Détection de la déglutition avec un microphone intra-auriculaire

par

Elyes Ben Cheikh

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 LA SANTÉ
M. Sc. A.

MONTRÉAL, LE 24 JUIN 2026

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE:

Prof Rachel E. Bouserhal, directrice de mémoire
Département de génie électrique, École de technologie supérieure

Prof Catherine Laporte, codirectrice
Département de génie électrique, École de technologie supérieure

Prof Nicola Hagemeister, présidente du jury
Département de génie des systèmes, École de technologie supérieure

Prof Julien Clément, membre du jury
Département de génie des systèmes, École de technologie supérieure

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

LE 10 JUIN 2026

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Mes premiers remerciements vont à ma directrice de recherche, Rachel Bouserhal. Je te suis profondément reconnaissant pour la confiance que tu m'as accordée dès le début, ainsi que pour ta bienveillance, ton respect, ton humilité et ton soutien indéfectible.

Un remerciement tout particulier à ma co-directrice, Catherine Laporte, pour son encadrement technique précieux tout au long de cette recherche. Je lui suis reconnaissant pour nos échanges stimulants. Merci pour ton soutien, ton expertise et ta gentillesse.

Je souhaite remercier ma famille pour son amour et son soutien. Merci à toi, Maman, pour ton amour sans limites, ton soutien constant, et pour avoir toujours rendu possible mes retours à la maison. Un merci sincère à mes frères pour leurs encouragements continus. J'espère qu'un jour j'aurai suffisamment de temps pour partager davantage de moments avec vous et concrétiser tous ces projets ensemble.

Enfin, ce travail a été soutenu par la Chaire de recherche Marcelle Gauvreau en génie sur le suivi multimodal de la santé et la détection précoce des maladies à l'École de technologie supérieure, ainsi que le programme OPSIDIAN financé par le CRSNG.

Détection de la déglutition avec un microphone intra-auriculaire

Elyes Ben Cheikh

RÉSUMÉ

La déglutition figure parmi les fonctions motrices les plus fréquemment sollicitées de l'organisme et constitue un marqueur clinique majeur de la santé neurologique et comportementale. Tant sa fréquence que sa dynamique temporelle sont altérées par la dysphagie, les maladies neurodégénératives et le vieillissement, et se trouvent également modulées par l'état émotionnel, la charge cognitive et la prise alimentaire.

Les outils cliniques de référence fournissent des observations brèves de la déglutition en environnement contrôlé et ne permettent pas une surveillance continue dans la vie quotidienne. Les capteurs portés au cou, bien que prometteurs, restent limités par leur faible acceptabilité sociale pour un usage prolongé.

Les *hearables*, dispositifs portables conçus pour l'oreille, constituent une plateforme socialement acceptable pour la surveillance physiologique non intrusive. En particulier, les microphones intra-auriculaires, placés dans le conduit auditif occlus, exploitent l'effet d'occlusion pour amplifier les sons transmis par voie osseuse et tissulaire, dont les sons de déglutition. Cependant, le signal capté est un mélange bioacoustique hétérogène (respiration, bruits cardiovasculaires, parole, mastication, mouvements physiques) au sein duquel la déglutition constitue une classe d'événement minoritaire, brève (durée moyenne ~860 ms) et fortement variable entre sujets. À ce défi s'ajoute l'absence de bases de données publiques d'audio intra-auriculaire synchronisé avec des annotations de déglutition validées par des mesures de référence.

Ce mémoire propose un cadre complet de détection de la déglutition à partir de signaux audio intra-auriculaires bilatéraux, structuré autour de deux contributions complémentaires. La première est la conception et la validation d'une base de données multimodale collectée auprès de 34 adultes sains (14 femmes, 20 hommes, âgés de 20 à 29 ans), comprenant des enregistrements synchronisés de signaux audio intra- et extra-auriculaires ainsi que des enregistrements multimodaux provenant d'une ceinture thoracique (électrocardiogramme, signaux respiratoires et accélérométrie), et d'imagerie échographique linguale servant de référence pour l'annotation. Le protocole couvre des tâches verbales et non verbales sous trois conditions acoustiques (silencieuse, bruit de foule, bruit d'usine) et plusieurs types de bolus (sec, eau, yaourt). En tant que preuve de concept, la première contribution présente un réseau entièrement connecté entraîné sur des représentations vectorielles de Yet Another Mobile Network (YAMNet) complétées par le taux de passages par zéro, atteint un score F1 de 0,875 ± 0,013 pour la classification binaire de la déglutition. La seconde contribution est un système de détection d'événements basé sur une architecture de réseau de neurones convolutif récurrent (CRNN) bilatérale opérant sur des spectrogrammes de normalisation d'énergie par canal (PCEN), avec fusion adaptative par confiance des deux oreilles et post-traitement par hystérésis.

Évalué sur les 34 participants par validation croisée à cinq plis disjoints au niveau sujet, le système atteint un score F1 de classification au niveau fenêtre de 0,892 ± 0,006 et un score F1 de détection au niveau événement de 0,835 ± 0,012, avec une précision de 0,779 ± 0,014 et un rappel de 0,900 ± 0,020, le rappel dépassant systématiquement 0,870 dans toutes les conditions de bruit. Ces résultats démontrent la faisabilité d'une détection robuste de la déglutition à partir d'audio intra-auriculaire en conditions acoustiques réalistes et ouvrent la voie à un suivi longitudinal de la dysphagie et de la sialorrhée à l'aide d'appareils auditifs.

Mots-clés: Bioacoustique, Déglutition, Microphones Intra-auriculaires, CRNN, Effet d'occlusion, PCEN, Fusion bilatérale

Swallowing Detection with an in-ear microphone

Elyes Ben Cheikh

ABSTRACT

Swallowing is one of the most frequently performed motor functions in the body and constitutes a major clinical marker of neurological and behavioural health. Both its frequency and temporal dynamics are altered by dysphagia, neurodegenerative diseases and ageing, and are also modulated by emotional state, cognitive load and food intake. Reference clinical tools provide brief observations of swallowing in controlled environments and do not allow continuous monitoring in daily life. Neck-worn sensors, although promising, remain limited due to limited social acceptability for prolonged use.

Hearables, wearable devices designed for the ear, constitute a socially acceptable platform for non-invasive physiological monitoring. In particular, in-ear microphones, placed in the occluded ear canal, exploit the occlusion effect to amplify sounds transmitted via bone and tissue conduction, including swallowing sounds. However, the captured signal is a heterogeneous bioacoustic mixture (breathing, cardiovascular sounds, speech, chewing, physical movements) within which swallowing constitutes a minority event class, brief (average duration ~860 ms) and highly variable between subjects. Added to this challenge is the absence of public databases of in-ear audio synchronised with swallowing annotations validated by reference measurements.

This thesis proposes a complete framework for swallowing detection from bilateral in-ear audio signals, structured around two complementary contributions. The first is the design and validation of a multimodal database collected from 34 healthy adults (14 women, 20 men, aged 20 to 29 years), comprising synchronised recordings of in-ear and outer-ear audio signals as well as multimodal recordings from a thoracic belt (electrocardiogram, respiratory signals and accelerometry), and tongue ultrasound imaging serving as annotation reference. The protocol covers verbal and non-verbal tasks under three acoustic conditions (silent, babble noise, factory noise) and several bolus types (dry, water, yoghurt). As a proof of concept, the first contribution presents a fully connected network trained on Yet Another Mobile Network (YAMNet) vector representations supplemented by the zero-crossing rate, which achieves an F1 score of 0.875 ± 0.013 for binary swallowing classification.

The second contribution is an event detection system based on a bilateral Convolutional Recurrent Neural Network (CRNN) architecture operating on Per-Channel Energy Normalization (PCEN) spectrograms, with confidence-based adaptive fusion of both ears and hysteresis post-processing. Evaluated on the 34 participants by five-fold subject-disjoint cross-validation, the system achieves a window-level classification F1 score of 0.892 ± 0.006 and an event-level detection F1 score of 0.835 ± 0.012, with a precision of 0.779 ± 0.014 and a recall of 0.900 ± 0.020, recall consistently exceeding 0.870 across all noise conditions. These results demonstrate the feasibility of robust swallowing detection from in-ear audio in realistic acoustic conditions and pave the way for longitudinal monitoring of dysphagia and sialorrhoea using hearables.

