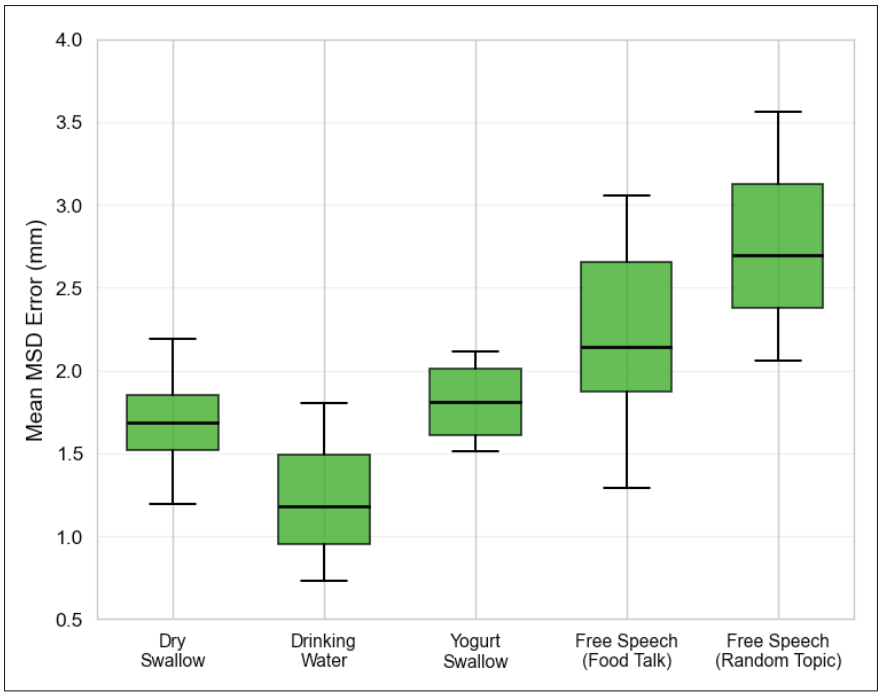


Figure 3.8 Palate tracking error by task

to the proposed movement mechanism, as swallowing typically involves less jaw movement than continuous speech.

Notably, the correlation analysis reveals a moderate positive relationship (r=0.64, p=0.001) between palate and tendon tracking errors, confirming that the accuracy of tendon tracking influences palate position estimation. This relationship is visually evident in Figure 3.9, where lower tendon tracking errors are associated with improved palate tracking performance. This suggests that incorporating visible palate features could help constrain or correct tendon position estimates when the palate boundary is well defined. A method for doing this was proposed in Section 3.4.7 and is evaluated in the following section.

Successful tracking examples (Figure 3.10) demonstrate the system's capability under optimal conditions. In case 109_block3_post (Figure 3.10a), the palate tracking error of 1.06 mm coincides with accurate tendon tracking (1.43 mm MED), illustrating the direct relationship between palate and tendon accuracies. Similarly, cases P06 and P08 (Figures 3.10c-3.10d)

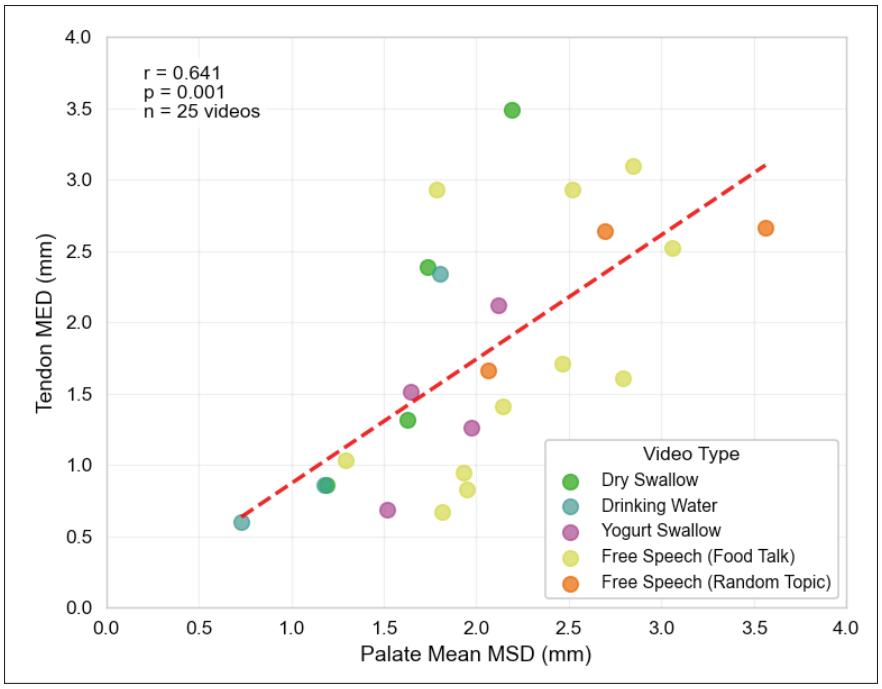


Figure 3.9 Error correlation

achieve sub-millimeter palate tracking errors during free speech, showing that reliable tracking is achievable in good imaging conditions, i.e., when the tendon is hyperechoic.

Failure cases (Figure 3.11) reveal systematic sources of error. The P02 free speech example (Figure 3.11a) with an 6.08 mm MSD represents a major failure, where degradation in tendon tracking cascades through the transformation mechanism. This occurs because other hyperechoic structures near the tendon mislead the tendon tracker. More moderate failures occur in P10, P13, and P14, with errors ranging from 3.05 to 4.15 mm.