Keywords: Bioacoustics, Swallowing, In-ear microphones, CRNN, Occlusion effect, PCEN, Bilateral fusion

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

Adam	Estimation adaptative des moments
ANC	Contrôle actif du bruit
AUPRC	Aire sous la courbe précision-rappel
AUROC	Aire sous la courbe ROC
BCE	Entropie croisée binaire
BiGRU	Bidirectional Gated Recurrent Unit
BiLSTM	Bidirectional Long Short-Term Memory
BoAW	Sac de mots audio
BTS	Son de transit du bolus
CER	Comité d'éthique pour la recherche
CNN	Réseau de neurones convolutif
CRNN	Réseau de neurones convolutif récurrent
CWT	Transformée en ondelettes continue
ECG	Électrocardiogramme
EEG	Électroencéphalogramme
EMA	Moyenne mobile exponentielle
EMG	Électromyographie
FCNN	Réseau de neurones entièrement connecté
FDS	Son discret final
FEES	Endoscopie par fibre optique
FN	Faux négatif
FP	Faux positif
GMM	Modèle de mélange de gaussiennes
GRU	Gated Recurrent Unit
HHT	Transformée de Hilbert-Huang

HMM	Modèle de Markov caché
HRCA	Auscultation cervicale haute résolution
iBad	In-ear Biosignal Audio Database
ICC	Coefficient de corrélation intraclasse
IDS	Son discret initial
IEM	Microphone intra-auriculaire
IMU	Unité d'inertie
IoU	Intersection sur union
IRB	Comité d'éthique institutionnel
LSTM	Long Short-Term Memory
MFCC	Coefficients cepstraux sur échelle de Mel
MLP	Perceptron multicouche
MTF	Champ de transition de Markov
OBS	Open Broadcaster Software
OEM	Microphone extra-auriculaire
PCEN	Normalisation d'énergie par canal
PPG	Photopléthysmographie
ReLU	Unité linéaire rectifiée
RMS	Moyenne quadratique
sEMG	Électromyographie de surface
SNR	Rapport signal-sur-bruit
SSF	Fréquence de déglutition spontanée
STFT	Transformée de Fourier à court terme
SVM	Machine à vecteurs de support
TN	Vrai négatif
TP	Vrai positif
UES	Sphincter œsophagien supérieur

VFSS	Vidéofluoroscopie de la déglutition
YAMNet	Yet Another Mobile Network
ZCR	Taux de passages par zéro

LISTE DES SYMBOLES ET UNITÉS DE MESURE

dB	Décibel
Hz	Hertz
kHz	Kilohertz
s	Seconde
ms	Milliseconde
min	Minute
mL	Millilitre
fps	Images par seconde
T_w	Durée de la fenêtre d'analyse
T_h	Pas (*hop*) de la fenêtre glissante
f_{min}	Fréquence minimale du banc de filtres Mel
f_{max}	Fréquence maximale du banc de filtres Mel
n_{mels}	Nombre de bandes Mel
n_{fft}	Taille de la transformée de Fourier
τ_{low}	Seuil bas de l'hystérésis de détection
τ_{high}	Seuil haut de l'hystérésis de détection

INTRODUCTION

0.1 Contexte

La déglutition est l'une des fonctions motrices les plus sollicitées de l'organisme : un adulte sain effectue en moyenne 580 à 600 déglutitions au cours d'une journée, tant pour s'alimenter que pour évacuer en continu sa salive et assurer la protection des voies aériennes (Lear, Jr & Moorrees, 1965; Kapila, Dodds, Helm & J., 1984). Cette fonction repose sur la coordination fine de près de trente paires de muscles et de six paires de nerfs crâniens, organisée en trois phases séquentielles : orale, pharyngée et œsophagienne dont la bonne articulation conditionne à la fois la nutrition, l'hydratation et la clairance des voies aériennes supérieures (Matsuo & Palmer, 2008). La complexité neuromusculaire sous-jacente rend la fonction de déglutition particulièrement vulnérable : toute lésion le long de l'axe cortico-bulbaire ou toute altération des effecteurs périphériques se traduit par une perturbation qualitative ou quantitative du geste.

Ce trouble, regroupé sous le terme de dysphagie, ne constitue pas une pathologie isolée mais un symptôme transversal à un large spectre de conditions cliniques. Dans les maladies neurologiques, la dysphagie est particulièrement fréquente après un accident vasculaire cérébral, où elle touche jusqu'à la moitié des patients en phase initiale et s'associe à un risque accru de pneumonie d'aspiration, de malnutrition et de mortalité (Martino *et al.*, 2005). Dans les maladies neurodégénératives, la dysphagie est rapportée chez 80 % ou plus des patients atteints de la maladie de Parkinson à un stade avancé, avec pour corollaire une insuffisance de la clairance salivaire fréquemment confondue à tort avec une hypersalivation (Chou, Evatt, Hinson & Kompoliti, 2007; Suttrup & Warnecke, 2016). Elle constitue également un élément diagnostique et pronostique majeur de la sclérose latérale amyotrophique (Ruoppolo *et al.*, 2013), accompagne la progression de la maladie d'Alzheimer et d'autres démences (Alagiakrishnan, Bhanji & Kurian, 2013) et s'observe de manière fréquente dans la sclérose en plaques (Calcagno, Ruoppolo, Grasso, Vincentiis & Paolucci, 2002). Au-delà du cadre neurologique, la dysphagie

est une complication attendue des traitements des cancers de la tête et du cou, qu'il s'agisse des séquelles chirurgicales ou de la toxicité de la radiothérapie (Nguyen *et al.*, 2004). Enfin, elle peut résulter du vieillissement lui-même sous la forme de la *presbyphagie*, qui résulte du déclin musculaire, de la modification des seuils de sensibilité sensorielle et de la perte de coordination neuromusculaire liés à l'âge, et qui prédispose à la dysphagie franche en présence de tout événement surajouté (Humbert & Robbins, 2008; Namasivayam & Steele, 2015). L'ensemble de ces situations cliniques partage un faisceau de conséquences délétères similaires : aspiration silencieuse, pneumopathies, malnutrition, déshydratation et réduction marquée de la qualité de vie. Sur le plan sanitaire et économique, le coût global de la dysphagie a été estimé à plus d'un milliard de dollars américains par an aux seuls États-Unis, essentiellement en raison des hospitalisations prolongées qu'elle occasionne (Altman, Yu & Schaefer, 2010).

La surveillance clinique de la déglutition n'est toutefois pas limitée au diagnostic de la dysphagie. La fréquence et la dynamique temporelle des déglutitions constituent également des indicateurs comportementaux et autonomiques : le rythme des déglutitions spontanées est modulé par l'état émotionnel, l'anxiété, la charge cognitive, l'attention ou encore la prise alimentaire et la stimulation olfactive (Cuevas, III, Richter, McCutcheon & Taub, 1995; Kapila *et al.*, 1984). Cette dimension fait de la déglutition un marqueur d'intérêt non seulement pour la rééducation en phoniatrie et la nutrition clinique, mais aussi pour la caractérisation d'états comportementaux dans des contextes ambulatoires. La fréquence de déglutition spontanée est ainsi un indicateur potentiellement sensible mais très variable. Chez des sujets sains, les taux rapportés dans la littérature s'étendent d'environ 0,1 à près de 7 déglutitions par minute pendant l'éveil, selon les populations et les méthodes de mesure (Bulmer, Ewers, Drinnan & Ewan, 2021).

Les enjeux liés au dépistage de la dysphagie dans diverses pathologies, au suivi longitudinal des patients à risque, ainsi qu'à la caractérisation autonomique et comportementale en contexte non contrôlé plaident pour le développement de solutions de surveillance : (i) non invasives,

compatibles avec un usage quotidien, (ii) robustes à la variabilité inter-sujets, (iii) capables de fonctionner en continu en dehors de l'environnement clinique. La plateforme des *hearables*, déjà socialement acceptée à travers l'usage massif d'écouteurs intra-auriculaires, constitue à cet égard un vecteur privilégié, dont l'exploitation pour la détection de la déglutition motive le présent travail.

0.2 Problématique

L'évaluation clinique de la déglutition repose aujourd'hui sur des techniques instrumentales telles que la vidéofluoroscopie de la déglutition (VFSS), l'endoscopie par fibre optique (FEES), la manométrie pharyngée, l'échographie sous-mentale, l'électromyographie de surface ou l'auscultation cervicale, qui demeurent confinées à des environnements spécialisés et ne fournissent que de brèves observations dans des conditions contrôlées (Logemann, 1998; Langmore, Schatz & Olsen, 1988; Hiramatsu, Kataoka, Osaki & Hagino, 2015). Au-delà de ces évaluations de référence, la fréquence de déglutition spontanée a été proposée comme indicateur non invasif du risque de dysphagie. Toutefois, une revue systématique de 19 études a révélé qu'aucune méthode standardisée n'existe pour mesurer cette fréquence, et que les taux rapportés diffèrent si substantiellement entre les technologies de mesure et les protocoles que les données provenant de différents centres ne peuvent être comparées de manière fiable (Bulmer *et al.*, 2021). Ce manque de standardisation souligne le besoin de systèmes de détection de la déglutition robustes et automatisés, capables de fournir des mesures cohérentes et reproductibles entre les séances et les individus.