In the case of Figure 3.11b, even though the tendon is accurately tracked, the palate tracking error is proportionally high. This occurs because the spatial relationship assumption breaks down, due to significant jaw opening that is not captured by the rigid transformation model. In Figure 3.11c, the tendon tracker fails because the tendon is not clearly visible, making annotation challenging even for experts. Finally, in Figure 3.11d, the tendon tracker also fails, with a tendon

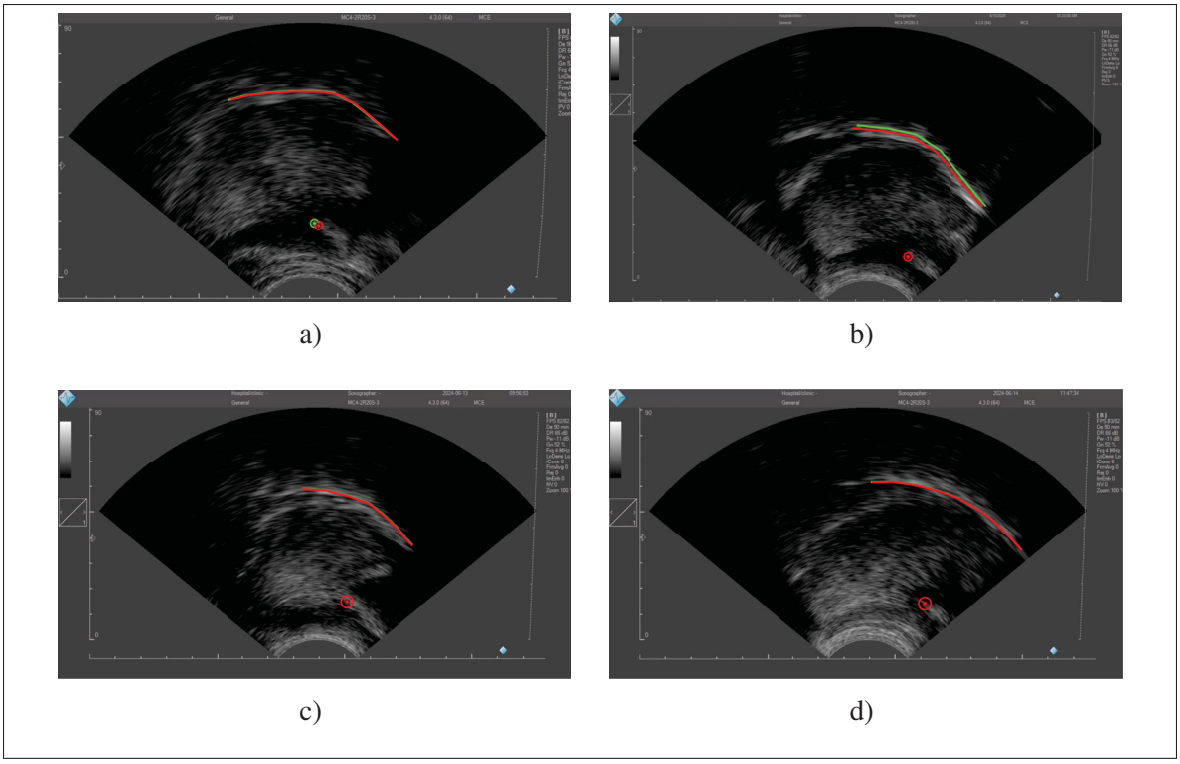


Figure 3.10 Successful examples of palate tracking using automatic tendon tracking. (a) 109_block3_post frame 388, Palate mean MSD = 1.06 mm, Tendon MED = 1.43 mm. (b) P04_free_speech frame 3010, Palate mean MSD = 1.36 mm. (c) P06_free_speech frame 632, Palate mean MSD = 0.09 mm. (d) P08_free_speech frame 1549, Palate mean MSD = 0.21 mm

tracking error of 3.47 mm, leading to a palate tracking error of 4.09 mm. This further highlights the problem of error propagation in the system.

The temporal consistency of the tracking system is further illustrated in Figure 3.12. The stability of the mean error, which remains consistently low across multiple reappearances, validates the tendon-based inference strategy. This demonstrates that the system maintains tracking continuity even through extended periods of palate invisibility.