Pour permettre une surveillance plus continue, plusieurs approches de détection ont été proposées, dont des capteurs placés sur le cou, notamment des microphones, des accéléromètres et des capteurs d'électromyographie de surface, afin de capter les signaux acoustiques, vibratoires ou musculaires associés à la déglutition (So *et al.*, 2023). Bien que ces approches soient prometteuses en conditions contrôlées, elles présentent des limites pratiques : la dépendance au

positionnement et la faible acceptabilité d'un port visible prolongé restreignent leur usage en dehors du laboratoire.

La détection intra-auriculaire a récemment émergé comme une alternative prometteuse pour la surveillance physiologique non intrusive. Les dispositifs auriculaires, souvent désignés sous le terme de *hearables*, offrent une plateforme discrète et socialement acceptable pour la détection continue. Les études sur la déglutition captée par des microphones intra-auriculaires demeurent cependant limitées en portée : les travaux existants ont principalement validé le conduit auditif comme site de captation et caractérisé l'acoustique intra-auriculaire de la déglutition (Yoshimoto, Taniguchi, Kurose & Kimura, 2022; Haji & Yamaguchi, 2025; Asakura & Miwa, 2025). La détection auriculaire a montré son potentiel pour des tâches connexes telles que la détection d'épisodes alimentaires (Bi *et al.*, 2018), mais la détection robuste d'événements de déglutition, à partir d'audio intra-auriculaire dans des conditions quotidiennes réalistes, reste largement inexplorée. Les travaux exploitant le microphone intra-auriculaire se sont jusqu'ici concentrés sur d'autres événements tels que la mastication, la prise alimentaire, la parole, les événements non verbaux génériques, ou se sont limités à caractériser l'acoustique intra-auriculaire de la déglutition isolée (Papapanagiotou, Diou & Delopoulos, 2017a; Bi *et al.*, 2018; Chabot, Bouserhal, Cardinal & Voix, 2021), sans proposer de cadre de détection bout-en-bout sur flux continu.

La Figure 0.1 illustre précisément ce défi : placé dans le conduit auditif, le microphone intra-auriculaire capte simultanément un mélange de signaux bioacoustiques hétérogènes (respiration, bruits cardiovasculaires, parole) au sein duquel l'événement de déglutition doit être détecté.

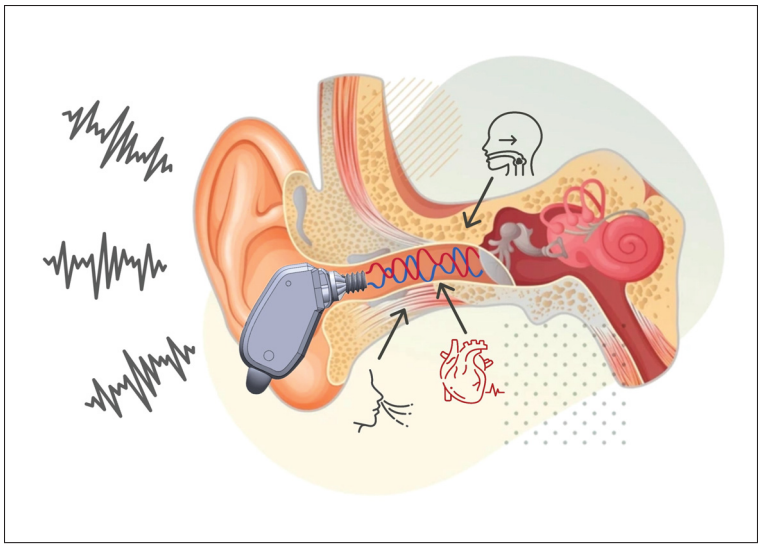


Figure 0.1 Illustration d'un IEM captant des signaux bioacoustiques dans le conduit auditif, produisant un mélange de respiration, de bruit physiologique, de parole et d'autres sons non verbaux

Cette problématique s'accompagne de plusieurs difficultés intrinsèques. Les événements de déglutition sont brefs (durée moyenne ~860 ms, rarement supérieure à 1,5 s (Khalifa, Coyle & Sejdić, 2020)), irréguliers et rares en regard des autres activités captées dans l'oreille, ce qui en fait une classe minoritaire et dispersée dans le temps (Bulmer *et al.*, 2021). À cela s'ajoute une forte variabilité inter-sujets de l'acoustique déglutitoire. Enfin, l'absence de bases de données publiques d'audio intra-auriculaire synchronisé avec des annotations de déglutition validées constitue un frein méthodologique supplémentaire au développement et à l'évaluation d'algorithmes de détection.

0.3 Objectifs

L'objectif principal de cette recherche est de développer un système de détection de la déglutition à partir d'audio intra-auriculaire bilatéral, capable d'identifier et de localiser les événements de déglutition dans des enregistrements continus. Ce système vise à exploiter les signaux des

deux oreilles et l'effet d'occlusion pour atteindre une détection robuste face aux interférences acoustiques, à la superposition temporelle des événements, au bruit et à la variabilité inter-sujets. Cet objectif principal se décline en quatre sous-objectifs :

1. **Créer et valider une base de données multimodale dédiée à la déglutition, intégrant plusieurs événements non verbaux :**
   Concevoir et mettre en œuvre une base de données multimodale synchronisée intégrant : (i) des enregistrements audio bilatéraux issus de microphones intra-auriculaires et extra-auriculaires, (ii) des signaux physiologiques comprenant l'électrocardiogramme (ECG), la respiration et l'accélérométrie, ainsi que (iii) des données d'imagerie échographique linguale utilisées comme référence pour l'annotation précise des événements de déglutition. Cette base de données doit couvrir un ensemble diversifié de tâches verbales et non verbales, incluant différents types de bolus (salive, eau, yaourt, biscuit), et être enregistrée dans des conditions acoustiques variées (calme, bruit de foule, bruit d'usine) afin de refléter des scénarios réalistes.
2. **Établir une preuve de concept pour la classification de la déglutition à partir d'audio intra-auriculaire :**
   Confirmer que l'audio intra-auriculaire porte une information suffisante pour discriminer la déglutition des autres événements physiologiques, et établir une référence de performance pour les travaux subséquents.
3. **Développer un algorithme de détection de la déglutition à partir d'audio bilatéral :**
   Proposer un système bout-en-bout capable de localiser temporellement les événements de déglutition dans un enregistrement continu, en exploitant la complémentarité des deux canaux intra-auriculaires pour améliorer la robustesse de la détection.
4. **Évaluer le système proposé :**

Quantifier les performances de détection sous différentes conditions acoustiques, justifier la contribution de chaque composante et démontrer la capacité du système à généraliser à des participants non vus lors de l'entraînement.

0.4 Structure du document

Ce mémoire est structuré selon une formule par articles et s'articule autour de deux contributions scientifiques : un article publié dans la revue *Sensors* (numéro spécial *Sensors for Physiological Monitoring and Digital Health : 2nd Edition*), intitulé « A Multimodal In-Ear Audio and Physiological Dataset for Swallowing and Non-Verbal Event Classification », et un article soumis à la revue *Applied Acoustics*, intitulé « Swallowing Detection from Bilateral In-Ear Audio ». Ces deux articles sont encadrés par une revue de la littérature et un chapitre de conclusions et perspectives.

0.4.1 Chapitre 1 : Revue de la littérature

Le premier chapitre fournit une revue approfondie de la littérature existante en lien avec cette recherche. Il commence par examiner les propriétés acoustiques du signal de déglutition, en distinguant son origine biomécanique, ses caractéristiques temporelles et spectrales, ainsi que les sources de variabilité physiologique. La revue de littérature aborde ensuite les méthodes de détection automatique de la déglutition, en couvrant les approches classiques de traitement du signal, l'apprentissage automatique traditionnel et les architectures d'apprentissage profond. Ce chapitre examine également le rôle du microphone intra-auriculaire pour la détection, en traitant de la détection d'événements par IEM, des phénomènes de propagation des sons corporels et de l'effet d'occlusion, des approches de détection actuelles, ainsi que des bases de données disponibles et de leurs limites.