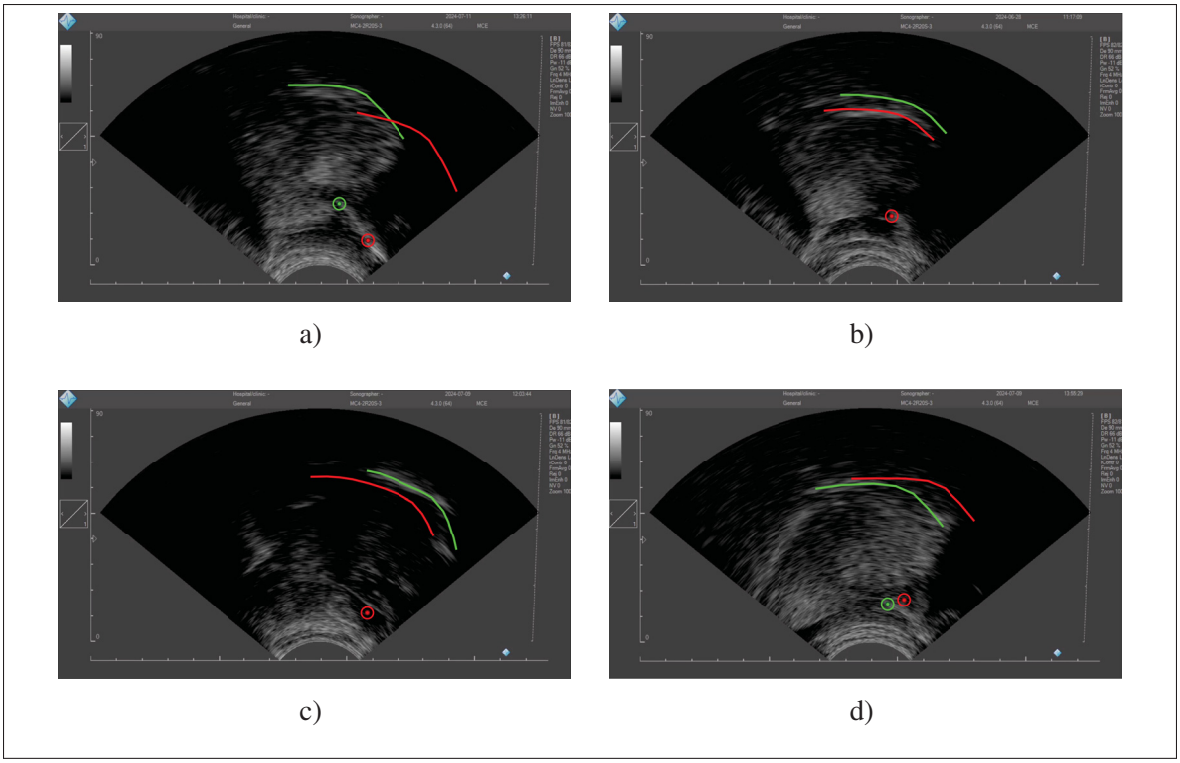


Figure 3.11 Example failures of palate tracking using automatic tendon tracking. (a) P17_dry_swallow frame 232, Palate mean MSD = 6.08 mm, Tendon MED = 9.62 mm. (b) P10_free_speech frame 424, Palate mean MSD = 3.05 mm. (c) P13_free_speech frame 1521, Palate mean MSD = 4.15 mm. (d) P14_free_speech frame 12, Palate mean MSD = 4.09 mm, Tendon MED = 3.47 mm

3.5.3 Evaluation of palate-guided tracking

To assess the impact of palate-guided correction described in Section 3.4.7, we compared two tracking approaches on the free speech dataset : the baseline tendon-based (TB) method, which relies solely on the tendon tracker and spatial transformation model (as evaluated in Section 3.5.2), and the palate-corrected (PC) method, which incorporates drift mitigation using visible palate features when available. Table 3.3 presents the comparative performance of both approaches on food talk and random topic videos.

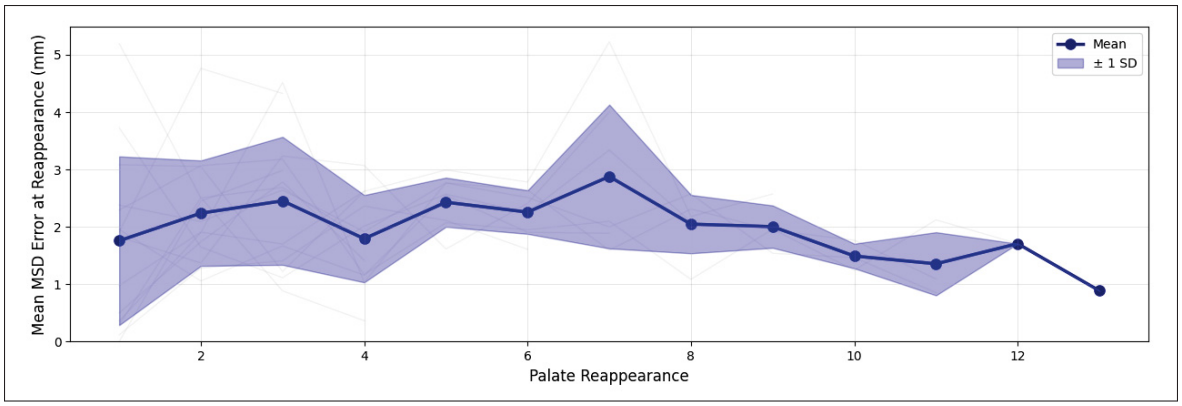


Figure 3.12 Comparison of mean MSD error across palate reappearance by video

Tableau 3.3 Comparison of tendon (MED) and palate (MSD) errors for the TB and PC methods across free speech videos

Method	**Video type**	**MED (mm)**	**MSD (mm)**
PC	Food talk	2.41 ± 2.63	3.09 ± 2.02
TB	Food talk	2.66 ± 3.08	2.71 ± 1.51
PC	Random topic	1.58 ± 0.70	2.70 ± 0.63
TB	Random topic	1.65 ± 0.78	2.90 ± 1.08

Table 3.3 reveals that the two methods yield broadly comparable results. However, the palate correction method does not represent a universal improvement; rather, its utility is best characterized as a corrective method for challenging scenarios.

The method demonstrates significant efficacy in high-error scenarios, for instance reducing tendon tracking error in P02 (from 7.06 mm to 4.78 mm) (Figure 3.13a) and palate tracking error in P03 (from 7.18 mm to 3.74 mm) (Figure 3.13b). Conversely, the method's reliability is questionable, as it can degrade high-quality tracks; this is evidenced by P05 (Figure 3.13c), where palate error increased from 2.84 mm to 4.22 mm. This unreliability stems from the palate tracing algorithm, which may generate spurious traces when the palatal echo is not acoustically well-defined (e.g., when it incorrectly detects the tongue) or when the thresholding process divides the palate into multiple clusters. In the latter case, the algorithm may trace only a