0.4.2 Chapitre 2 : A Multimodal In-Ear Audio and Physiological Dataset for Swallowing and Non-Verbal Event Classification

Ce chapitre présente la première contribution du mémoire, publiée dans le journal *Sensors* (numéro spécial *Sensors for Physiological Monitoring and Digital Health : 2nd Edition*). L'article décrit la conception et la réalisation d'une base de données multimodale dédiée à la déglutition, collectée auprès de 34 adultes sains (14 femmes, 20 hommes) âgés de 20 à 29 ans. Les enregistrements comprennent de l'audio intra-auriculaire et extra-auriculaire synchronisé, des signaux physiologiques captés par une ceinture thoracique (électrocardiogramme, respiration, accélérométrie) et de l'imagerie échographique linguale servant de référence pour l'annotation des événements de déglutition. Le protocole couvre des tâches verbales et non verbales variées sous trois conditions acoustiques. En tant que preuve de concept, un réseau de neurones entièrement connecté entraîné sur des *embeddings* Yet Another Mobile Network (YAMNet) complétés par le taux de passages par zéro, atteint un score F1 moyen de 0,875 ± 0,013 pour la classification binaire de la déglutition, démontrant la faisabilité de la détection à partir d'audio intra-auriculaire.

0.4.3 Chapitre 3 : Swallowing Detection from Bilateral In-Ear Audio

Ce chapitre présente la seconde contribution du mémoire, soumise à la revue *Applied Acoustics*. L'article propose un cadre complet de détection hors-ligne d'événements de déglutition à partir d'audio bilatéral intra-auriculaire à un taux d'échantillonnage réduit de 4 kHz. Le système repose sur une représentation par normalisation d'énergie par canal (PCEN), une architecture CRNN à double oreille intégrant un encodeur partagé et une fusion adaptative par confiance des deux oreilles. Un détecteur par hystérésis, calibré à partir d'un segment de référence d'au moins 45 s sans déglutition, permet de convertir les probabilités estimées à l'échelle des fenêtres en événements discrets. Évalué sur 34 participants par validation croisée à cinq plis disjoints au niveau sujet, le système atteint un score F1 de classification au niveau fenêtre de 0,892 ± 0,006

et un score F1 de détection au niveau événement de 0,835 ± 0,012, avec un rappel dépassant systématiquement 0,870 dans toutes les conditions de bruit.

0.4.4 Chapitre 4 : Conclusions et perspectives

Le dernier chapitre résume les résultats de la recherche et en dégage les contributions principales. Il discute les implications cliniques potentielles du système proposé, notamment pour la surveillance longitudinale de la dysphagie et de la sialorrhée. Ce chapitre présente également les pistes de recherche futures envisagées, incluant la validation du système sur des populations cliniques présentant des troubles de la déglutition, l'extension de l'évaluation à des enregistrements ambulatoires non contraints, et l'investigation de reformulations causales pour un déploiement en temps réel.

CHAPITRE 1

REVUE DE LA LITTÉRATURE

Le chapitre précédent a établi le contexte clinique, le rôle de la déglutition dans plusieurs pathologies neurologiques et gériatriques, ainsi que les limites des modalités d'examen instrumental. La présente revue se concentre sur la partie acoustique du problème : comment la déglutition produit des sons exploitables, comment ces sons ont été caractérisés et détectés automatiquement, et quelle est la place du microphone intra-auriculaire dans ce champ. L'objectif est de dégager, à partir de travaux convergents, les choix méthodologiques et les verrous techniques qui structurent encore le domaine.

1.1 Caractérisation acoustique du signal de déglutition

1.1.1 Origine biomécanique et phases sonores

Le son de la déglutition n'est pas un événement acoustique unitaire mais la somme de plusieurs sources biomécaniques activées de manière séquentielle au cours du transit du bolus. La littérature s'accorde classiquement sur un découpage en trois composantes acoustiques successives (Hamlet, Patterson, Fleming & Jones, 1992; Cichero & Murdoch, 2002; Lazareck & Moussavi, 2004), dont la correspondance avec les événements physiologiques sous-jacents a été examinée par Honda *et al.* (2016) chez 33 sujets sains : 20 sujets ont permis de comparer huit sites cervicaux d'enregistrement, 10 sujets ont caractérisé l'effet du volume de bolus, et 3 sujets ont fait l'objet d'un enregistrement simultané microphone et vidéofluoroscopie permettant d'établir la correspondance entre composantes acoustiques et événements physiologiques. Un premier son transitoire bref, dit son discret initial (IDS), correspond à la période orale et à l'amorce du transit pharyngé : il coïncide avec le mouvement postérieur de la langue, l'élévation hyo-laryngée et la fermeture du vestibule laryngé, et il est attribué à la mise en mouvement des structures supra-glottiques et à l'éjection du bolus hors de la cavité orale. Lui succède le son de transit du bolus (BTS), composante plus continue et de plus longue durée, associée à la phase pharyngée. Son origine acoustique est encore débattue : elle est tantôt rapportée au passage du bolus à

travers la jonction pharyngo-œsophagienne et au frottement contre les parois pharyngées, tantôt à la perturbation des flux aériens résiduels et aux contractions séquentielles des constricteurs pharyngés (Cichero & Murdoch, 2002; Hadjileontiadis, 2005). Enfin, un second son transitoire, le son discret final (FDS), marque la période de repositionnement : il accompagne le relâchement du sphincter œsophagien supérieur, la redescente laryngée et l'expiration post-déglutition qui rétablit la perméabilité des voies aériennes. Cette structure tri-phasique n'est toutefois pas toujours discernable dans les enregistrements bruts : sa visibilité dépend fortement du capteur, de la posture et du bolus.

1.1.2 Propriétés temporelles et spectrales

Sur le plan temporel, la déglutition est un événement bref : les premiers travaux de chronométrage acoustique rapportaient des durées totales comprises entre 0,5 et 1,3 s chez l'adulte sain, avec une médiane proche de 800 ms (Cichero & Murdoch, 2002; Lazareck & Moussavi, 2004), ordre de grandeur confirmé par les corpus d'auscultation cervicale haute résolution (HRCA) de plus grande taille (Khalifa *et al.*, 2020; Dudik, Jestrović, Luan, Coyle & Sejdić, 2015). Les études récentes en microphone intra-auriculaire précisent cette fenêtre : Asakura & Miwa (2025), sur 28 adultes équipés simultanément d'un microphone cervical et d'un IEM, rapportent une durée moyenne de 720 ms en cervical et 641 ms en intra-auriculaire pour un bolus de 5 ml d'eau, illustrant que le site de capture raccourcit légèrement la durée mesurée. Chez le patient dysphagique post-AVC ou porteur d'une atteinte neuromusculaire, un allongement significatif des phases pharyngées, parfois au-delà de 2 s, a été documenté (Kurosu, Coyle, Dudik & Sejdić, 2019; Cichero & Murdoch, 2002), et Konoike *et al.* (2025) confirment un effet analogue chez les sujets âgés, en particulier les hommes face à des bolus de viscosité élevée.

Indépendamment de la durée individuelle, la fréquence de déglutition spontanée reste fortement variable : une revue systématique rapporte des taux compris entre 0,1 et 7 déglutitions par minute pendant l'éveil, sans méthode de mesure standardisée (Bulmer *et al.*, 2021). Dans le cadre d'une détection automatique, la combinaison d'un faible taux d'occurrence et d'une durée brève des déglutitions produit un déséquilibre structurel entre les classes (déglutition et les autres signaux

physiologiques audibles) dans tout corpus d'enregistrement continu, avec des conséquences directes sur l'entraînement et l'évaluation des détecteurs.

Sur le plan spectral, les sons de la déglutition concentrent la majeure partie de son énergie dans les basses fréquences. Pour les enregistrements réalisés sur le cou, un pic principal se situe autour de 300 Hz dans la bande 100–500 Hz (Asakura & Miwa, 2025; Jestrović, Dudik, Luan, Coyle & Sejdić, 2013). Pour les enregistrements réalisés au niveau du canal auditif, Asakura & Miwa (2025) montrent que ce pic basse fréquence est atténué mais que le spectre inclut systématiquement des composantes au-delà de 1 kHz, reflet de la double voie de transmission propre à l'enregistrement intra-auriculaire conduction aérienne via la trompe d'Eustache et conduction osseuse. La viscosité du bolus module fortement ces caractéristiques : Jestrović *et al.* (2013), sur 56 adultes sains et trois consistances (eau, nectar, miel), montrent que l'augmentation de viscosité abaisse le contenu en hautes fréquences et accroît la régularité du signal. Konoike *et al.* (2025) ajoutent que la puissance moyenne diminue avec la viscosité, avec un effet modulé par l'âge et le sexe, les sujets masculins présentant des amplitudes systématiquement supérieures et un effet d'âge plus marqué sur les bolus épais. Asakura & Miwa (2025) retrouvent ces effets du bolus sur la puissance, l'intensité moyenne et la fréquence de crête, mais constatent qu'ils sont plus prononcés dans le signal cervical, le signal intra-auriculaire étant moins directement couplé au mouvement du bolus.