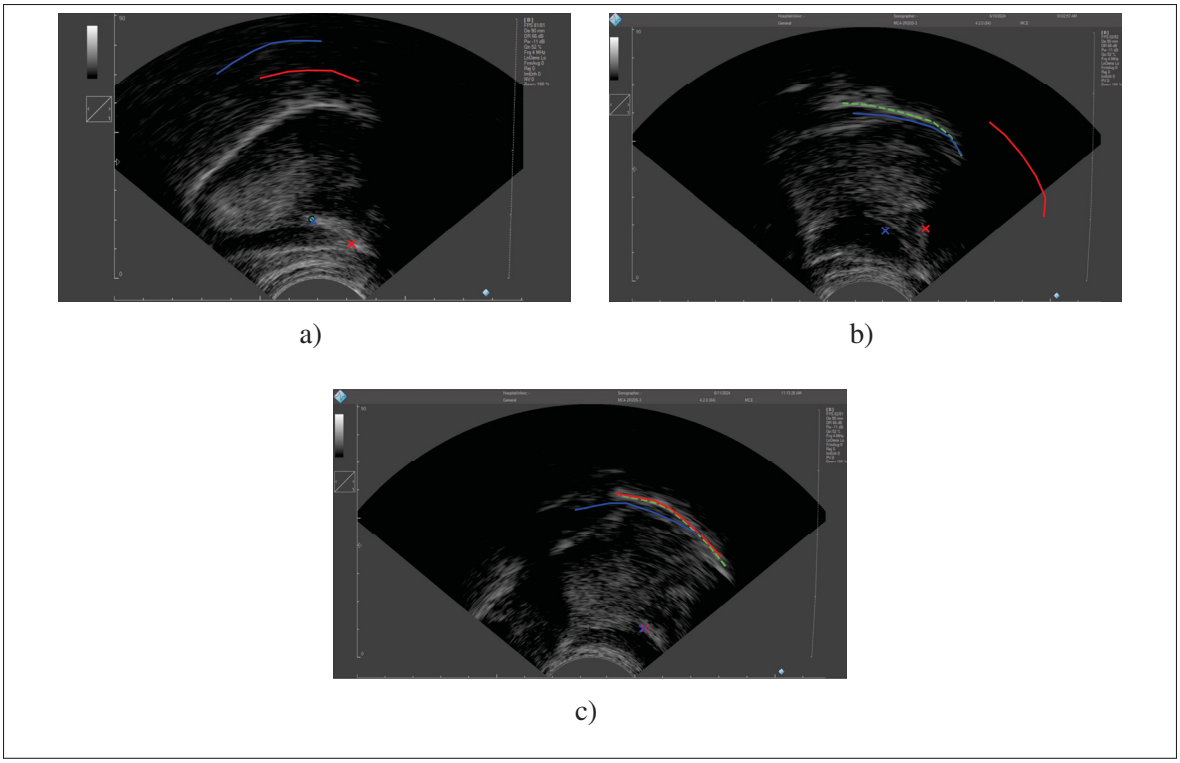


Figure 3.13 Visual comparison of Tendon-Based (TB) and Palate-Corrected (PC) tracking methods in free speech videos. Ground Truth : Green ; TB Prediction : Red ; PC Prediction : Blue. (a) P02 frame 4142. (b) P03 frame 820. (c) P05 frame 709

portion of the palate, causing the reference palate trace to be incorrectly aligned to this partial segment (as shown in Figure 3.13c). Consequently, the tendon-based method should be retained as the default processing pipeline. The palate correction method should be reserved for targeted intervention in specific challenging videos (e.g., P02, P04) where the tendon is poorly tracked but a stable palate contour is visibly available to serve as a reference.

3.6 Conclusions

To our knowledge, this study presents the first automated method for palate tracking in ultrasound speech imaging. By anchoring the displacement of the tendon of the genioglossus and palatal motion through our proposed two-step transformation mechanism (Equation 3.1), we demonstrate

that reliable palate tracking is achievable even when the palate remains acoustically invisible for over 95% of the recording duration. The proposed hybrid tracking system, combining YOLO tendon detection with particle filtering, achieves mean tracking errors between 1.24 mm and 2.77 mm across diverse speech and swallowing tasks, with performance strongly correlated to tendon tracking accuracy. The method's robustness is particularly evident in its ability to maintain tracking continuity through extended periods of palate invisibility during free speech, where traditional direct visualization approaches would fail entirely.

This work has meaningful implications for both clinical practice and research. Given the limitations in validating absolute positional accuracy during speech in the presence of significant jaw motion, the most immediate and suitable application is for enhanced visual biofeedback. For speech-language pathologists and their clients, the automated tracking could potentially provide a continuous, stable palatal reference without repeated calibration procedures. This may help reduce cognitive load for the learner and can improve the pedagogical effectiveness of ultrasound biofeedback, even if the tracked palate's absolute position is slightly offset by jaw movement. Particularly in L2 pronunciation training, having a stable palatal reference may allow learners to focus on functional tongue-palate relationships rather than compensating for image shifts, thereby improving their ability to establish accurate articulatory targets and develop self-monitoring skills for target speech sounds. The system could be further enhanced with real-time quality feedback, such as visual alerts when tendon visibility degrades, to prompt users to adjust probe positioning, ensuring optimal imaging quality during biofeedback sessions. In phonetics research, the method facilitates large-scale quantitative studies by eliminating manual annotation bottlenecks while providing frame-by-frame palatal reference over short time scales for coarticulation analysis and cross-speaker normalization. The significant intra and inter participant variability observed (ranging from sub-millimeter to 16 mm errors) highlights the method's sensitivity to both imaging conditions and anatomical differences in ultrasound-based speech analysis.

Future work should address several limitations. First, the tendon–palate motion model requires a more robust validation method. Our accuracy assessment is fundamentally limited, as validation

is only possible during brief moments of palate visibility. These depend on acoustic coupling, a condition that is often broken by larger jaw movements characteristic of free speech. Future studies could use external measurement devices, such as Electromagnetic Articulography, to independently track jaw and palate movement and precisely quantify the model's error. Second, the tendon detection model's robustness could be enhanced by exploring multi-landmark tracking to improve performance in poor imaging conditions. Third, generalizability should be established by testing the system on larger, more diverse datasets, including populations with speech disorders. Finally, integrating the palate tracker with existing tongue tracking systems would enable more comprehensive automated articulatory analysis. Despite these areas for improvement, this work establishes tendon-based inference as a viable paradigm for palate tracking in ultrasound tongue imaging.

CONCLUSION ET RECOMMANDATIONS

Ce mémoire a abordé deux défis fondamentaux dans l'analyse articulatoire par imagerie échographique. Premièrement, il a établi des pratiques exemplaires pour le traçage fiable du palais dur dans les images échographiques. Deuxièmement, il a proposé une méthode novatrice de suivi automatique et continu du palais, permettant de maintenir une référence anatomique stable même lorsque le palais est invisible acoustiquement pendant la parole. Ce travail a permis de franchir une étape clé vers des systèmes de biofeedback capables de générer une rétroaction corrective pour l'apprentissage et la prononciation d'une L2 et l'intervention clinique en orthophonie.

Les contributions principales de ce mémoire peuvent être résumées en trois volets. Premièrement, l'évaluation systématique des pratiques exemplaires pour le traçage manuel du palais a démontré que la déglutition sèche, couplée à l'utilisation du CES, maximise la reproductibilité inter-juges et la qualité des contours extraits. Cette découverte est particulièrement pertinente, car la déglutition sèche se produit spontanément et fréquemment en parole libre, offrant une opportunité naturelle pour l'initialisation ou la validation d'un système de suivi.

Deuxièmement, la validation de la méthode CES montre une erreur moyenne de 2,63 mm lors de la déglutition sèche, une valeur du même ordre de grandeur que la variabilité inter-juges. Ces résultats soutiennent l'utilisation du CES pour la délinéation statique du palais, tout en mettant en évidence une sensibilité aux artefacts liés aux milieux de couplage, tels que l'eau ou le yaourt.

Troisièmement ce mémoire a introduit une méthode novatrice de suivi automatique du palais, fondée sur l'inférence à partir du tendon du muscle génioglosse. En modélisant la relation entre le déplacement de ce tendon, constamment visible, et celui du palais comme une transformation rigide, et en combinant un détecteur YOLO avec un filtre particulaire pour le suivi temporel, le système parvient à maintenir un suivi continu du palais qui, lorsqu'il est validé aux rares moments où le palais réapparaît (moins de 5% des images en parole libre). Les expériences

menées sur 71 vidéos (51 déglutitions, 20 extraits de parole libre) montrent des erreurs MSD moyennes de 1,34 à 2,68 mm selon les tâches, avec une corrélation significative ($r = 0,64$, $p = 0,001$) entre la précision du suivi du tendon et celle du palais, validant ainsi l'hypothèse centrale de cette approche.

Ces avancées ouvrent la voie à des applications concrètes. Notre travail a mené au développement de ReaPT (Real-time Palate Tracker) (Aalto *et al.*, 2025a), une application transparente qui se superpose aux logiciels d'échographie, tels qu'EchoWave, afin de fournir un biofeedback visuel en temps réel. ReaPT permet aux orthophonistes de visualiser le contour du palais (extrait automatiquement à partir de vidéos de déglutition ou tracé manuellement) et d'ajouter des annotations personnalisées (flèches, points, contours) directement sur l'image échographique pendant les séances d'intervention. Lors d'une étude avec neuf apprenants adultes de l'anglais L2 (locuteurs natifs arabophones), les participants ont attribué des notes très élevées à la visualisation du palais (moyenne de 4,44/5), aux annotations personnalisées (5/5), et à l'utilité pédagogique globale (4,67/5), rapportant une meilleure conscience de leurs gestes articulatoires. L'orthophoniste qui administrait le biofeedback a observé que les apprenants développaient rapidement un sentiment de propriété sur la forme de leur propre palais, demandant systématiquement à ce qu'on utilise leur contour plutôt que celui d'un autre participant, et que les annotations ont particulièrement facilité l'acquisition des sons complexes comme /ɹ/ et /ɬ/, avec des progrès maintenus même après une pause de trois mois. Ce prototype fonctionnel démontre comment notre recherche fondamentale sur le suivi automatique du palais peut se traduire en outils cliniques accessibles, rendant le biofeedback échographique plus intuitif et pédagogiquement efficace pour l'apprentissage des langues et l'intervention orthophonique.

Malgré les résultats prometteurs, plusieurs pistes d'amélioration émergent clairement. L'hypothèse de transformation rigide, bien que validée dans les conditions de déglutition, peut être mise à mal par les mouvements mandibulaires importants en parole libre. Des travaux futurs devraient

explorer des modèles non rigides ou hybrides, potentiellement appris à partir de données multimodales, pour mieux capturer l'influence de la mâchoire. La performance du système repose fortement sur la détection fiable du tendon ; dans les cas où l'échogénicité du tendon est faible ou masquée par d'autres structures, le suivi peut dériver. L'exploration de modèles multi-points ou l'intégration de connaissances anatomiques a priori dans le réseau de neurones pourraient renforcer cette étape critique.

Bien que nous ayons exploré une forme de ré-initialisation par correction guidée par le palais, où le système exploite les apparitions spontanées du palais pour corriger les erreurs accumulées, cette approche présente des limitations liées à la fiabilité de l'extraction automatique du contour du palais à partir d'images uniques. Une stratégie de ré-initialisation alternative et potentiellement plus robuste consisterait à demander explicitement à l'apprenant d'effectuer une déglutition sèche lorsque le système détecte une incertitude élevée. Cette approche permettrait de recalibrer le contour en exploitant la méthode CES, qui s'est révélée plus fiable que l'extraction à partir d'images uniques. La déglutition étant un geste simple, non intrusif et fréquent durant la parole libre, cette stratégie s'intégrerait naturellement dans les séances de biofeedback tout en offrant une ré-initialisation de meilleure qualité.

Toutes les expériences ont été menées sur des locuteurs sains. Il est essentiel de tester et d'adapter le système sur des populations présentant des troubles de l'articulation, où les schémas de mouvement de la langue et du tendon peuvent différer. Le suivi du palais gagne toute sa valeur lorsqu'il est combiné à un suivi précis de la langue. L'intégration de notre méthode avec des systèmes de suivi lingual de pointe permettrait de créer un outil d'analyse articulatoire complet, fournissant à la fois la dynamique de la langue et son cadre de référence. Bien que conçu pour le traitement hors ligne, le pipeline pourrait être optimisé pour fonctionner en temps réel, rendant le biofeedback immédiat et interactif.