1.1.3 Variabilité physiologique et sources de confusion

L'utilisation d'un signal acoustique comme indicateur de la déglutition suppose que cet indicateur soit suffisamment stable au sein d'un même sujet et entre sujets. Or la littérature documente une variabilité importante à plusieurs échelles. À l'échelle inter-sujets, la morphologie du conduit pharyngé, l'épaisseur des tissus cervicaux, l'âge, le sexe, la dentition et l'état de santé modifient les propriétés résonantes des voies et altèrent la signature observée (Cichero & Murdoch, 2002; Steele *et al.*, 2015). À l'échelle intra-sujet, la nature du bolus, son volume, la posture, la phase respiratoire au déclenchement et la fatigue suffisent à produire des variations notables d'amplitude, de durée et de contenu spectral (Cichero & Murdoch, 2002; Sazonov *et al.*, 2010;

Dudik *et al.*, 2015), dégradant significativement la généralisation des modèles de classification entraînés sur un sous-ensemble restreint de sujets (Khalifa *et al.*, 2020; Sejdić, Steele & Chau, 2013; Dudik *et al.*, 2015). À cette variabilité du signal-cible s'ajoute une seconde source d'erreur : la superposition temporelle d'événements tels que les battements cardiaques, bruits digestifs, raclements de gorge, qui se produisent dans la même fenêtre d'analyse et viennent masquer la déglutition.

1.2 Détection automatique de la déglutition

1.2.1 Approches par traitement du signal classique

Les premières méthodes de détection automatique reposent sur des descripteurs temps-fréquence calculés sur le signal cervical, suivis d'algorithmes à seuil ou de classifieurs linéaires. Lazareck & Moussavi (2004) équipent des sujets d'un accéléromètre piézoélectrique posé à l'échancrure suprasternale et identifient dans le signal deux sections caractéristiques : une phase d'*ouverture*, correspondant au relâchement du sphincter œsophagien supérieur, et une phase de *transit*, liée au passage du bolus. À partir de ces segments, une analyse discriminante permet de classifier correctement 13 sujets témoins sur 15 et 11 sujets dysphagiques sur 11. Walker & Bhatia (2011) étendent cette approche par seuillage adaptatif à l'énergie du signal d'un microphone laryngé. Olubanjo & Ghovanloo (2014) proposent une détection en quasi temps réel fondée sur des seuils combinés sur l'énergie, la fréquence de pic et les entropies de Shannon et d'ondelettes : sur 229 événements, leur système atteint un rappel de 79,9 % mais une précision de 67,6 %, ce qui illustre la difficulté à séparer la déglutition de la parole et de la respiration sans apprentissage.

1.2.2 Apprentissage automatique classique

L'introduction de classifieurs probabilistes a marqué une transition entre les approches fondées sur des règles heuristiques et les méthodes modernes d'apprentissage profond. Khlaifi, Badii, Istrate & Demongeot (2019) proposent un pipeline complet fondé sur des descripteurs cepstraux

et un modèle de mélange de gaussiennes (GMM) entraîné sur 27 volontaires sains : le modèle distingue déglutition, parole et bruits respiratoires avec une précision globale de 87,6 %, une sensibilité de 93,3 % pour la déglutition et de 91,9 % pour la parole. La position du microphone, fixée après comparaison entre l'échancrure jugulaire et la protubérance laryngée, contribue notablement aux performances. Wei, Pan, Zhong & Huan (2018) démontrent par ailleurs que la fusion d'une mesure d'accélération cervicale et d'un signal de photopléthysmographie (PPG) améliore la robustesse de la détection face aux artéfacts de mouvement et permet d'atteindre une précision supérieure à 90 %.

Shirazi, Buchel, Daun, Lenton & Moussavi (2012) ciblent un problème connexe, l'aspiration silencieuse, en analysant non le son de déglutition lui-même mais les caractéristiques spectrales du bruit respiratoire post-déglutition : sur 21 patients dysphagiques et 185 déglutitions validées par VFSS ou FEES, leur classifieur à partitionnement de données diffus (*Fuzzy C-Means clustering*) atteint 86,4 % d'exactitude. Makeyev, Lopez-Meyer, Schuckers, Besio & Sazonov (2012) proposent une évaluation sur un corpus de taille conséquente : 12 sujets enregistrés sur 44 heures de repas complets, totalisant 7 305 déglutitions. Le système extrait des descripteurs coefficients cepstraux sur échelle de Mel (MFCC) et les soumet à une machine à vecteurs de support (SVM) à noyau gaussien. Il atteint 87 à 90 % d'exactitude en validation intra-sujet, mais ce taux s'abaisse à 75–81 % en validation inter-sujets. Cet écart traduit une variabilité acoustique interindividuelle non négligeable et suggère que les modèles entraînés sur un individu donné peinent à se généraliser à de nouveaux sujets, ce qui pose la question de leur robustesse dans des conditions d'utilisation réelles.

1.2.3 Apprentissage profond

L'arrivée des réseaux de neurones profonds a marqué un changement d'échelle, à la fois dans la taille des corpus exploitables et dans la nature des descripteurs employés. Khalifa *et al.* (2020) ont publié le premier détecteur profond entraîné sur un corpus HRCA de grande taille : 3 144 déglutitions de 248 patients (dont 44 post-AVC) validées par VFSS simultanée, avec une durée moyenne de déglutition de 862,6 ± 277 ms. Leur réseau entièrement connecté, alimenté

par les spectrogrammes transformée de Fourier à court terme (STFT) du signal enregistré par le microphone et de l'accéléromètre, atteint au niveau fenêtre une exactitude de détection supérieure à 95 % pour la métrique globale rapportée par les auteurs. Gravellier, Le Coz, Farinas & Pinquier (2023) proposent un réseau de neurones convolutif (CNN) peu profond opérant sur des spectrogrammes à quatre canaux (un microphone, un accéléromètre tri-axe). Sur 42 adultes sains et 1 704 déglutitions annotées, le modèle atteint un score F1 de 87,2 %. Les auteurs rapportent un écart de plus de dix points entre déglutitions contrôlées (95,1 %) et déglutitions spontanées (84 %), ces dernières présentant une amplitude inférieure d'un ordre de grandeur. So, Tadj, Schwerin, Chang & Frakking (2025) proposent une approche d'apprentissage par transfert basée sur le modèle YAMNet pour la segmentation et la détection automatiques de déglutitions à partir de signaux d'auscultation cervicale chez l'enfant. Leur méthode atteint une précision globale de 91 %.

Chen & Kamavuako (2024) comparent systématiquement six méthodes de conversion 1D vers 2D (STFT, Mel-spectrogramme, champ de transition de Markov (MTF), transformée de Hilbert-Huang (HHT), transformée en ondelettes continue (CWT)) et six architectures pré-entraînées sur ImageNet (GoogLeNet, SqueezeNet, MobileNet-V2, EfficientNet-B0, ShuffleNet-V2, DeiT) pour la détection de la déglutition à l'aide d'un microphone placé sur la gorge. La combinaison STFT + MobileNet-V2 sur fenêtres de 0,5 s atteint 92,5 % de précision en détection continue, avec une latence d'anticipation d'environ 350 ms par rapport au pic du signal.

1.3 Le microphone intra-auriculaire pour la détection

1.3.1 Propagation des sons corporels et effet d'occlusion

Les enregistrements effectués par un microphone intra-auriculaire captent un spectre sonore complexe, formé par la superposition de multiples sources physiologiques : battements cardiaques, respiration (inspiration et expiration), mouvements musculaires (langue, mâchoire, yeux), mastication et déglutition. Ces sons corporels atteignent le conduit auditif par trois voies de transmission distinctes, illustrées à la figure 1.1 :

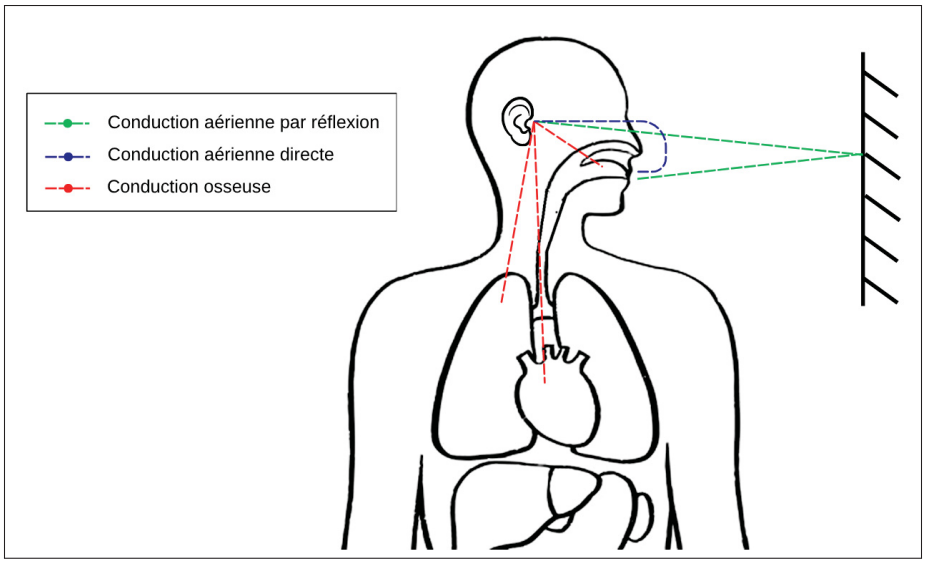


Figure 1.1 Voies de transmission des sons corporels vers le microphone intra-auriculaire : conduction aérienne directe (bleu), conduction aérienne par réflexion (vert) et conduction osseuse/tissulaire (rouge)