Une extension de ce travail consisterait à évaluer la fiabilité du suivi automatique dans des contextes phonétiques variés. Si le palais constitue une cible évidente pour les consonnes occlusives, son rôle demeure tout aussi important pour les sons sans contact linguo-palatal, tels que les voyelles ou les fricatives (comme /s/). Dans ces configurations, le palais délimite la paroi supérieure du conduit vocal, et sa morphologie individuelle influence les stratégies articulatoires mises en œuvre pour atteindre des géométries de conduit comparables. La validation de la précision du modèle de suivi sur une large gamme de phonèmes permettrait ainsi de mieux comprendre comment les apprenants adaptent la position de leur langue en fonction de leur propre cadre anatomique palatin.

En conclusion, ce mémoire aborde une limitation bien connue de l'imagerie échographique, à savoir l'invisibilité fréquente du palais. Les travaux présentés montrent qu'il est possible d'inférer un repère anatomique palatin à partir d'une structure systématiquement visible dans l'image. Cette approche ouvre la voie au développement d'outils automatisés d'analyse articulatoire, avec des applications potentielles en recherche et en pédagogie.

BIBLIOGRAPHIE

Aalto, E., Ben Asker, H., Farazi, S., Ménard, W. C. L. & Laporte, C. (2025a). User experiences of palate display and annotation tools in lingual ultrasound biofeedback. *Proceedings of the first meeting on Methods and Techniques in Phonetic Sciences (MaTiPS)*.

Aalto, E. M., Ben Asker, H., Ménard, L., Cardoso, W. & Laporte, C. (2025b). Ultrasound imaging in second language research : Systematic review and thematic analysis. *Speech Communication*, 175, 103324. doi : https://doi.org/10.1016/j.specom.2025.103324.

Akgul, Y., Kambhamettu, C. & Stone, M. (1999). Automatic extraction and tracking of the tongue contours. *IEEE Transactions on Medical Imaging*, 18(10), 1035-1045. doi : 10.1109/42.811315.

Al Ani, S., Cleland, J. & Zoha, A. (2025). Deep learning in ultrasound tongue imaging : a systematic review toward automated detection of speech sound disorders. *Frontiers in Artificial Intelligence*, 8, 1631134.

Ben Asker, H., Aalto, E. M. A., Ménard, L., Cardoso, W. & Laporte, C. (2025). Best practices for tracing the palate in ultrasound images. *Proceedings of the 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*.

Bernhardt, B. M., Gick, B., Bacsfalvi, P. & Ashdown, J. (2003). Speech habilitation of hard of hearing adolescents using electropalatography and ultrasound as evaluated by trained listeners. *Clinical Linguistics & Phonetics*, 17(3), 199–216.

Bernhardt, B. M., Bacsfalvi, P. & Gick, B. (2005). Exploring the use of electropalatography and ultrasound in speech habilitation. *The Journal of Speech and Language Pathology and Audiology*, 29(4), 141–188.

Bliss, H., Abel, J. & Gick, B. (2018). Computer-assisted visual articulation feedback in L2 pronunciation instruction : A review. *Journal of Second Language Pronunciation*, 4(1), 129–153.

Brunner, J., Fuchs, S. & Perrier, P. (2009). On the relationship between palate shape and articulatory behavior. *The Journal of the Acoustical Society of America*, 125(6), 3936–3949.

Bryfonski, L. (2023). Is seeing believing ? The role of ultrasound tongue imaging and oral corrective feedback in L2 pronunciation development. *Journal of Second Language Pronunciation*, 9(1), 103–129.

Cakir, A., Amasyali, M. F. & Genc, A. E. (2021). Confidence-based MOT : A Score-based Tracklet Management. *arXiv preprint arXiv :2107.04327*.

Chen, S., Whalen, D. & Mok, P. P. K. (2024). Production of the English /ɹ/ by Mandarin–English Bilingual Speakers. *Language and Speech*, 00238309241230895.

Chen, W.-R., Tiede, M. & Whalen, D. (2020). DeepEdge : Automatic ultrasound tongue contouring combining a deep neural network and an edge detection algorithm. *12th International Seminar on Speech Production (ISSP 2020)*.

Cleland, J. & Scobbie, J. M. (2021). The dorsal differentiation of velar from alveolar stops in typically developing children and children with persistent velar fronting. *Journal of Speech, Language, and Hearing Research*, 64(6S), 2347–2362.

Cleland, J., Scobbie, J. M., Nakai, S. & Wrench, A. A. (2015a). Helping children learn non-native articulations : The implications for ultrasound-based clinical intervention. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS), Glasgow, 10-14 August 2015*.

Cleland, J., Scobbie, J. M. & Wrench, A. A. (2015b). Using ultrasound visual biofeedback to treat persistent primary speech sound disorders. *Clinical linguistics & phonetics*, 29(8-10), 575–597.

Csapó, T. G. & Lulich, S. M. (2015). Error analysis of extracted tongue contours from 2D ultrasound images. *Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015)*, pp. 2157–2161. doi : 10.21437/Interspeech.2015-486.

Csapó, T. G., Grósz, T., Gosztolya, G., Tóth, L. & Markó, A. (2017). DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface. *INTERSPEECH*, pp. 3672–3676.

Davidson, L. (2006). Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *The Journal of the Acoustical Society of America*, 120(1), 407–415.

Di Bella, L., Lyu, Y., Cornelis, B. & Munteanu, A. (2025). HybridTrack : A Hybrid Approach for Robust Multi-Object Tracking. *arXiv preprint arXiv :2501.01275*.

Diekhoff, M. R. & Lulich, S. M. (2022). Conceptualizations of the articulation of rhotic sounds in American English and the role of clinical experience in their formation. *Perspectives of the ASHA Special Interest Groups*, 7(4), 1256–1274.

Epstein, M. A. & Stone, M. (2005). The tongue stops here : Ultrasound imaging of the palate. *The Journal of the Acoustical Society of America*, 118(4), 2128–2131.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *kdd*, 96(34), 226–231.

Fasel, I. & Berry, J. (2010). Deep belief networks for real-time extraction of tongue contours from ultrasound during speech. *2010 20th International Conference on Pattern Recognition*, pp. 1493–1496.

Faucher, G., Karimi, E., Ménard, L. & Laporte, C. (2019). Automatic palate delineation in ultrasound videos. *Proceedings of the 19th International Congress of Phonetic Sciences. Editors S. Calhoun, P. Escudero, M. Tabain, and P. Warren (Melbourne)*, pp. 422–426.

Feng, M., Wang, Y., Xu, K., Wang, H. & Ding, B. (2021). Improving ultrasound tongue contour extraction using U-Net and shape consistency-based regularizer. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6443–6447.

Fischler, M. A. & Bolles, R. C. (1981). Random sample consensus : a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 381–395.

Frangi, A. F., Niessen, W. J., Vincken, K. L. & Viergever, M. A. (1998). Multiscale vessel enhancement filtering. *Medical Image Computing and Computer-Assisted Intervention—MICCAI'98 : First International Conference Cambridge, MA, USA, October 11–13, 1998 Proceedings 1*, pp. 130–137.

Gibbon, F. E. (1999). Undifferentiated lingual gestures in children with articulation/phonological disorders. *Journal of Speech, Language, and Hearing Research*, 42(2), 382–397.

Gick, B., Bird, S. & Wilson, I. (2005). Techniques for field application of lingual ultrasound imaging. *Clinical linguistics & phonetics*, 19(6-7), 503–514.

Gick, B., Bernhardt, B., Bacsfalvi, P., Wilson, I., Zampini, M. et al. (2008). Ultrasound imaging applications in second language acquisition. *Phonology and second language acquisition*, 36(6), 309–322.

Hardcastle, W. J., Gibbon, F. E. & Jones, W. (1991). Visual display of tongue-palate contact : Electropalatography in the assessment and remediation of speech disorders. *British Journal of Disorders of Communication*, 26(1), 41–74.

Heyne, M. & Derrick, D. (2015). Using a radial ultrasound probe's virtual origin to compute midsagittal smoothing splines in polar coordinates. *The Journal of the Acoustical Society of America*, 138(6), EL509–EL514.

Hixon, T. J., Weismer, G. & Hoit, J. D. (2018). *Preclinical speech science : Anatomy, physiology, acoustics, and perception*. Plural Publishing.

Jocher, G., Chaurasia, A. & Qiu, J. (2023). Ultralytics YOLOv8 (Version 8.0.0). Repéré à https://github.com/ultralytics/ultralytics.

Karimi, E., Ménard, L. & Laporte, C. (2019). Fully-automated tongue detection in ultrasound images. *Computers in biology and medicine*, 111, 103335.

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H. & Golestani, N. (2015). The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *The journal of the acoustical society of America*, 138(2), 817–832.

Kochetov, A. (2020). Research methods in articulatory phonetics I : Introduction and studying oral gestures. *Language and Linguistics Compass*, 14(4), e12368. doi : https://doi.org/10.1111/lnc3.12368.

Kovesi, P. et al. (1999). Image features from phase congruency. *Videre : Journal of computer vision research*, 1(3), 1–26.

Lammert, A., Proctor, M. & Narayanan, S. (2013). Interspeaker variability in hard palate morphology and vowel production. *Journal of Speech, Language, and Hearing Research*, 56(6), 1924–1933.

Laporte, C., Takla, R., Tiede, M. & Ménard, L. (2022). Imaging the palate : best methods and practices.

Laporte, C. & Ménard, L. (2018). Multi-hypothesis tracking of the tongue surface in ultrasound video recordings of normal and impaired speech. *Medical Image Analysis*, 44, 98–114.

Lee, A., Liker, M., Fujiwara, Y., Yamamoto, I., Takei, Y. & Gibbon, F. (2023). EPG research and therapy : further developments. *Clinical linguistics & phonetics*, 37(8), 701–721.

Lee, e. a. (2023). Effects of Production Training With Ultrasound Biofeedback on Second-Language (L2) Production Training of English Tense–Lax Vowel Contrasts. Repéré à PublishedinJournalofSpeech,Language,andHearingResearch.

Li, J. J., Ayala, S., Harel, D., Shiller, D. M. & McAllister, T. (2019). Individual predictors of response to biofeedback training for second-language production. *The Journal of the Acoustical Society of America*, 146(6), 4625–4643.

Li, M., Kambhamettu, C. & Stone, M. (2005). Automatic contour tracking in ultrasound images. *Clinical linguistics & phonetics*, 19(6-7), 545–554.

Lin, S. (2021). Observing and Measuring Speech Articulation. Dans Knight, R.-A. & Setter, J. (Éds.), *The Cambridge Handbook of Phonetics* (pp. 362–386). Cambridge University Press.

Martinez-del Rincon, J., Nebel, J.-C., Makris, D. & Orrite-Urunuela, C. (2008). Tracking Human Body Parts Using Particle Filters Constrained by Human Biomechanics. *Proceedings of the British Machine Vision Conference (BMVC)*.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W. & Bethge, M. (2018). DeepLabCut : markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9), 1281–1289.

McAllister, T., Preston, J. L., Hitchcock, E. R. & Hill, J. (2020). Protocol for correcting residual errors with spectral, ultrasound, traditional speech therapy randomized controlled trial (C-RESULTS RCT). *BMC pediatrics*, 20(1), 66.