C'est cette troisième voie qui domine le signal capté par un microphone intra-auriculaire. Dans ce contexte de configuration occlusive, l'effet d'occlusion se manifeste de manière significative. Il désigne l'amplification des sons de basse fréquence générés à l'intérieur du corps lorsque le conduit auditif est obstrué par un dispositif occlusif : bouchon d'oreille, écouteur intra-auriculaire ou embout sur mesure. En conditions normales (oreille ouverte), les vibrations produites par conduction osseuse et tissulaire, parole, mastication, battements cardiaques, déglutition, empruntent le chemin de moindre impédance et rayonnent en partie vers l'extérieur du conduit avant d'atteindre la membrane tympanique, ce qui atténue leur intensité perçue. Lorsque le conduit est occlus, ces vibrations ne peuvent plus s'échapper : elles sont réfléchies vers le tympan et piégées dans le volume résiduel entre le dispositif et la membrane, créant une augmentation significative de la pression acoustique dans les basses fréquences tandis que les composantes hautes fréquences sont atténuées. La sensation résultante est souvent comparée à celle de parler dans un tonneau, où les réflexions sonores dans une cavité confinée créent une perception d'écho (Liebich & Jax, 2022).

Sur le plan physique, le phénomène s'explique principalement par une augmentation de l'impédance acoustique du conduit auditif (Zurbrügg, Stirnemann, Kuster & Lissek, 2014) conjuguée à une transmission par conduction osseuse renforcée. Son amplitude dépend de plusieurs facteurs, parmi lesquels la géométrie du conduit, la fréquence du son interne et, de manière prépondérante, la profondeur d'insertion du dispositif occlusif (Dean & Martin, 2000; Branda, 2012; Nielsen & Darkner, 2011). Une insertion peu profonde, si elle assure un sceau acoustique, piège le son dans un volume résiduel important et produit un effet d'occlusion marqué. En revanche, une insertion très profonde, qui scelle le conduit à proximité de sa portion osseuse, réduit considérablement l'effet. L'absence de sceau, quelle que soit la profondeur, laisse le conduit acoustiquement ouvert et supprime l'amplification. La modélisation par éléments finis de (Brummund, Sgard, Petit & Laville, 2014) pour l'oreille externe humaine confirme que l'amplification est principalement médiée par la conduction osseuse et qu'elle se concentre sous environ 1 à 2 kHz. Les mesures expérimentales rapportent des amplifications typiques de 20 à 30 dB, pouvant atteindre jusqu'à 40 dB sous 1 kHz en conditions de sceau optimal (Bernier & Voix, 2013).

Par ailleurs, si l'effet d'occlusion est un atout pour la captation de biosignaux, il est généralement perçu comme inconfortable lors d'un port prolongé en raison de l'amplification de la propre voix de l'utilisateur. Plusieurs stratégies de mitigation ont été proposées. L'ajustement de la profondeur d'insertion offre un compromis : une insertion plus profonde réduit l'effet mais peut occasionner un inconfort en raison de la pression exercée sur les portions déformables du conduit lors de la mastication ou de la parole (Benacchio *et al.*, 2018). Une insertion plus superficielle diminue l'isolation vis-à-vis du bruit externe, affaiblissant la captation des sons corporels. Les solutions les plus avancées reposent sur des principes de contrôle actif du bruit, et plus spécifiquement (Mejía, Dillon & Fisher, 2008; Liebich & Jax, 2022), qui utilisent un microphone interne pour mesurer l'énergie acoustique indésirable et générer un signal de compensation en temps réel. Cette approche présente le double avantage de réduire l'inconfort perceptif lié à l'amplification de la voix propre tout en préservant la capacité de captation des sons corporels d'intérêt pour le monitoring physiologique.

1.3.2 Détection d'événements par un microphone intra-auriculaire

Le microphone intra-auriculaire enregistre les sons présents dans le canal auditif occlus d'un sujet. Ce site capte les sons verbaux (parole) ou non verbaux (mastication, déglutition, claquements de langue, claquements de dents, raclements de gorge, soupirs, respiration profonde). Plusieurs travaux ont étudié ces événements pour des finalités variées (amélioration de la parole, reconnaissance d'activité, suivi de la nutrition), formant un corpus méthodologique cohérent dont la déglutition n'est qu'une cible parmi d'autres. Röddiger *et al.* (2022), dans une revue systématique de la littérature *hearables*, recensent plus de soixante phénomènes physiologiques et comportementaux détectables à l'oreille.

Goverdovsky *et al.* (2017) décrivent un dispositif, *hearable*, multimodal mesurant l'électroencéphalogramme (EEG), le rythme cardiaque et la respiration. Les auteurs mentionnent explicitement les sons de mastication et de déglutition comme artefacts récurrents, documentant indirectement la sensibilité du capteur à ces événements.

Chabot *et al.* (2021) proposent un pipeline de détection et de classification d'événements non verbaux à partir d'un microphone intra-auriculaire. Leur architecture combine des descripteurs MFCC, des coefficients de modulation d'amplitude inspirés de l'audition et une représentation PCEN, agrégés selon un schéma de sac de mots audio (BoAW), avant classification par une SVM à noyau polynomial. Sur dix classes (claquement de dents et de langue, clignement forcé des yeux, fermeture des yeux, fermeture forcée des yeux, grincement de dents, raclement de gorge, bruit de salive, toux et parole) ce pipeline atteint 81,5 % de sensibilité et 83,0 % de précision en environnement silencieux. Les performances chutent à 55–70 % en présence d'un bruit de babillage ou d'usine à 70 dBA, ce qui illustre la robustesse modérée des descripteurs construits à la main lorsque le signal intra-auriculaire est superposé à du bruit. Papapanagiotou *et al.* (2017a) démontrent que le même type de capteur permet une détection efficace de la mastication par CNN appliqué directement au signal sous-échantillonné à 2 kHz, avec un score F1 de 0,908.

Cet ensemble de travaux montre que le microphone intra-auriculaire constitue un instrument de capture polyvalent pour les événements oro-faciaux, mais qu'il est difficile, en pratique, de séparer la déglutition des autres signaux générés par le sujet, qu'ils soient verbaux ou pas.

1.3.3 Le microphone intra-auriculaire pour la détection de la déglutition

1.3.3.1 Approches actuelles de détection

Les travaux centrés explicitement sur la déglutition à partir d'un microphone intra-auriculaire restent peu nombreux et de portée limitée. L'idée remonte au travail pionnier de Firmin, Reilly & Fourcin (1997), qui combinaient une sonde auriculaire avec un électrolaryngographe et un microphone aérien afin d'observer les sons intrinsèques de la déglutition. Päßler & Fischer (2011) introduisent une prothèse auditive logeant un microphone intra-canal et un microphone de référence captant le bruit ambiant. Päßler, Wolff & Fischer (2012) l'appliquent ensuite à 51 sujets sains consommant sept aliments solides et une boisson, à l'aide d'un modèle de Markov caché (HMM) avec algorithme de Viterbi pour segmenter les épisodes de mastication et de déglutition. Le système atteint 83 % de précision en détection d'ingestion et 79 % en classification. La déglutition y est traitée comme un événement parmi d'autres dans une chaîne alimentaire et non comme une cible clinique en soi.

Chabot *et al.* (2021), déjà présentés à la section 1.3.2, atteignent 81,5 % de sensibilité et 83,0 % de précision en environnement silencieux sur dix classes d'événements, avec une dégradation notable en validation temps réel (69,9 % / 78,9 %).

Aucune étude publiée à notre connaissance ne propose, à ce jour, une chaîne complète de détection de la déglutition à partir d'un microphone intra-auriculaire validée contre un étalon.