McAllister Byun, T. & Campbell, H. (2016). A Case Study on the Efficacy of Ultrasound Biofeedback in Voice Pedagogy. *Journal of Singing*, 73(1), 39–49.

McCabe, P., Preston, J. L., Evans, P. & Heard, R. (2023). A pilot randomized control trial of motor-based treatments for childhood apraxia of speech : Rapid Syllable Transition Treatment and ultrasound biofeedback. *American Journal of Speech-Language Pathology*, 32(2), 629–644.

Micucci, M. & Iula, A. (2022). Recent advances in machine learning applied to ultrasound imaging. *Electronics*, 11(11), 1800.

Mielke, J. (2015). An ultrasound study of Canadian French rhotic vowels with polar smoothing spline comparisons. *The journal of the Acoustical Society of America*, 137(5), 2858–2869.

Mielke, J., Baker, A., Archangeli, D. & Racy, S. (2005). Palatron : a technique for aligning ultrasound images of the tongue and palate.

Mozaffari, M. H. & Lee, W.-S. (2020). Encoder-decoder CNN models for automatic tracking of tongue contours in real-time ultrasound data. *Methods*, 179, 26–36.

Noiray, A., Ménard, L. & Iskarous, K. (2013). The development of motor synergies in children : Ultrasound and acoustic measurements. *The Journal of the Acoustical Society of America*, 133(1), 444–452.

Nota, Y., Kitamura, T., Takemoto, H. & Maekawa, K. (2024). Mapping palatal shape to electromagnetic articulography data : An approach using 3D scanning and sensor matching. *JASA Express Letters*, 4(1).

Ostu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Trans SMC*, 9, 62.

Ouni, S. (2014). Tongue control and its implication in pronunciation training. *Computer Assisted Language Learning*, 27(5), 439–453.

Patrick, K., Fricke, S., Rutter, B. & Cleland, J. (2024). Clinical application of usage-based phonology : treatment of cleft palate speech using usage-based electropalatography. *International Journal of Speech-Language Pathology*, 26(4), 595–610.

Preston, J. L., Brick, N. & Landi, N. (2013). Ultrasound biofeedback treatment for persisting childhood apraxia of speech. *American Journal of Speech-Language Pathology*, 22(4), 627–643.

Preston, J. L., Leece, M. C. & Maas, E. (2016). Intensive treatment with ultrasound visual feedback for speech sound errors in childhood apraxia. *Frontiers in human neuroscience*, 10, 440.

Preston, J. L., Leece, M. C. & Maas, E. (2017). Ultrasound Biofeedback in Clinical Practice. Repéré à ClinicalReport.

Ribeiro, M. S., Cleland, J., Eshky, A., Richmond, K. & Renals, S. (2021). Exploiting ultrasound tongue imaging for the automatic detection of speech articulation errors. *Speech Communication*, 128, 24–34.

Ronneberger, O., Fischer, P. & Brox, T. (2015). U-net : Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241.

Roussos, A., Katsamanis, A. & Maragos, P. (2009). Tongue tracking in ultrasound images with active appearance models. *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 1733–1736.

Song, J. Y. & Eckman, F. (2021). Using ultrasound tongue imaging to study covert contrasts in second-language learners' acquisition of English vowels. *Language Acquisition*, 28(4), 344–369.

Stone, M. (2005). A guide to analysing tongue motion from ultrasound images. *Clinical linguistics & phonetics*, 19(6-7), 455–501.

Sugden, E., Lloyd, S., Lam, J. & Cleland, J. (2019). Systematic review of ultrasound visual biofeedback in intervention for speech sound disorders. *International journal of language & communication disorders*, 54(5), 705–728.

Tang, L., Bressmann, T. & Hamarneh, G. (2012). Tongue contour tracking in dynamic ultrasound via higher-order MRFs and efficient fusion moves. *Medical Image Analysis*, 16(8), 1503–1520.

Toutios, A., Byrd, D., Goldstein, L. & Narayanan, S. (2019). Advances in vocal tract imaging and analysis. Dans *The Routledge handbook of phonetics* (pp. 34–50). Routledge.

Verhoeven, J., Miller, N. R., Daems, L. & Reyes-Aldasoro, C. C. (2019). Visualisation and analysis of speech production with electropalatography. *Journal of Imaging*, 5(3), 40.

Wrench, A. & Balch-Tomes, J. (2022). Beyond the Edge : Markerless Pose Estimation of Speech Articulators from Ultrasound and Camera Images Using DeepLabCut. *Sensors*, 22(3). doi : 10.3390/s22031133.

Wrench, A. A. (2017). Real-time tongue contour fitting and vocal tract carving. *Ultrafest VIII*, pp. 99–100.

Xu, K., Yang, Y., Stone, M., Jaumard-Hakoun, A., Leboullenger, C., Dreyfus, G., Roussel, P. & Denby, B. (2016). Robust contour tracking in ultrasound tongue image sequences. *Clinical linguistics & phonetics*, 30(3-5), 313–327.

Zhao, C., Zhang, P., Zhu, J., Wu, C., Wang, H. & Xu, K. (2019). Predicting tongue motion in unlabeled ultrasound videos using convolutional LSTM neural networks. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5926–5930.

Zharkova, N. (2018). An ultrasound study of the development of lingual coarticulation during childhood. *Phonetica*, 75(3), 245–271.

Zharkova, N., Hewlett, N. & Hardcastle, W. J. (2011). Coarticulation as an indicator of speech motor control development in children : An ultrasound study. *Motor Control*, 15(1), 118–140.

Zharkova, N., Hewlett, N. & Hardcastle, W. J. (2012). An ultrasound study of lingual coarticulation in /sV/ syllables produced by adults and typically developing children. *Journal of the International Phonetic Association*, 42(2), 193–208.

Zhu, J., Styler, W. & Calloway, I. (2019). A CNN-based tool for automatic tongue contour tracking in ultrasound images. *arXiv preprint arXiv :1907.10210*.