1.3.3.2 Bases de données disponibles et limites

Les bases de données publiques exploitables pour la détection de la déglutition par microphone intra-auriculaire sont inexistantes à ce jour. Les corpus IEM disponibles dans la littérature tels

qu'iBad (Chabot *et al.*, 2021), SPLENDID (Papapanagiotou *et al.*, 2017a) ou les travaux de Goverdovsky *et al.* (2017) incluent bien la déglutition parmi les événements enregistrés, mais sans annotation accompagnée de mesures objectives de référence. La déglutition y constitue une classe parmi d'autres événements non vocaux, annotée manuellement à partir du signal audio ou vidéo, sans validation indépendante.

Le tableau 1.1 récapitule les corpus IEM ou auriculaires audio dans lesquels une déglutition (au sens large) est identifiable. Les corpus cervicaux à validation cinématique (VFSS / FEES) qui dominent par ailleurs la littérature de la détection de déglutition ne sont pas listés ici car ils ne relèvent pas du même cadre de mesure. Les corpus IMU sans audio (FIC (Kyritsis, Tatli, Diou & Delopoulos, 2017), FreeFIC (Kyritsis, Diou & Delopoulos, 2020)), les microphones péri-auriculaires non strictement intra-canal (Auracle (Bi *et al.*, 2018), EarBit (Bedri *et al.*, 2017)) ont également été écartés.

Tableau 1.1 Corpus basés sur des capteurs intra-auriculaires contenant des événements de déglutition, sans validation par un étalon clinique objectif

Corpus	*N*	*Annotation déglutition*	*Disponibilité*
iBad (Chabot *et al.*, 2021)	25	Classe *saliva noise* annotée	Privé
SPLENDID (Papapanagiotou *et al.*, 2017a)	14	Mastication annotée, déglutitions présentes mais non isolées	Privé
EarSAVAS (Zhang *et al.*, 2024)	42	Déglutition non annotée	Public
Vibravox (Hauret *et al.*, 2024)	188	Déglutition non annotée	Public
ACE (Merck *et al.*, 2016)	6	Mastication et déglutition annotées (vérité terrain vidéo)	Privé
(Päßler *et al.*, 2012)	51	Annotation manuelle des mâchements et des déglutitions	Privé
(Papapanagiotou *et al.*, 2021)	9	Déglutition non annotée	Privé
(Goverdovsky *et al.*, 2017)	NA	Déglutition mentionnée comme artefact uniquement (sans annotation)	Privé

CHAPITRE 2

A MULTIMODAL IN-EAR AUDIO AND PHYSIOLOGICAL DATASET FOR SWALLOWING AND NON-VERBAL EVENT CLASSIFICATION

Elyes Ben Cheikh[1] , Yassine Mrabet[1,2] , Catherine Laporte[1] , Rachel E. Bouserhal[1,2]

[1] Department of Electrical Engineering, École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Centre for Interdisciplinary Research in Music Media and Technology,
527 Rue Sherbrooke O, Montréal, Québec, Canada H3A 1E3

Cet article a été publié en 2026 dans la revue Sensors (vol. 26, n° 7, art. 2019), dans le numéro spécial *«Sensors for Physiological Monitoring and Digital Health : 2nd Edition»*

2.1 Abstract

Swallowing is a critical marker of neurological and emotional health. The ability to monitor it continuously and non-invasively, especially through smart ear-worn devices, holds significant promise for clinical applications. Despite this potential, no public audio datasets currently support reliable swallowing sound detection. Existing datasets focus primarily on speech and breathing, offering limited coverage and lacking detailed annotations for swallowing events. To address this gap, we introduce an in-ear audio dataset specifically designed to capture a wide range of verbal and non-verbal sounds. It includes comprehensive labeling focused on swallowing. The dataset was collected from 34 healthy adults (14 females and 20 males) between the ages of 20 and 29. Each participant performed a series of predefined tasks involving both non-verbal and verbal events. Non-verbal tasks included swallowing, clicking, forceful blinking, touching the scalp, and physical movements such as squatting or walking in place. Verbal tasks consisted of speaking (e.g., describing an image). Recordings were conducted in both quiet and noisy environments to better reflect real-world conditions. Data were captured using a combination of in-/outer-ear microphones, a chest belt to record ECG, respiration and acceleration signals, and an ultrasound probe to track tongue movement, which served as a reference for swallowing annotation. All signals were precisely synchronized. To ensure high data quality, the recordings were reviewed using both algorithmic analysis and manual inspection. Swallowing events were

identified based on ultrasound signals and validated by an expert to guarantee accurate labeling. As a proof of concept that in-ear audio supports swallow classification, we fine-tune a fully connected neural network on YAMNet embeddings plus ZCR features. Across the completed folds, the model reaches an F1 score of 0.875 ± 0.013.

2.2 Introduction

Swallowing is a vital physiological function that ensures the safe transport of food from the oral cavity to the esophagus while protecting the airway (Matsuo & Palmer, 2008). A disruption of this mechanism can cause dysphagia, which affects many adults in relation to various neurological disorders (Kang, 2024). In Parkinson's disease, sialorrhea, or excessive salivation, is mainly due to a lower frequency of swallowing rather than increased saliva production (Chou *et al.*, 2007). This impaired swallowing in Parkinson's disease patients leads to social embarrassment and promotes saliva inhalation into the lungs, posing a risk of respiratory complications (Chou *et al.*, 2007). Similarly, patients with motor neuron disease exhibit a reduction in swallowing ability compared to healthy subjects (Hughes & Wiles, 1996). Beyond neurological disorders, the frequency of spontaneous swallowing is also modulated by emotional state or stress. Studies in healthy volunteers have shown that stress or anxiety significantly increases the rate of swallowing compared to a relaxed state (Cuevas *et al.*, 1995). Moreover, sensory or cognitive factors such as hunger, attention, or exposure to food-related stimuli like pleasant smells or visual presentation of food can also influence swallowing activity (Loret, 2015; Ashiga *et al.*, 2019). Thus, swallowing is not only a marker of neurological or physiological health but also reflects behavioral, sensory, and emotional states. Therefore, monitoring swallowing has a dual clinical value : it allows for detecting and tracking the progression of swallowing disorders, while also providing insights into the patient's autonomic and behavioral states (stress, pain or hunger) through the frequency and context of spontaneous swallowing.

Physiological factors related to the swallowing act itself also influence the acoustic characteristics of swallowing sounds. In particular, variations in bolus properties and subject age have been shown to significantly affect the spectral and temporal features recorded from both ear and

cervical sensors (Asakura & Miwa, 2025). Such variability introduces additional challenges for automatic swallowing detection systems, as swallowing signals may differ substantially both between subjects and across repeated swallows within the same individual.

Given the clinical relevance of swallowing, researchers are exploring technological approaches to continuously monitor this function in daily life.

Multimodal measurements are increasingly viewed as essential for reliable physiological monitoring. Rather than relying on a single channel, recent systems deliberately combine complementary modalities that capture different aspects of swallowing physiology. Shin et al., for example, integrated two surface electromyography (EMG) electrodes and a microphone on a kirigami-structured neck wearable, achieving about 89.5 % accuracy in swallowing and silent aspiration detection with clinically competitive performance (Shin *et al.*, 2024). Similarly, Song et al. introduced a soft, skin-attachable throat vibration sensor and demonstrated deep-learning-based classification of multiple throat events (e.g., cough, speech, swallowing), reporting an accuracy of approximately 96 % (Song, Yun, Giovanoli, Easthope & Chung, 2025). Beyond performance gains, these studies highlight a key motivation for multimodal sensing : combining heterogeneous signals reduces ambiguity, improves robustness to noise and motion artifacts, and mitigates inter-subject variability by allowing one modality to compensate when another becomes unreliable. However, despite the clear advantages of multimodal fusion, progress is constrained by the limited availability of high-quality, synchronized multimodal datasets with reliable reference annotations.

Recent work on automatic swallow detection has primarily relied on neck-based sensing, including cervical auscultation microphones, accelerometers, and surface EMG. Using these modalities, Gravellier et al. reported multi-event pharyngolaryngeal activity detection in ecological conditions with a depthwise CNN–LSTM model (F1 = 0.88 for swallowing on 42 participants) (Gravellier, Coz, Farinas & Pinquier, 2024). Kimura *et al.* (2024) achieved F1 $\approx$ 0.92 and 95.2 % accuracy on videofluoroscopic swallowing study VFSS synchronized recordings using MFCC features and ensemble classifiers. More recently, So et al. extended cervical auscultation approaches to

pediatric and neonatal cohorts using transfer learning, reporting 91 % accuracy (So *et al.*, 2025). While these results demonstrate the effectiveness of neck-based pipelines, the required sensor placement may be obtrusive and therefore less suited to comfortable, long-term monitoring in daily life.

The recent emergence of smart wearable auditory devices, commonly known as hearables, offers new opportunities for the non-invasive monitoring of vital signs and physiological signals. Recent developments in hearable hardware platforms further highlight the ear as a promising location for physiological sensing. For example, the open-source OpenEarable 2.0 platform integrates multiple sensors within an ear-worn device, enabling multimodal physiological monitoring (Röddiger *et al.*, 2025). Hearables can operate with or without an acoustic seal. Open-fit earbuds admit external sound and produce little to no occlusion effect, whereas IEM devices form a tight seal in the ear canal when properly placed. Compared to conventional throat or neck placements, a sealed ear canal benefits from the occlusion effect. This seal plays a crucial role by reducing external noise. This effect results in the amplification of internally generated sounds such as breathing, swallowing, or muscle movements through bone and tissue conduction. As a result, vibrations are amplified within the sealed canal, allowing the IEM to capture subtle non-verbal physiological signals (Chabot *et al.*, 2021).

Several studies have demonstrated the feasibility of measuring respiratory rate and heart rate using a microphone placed in the ear, even in noisy environments (Butkow, Dang, Ferlini, Ma & Mascolo, 2021; Martin & Voix, 2018). For example, audio recordings in the ear canal allow the extraction of breathing and heartbeat sounds with acceptable accuracy after denoising, even at ambient noise levels up to 110 dB (Martin & Voix, 2018). Similarly, in-ear recordings of respiratory sounds have been used to classify respiratory phases (inspiration vs. expiration) and to distinguish nasal from oral breathing (Mehrban, Voix & Bouserhal, 2024). These findings confirm the potential of IEMs to capture physiological signals (Hauret, Joubaud, Zimpfer & Éric Bavu, 2023).

More recently, this approach has been extended to swallowing : sensors embedded in earpieces have shown the ability to detect certain movements associated with swallowing. In particular, Yoshimoto *et al.* (2022) showed that an earphone type optical sensor, consisting of an in-ear IR LED and phototransistor that convert changes in reflected light from the ear canal motion into electrical signals, can track soft-palate motion during swallowing when compared with simultaneous videofluoroscopy (the clinical gold standard). In another study, intra-aural acoustic recordings synchronized with videofluoroscopy demonstrated that ear-canal microphones can capture swallowing-related sound patterns that are temporally aligned with physiological swallowing events (Haji & Yamaguchi, 2025). Automated segmentation of swallowing sounds has also been validated against videofluoroscopic swallowing studies to estimate clinically relevant parameters such as pharyngeal clearance time (Jayatilake, Teramoto, Ueno, Matsuo & Suzuki, 2026). In-ear sensing has also been explored for detecting mastication. Papapanagiotou *et al.* (2017b) developed a chewing detection system combining audio, photoplethysmography (PPG), and accelerometry, demonstrating that multimodal fusion improves the automatic recognition of eating activity.

These advancements highlight how hearables, beyond their audio capabilities, are emerging as health monitoring platforms capable of continuously and discreetly tracking respiration, pulse, chewing, or swallowing.

Despite these promising developments, there are very few publicly available datasets suitable for training and validating automatic swallowing detection algorithms. As summarized in Table 2.1, existing ear-worn and IEM datasets can be broadly divided into speech-oriented, respiration-focused, or activity-based corpora. Although valuable for their respective objectives, these resources present important limitations when considered for precise swallowing analysis. Speech-oriented datasets were primarily developed for voice acquisition rather than physiological event detection.

Tableau 2.1 Comparison of Representative In-Ear and Ear-Worn Datasets

Dataset	Subjects	Modalities	Noise Diversity
SpEAR	25	IEM + OEM + Ref mic	Quiet + noisy
VibraVox	188	IEM + Bone conduction mic + Laryngophone	Quiet + noisy
iBad	25	IEM + OEM + ECG + Resp.	Quiet + noisy
RespEar	18	IEM + Resp. ref.	Sedentary + active
OESense	31	IEM	Head movement
Proposed Dataset	34	IEM + OEM + ECG + Resp. + IMU	Quiet + noisy

IEM : in-ear microphone, OEM : outer-ear microphone, ECG : electrocardiogram, Resp. : respiration, IMU : inertial measurement unit.

Existing datasets can be grouped into speech, respiration, and activity-oriented corpora, each addressing different objectives but remaining limited for swallowing validation. SpEAR (Bouserhal, Bernier & Voix, 2019) provides synchronized in-ear and outer-ear microphone recordings from 25 participants, yet focuses exclusively on speech without physiological or swallowing annotations. VibraVox (Hauret *et al.*, 2024) extends this approach to a large-scale multimodal corpus (188 participants) including IEMs, bone conduction sensors, and a laryngophone, but does not provide event-level or imaging-supported swallowing labels. On the respiration side, iBad (Chabot *et al.*, 2021) combines in-ear audio with ECG and respiratory belt signals for non-verbal in-ear event detection, although swallowing events are neither specifically annotated nor imaging-validated. Similarly, RespEar (Liu *et al.*, 2024) links IEM recordings to respiratory references but concentrates on respiration rate estimation rather than event detection. Finally, activity-oriented datasets such as OESense (Ma, Ferlini & Mascolo, 2021) demonstrate the feasibility of using the in-ear occlusion effect for human-motion sensing — including step counting and hand-to-face gesture detection but do not provide annotations for swallowing with

references. As a result, none of these publicly available datasets integrates synchronized in-ear audio, multimodal physiological measurements, and imaging-validated swallow annotations with precise temporal boundaries for benchmarking automatic swallowing detection.

In this context, we introduce a multimodal dataset covering varied verbal and non-verbal events, with comprehensive annotations and a focus on swallowing. Synchronized in-/outer-ear audio, respiration, ECG, acceleration and lingual ultrasound were recorded from healthy adult participants performing spontaneous (natural rate) or prompt swallows under three opposing acoustic conditions : one quiet environment and two noisy ones. Finally, we present a proof of concept demonstrating the relevance of this multimodal dataset for swallowing analysis through a binary swallow classification task.

2.3 Methods

In this study, data collection and analysis were conducted following ethical approval from the CER, the IRB of the École de technologie supérieure (H20221006). All procedures were carried out according to the relevant guidelines and regulations.

2.3.1 Data Collection

Participants were recruited through email invitations and word-of-mouth referrals. A total of 34 individuals (14 females and 20 males), aged between 20 and 29, were enrolled in the study. Before data collection, all participants underwent an otoscopic examination to assess the condition of the outer ear and the eardrum. This evaluation ensured the absence of foreign objects, excessive cerumen, or signs of infection or inflammation. Only participants deemed healthy and free of ear-related conditions were included.

2.3.1.1 Sensors and Instrumentation

Acoustic data were collected inside a double-wall audiometric booth (Eckel Noise Control Technologies, Morrisburg, ON, Canada) using a Roland OCTA-CAPTURE USB audio interface

(Roland Corp., Hamamatsu, Japan) operated via MATLAB 2022 (MathWorks, Natick, MA, USA). Recordings were made on four channels at a sampling rate of 48 kHz with 16-bit resolution, capturing signals simultaneously from the left and right ears using custom earpieces developed by the ÉTS-EERS Industrial Research Chair in In-Ear Technologies. Each earpiece was equipped with two microphones : one positioned inside the ear canal and the other outside the ear. Both microphones were Knowles FG45-32491-000 electret condenser capsules (Knowles Electronics LLC, Itasca, IL, USA). The nominal sensitivity is −58 dB re 1 V/Pa at 1 kHz, which corresponds to approximately 1.26 mV/Pa. The intra-aural microphone configuration is based on previously validated in-ear sensing technology, shown to reliably capture physiological acoustic signals (Bouserhal *et al.*, 2019). Background noise was delivered through four loudspeakers placed within the booth.

Physiological signals were captured using the BioHarness™3.0 wearable chest belt (Zephyr Technology Corporation, Annapolis, MD, USA), a clinically and experimentally validated system for ambulatory monitoring of multiple biosignals, with demonstrated accuracy and reliability for respiration and cardiac measurements across rest and exercise conditions (Panni, Cosoli, Antognoli & Scalise, 2024; Nazari & MacDermid, 2020). The device, equipped with a chest strap containing dry electrodes and a sensor module, recorded a single-lead ECG at 250 Hz, respiration rate through a pressure sensor at 25 Hz, and triaxial accelerometry at 100 Hz. Prior to each session, the device was configured in logging mode using the Zephyr Configuration Tool.

Based on evidence showing that ultrasound of the submental region allows reliable visualization of swallowing-related movements (Maeda *et al.*, 2023), ultrasound imaging of the tongue through the submental region was performed using a pocket-sized MicrUs USB-powered system (Telemed, Vilnius, Lithuania) equipped with an MC4-2R20S-3 probe, (4 MHz center frequency, 90 mm depth, 66 dB dynamic range). The system was operated with EchoWave 2 software (v4.4) on a laptop with an Intel Core i7 processor, 16 GB of RAM and a 64-bit operating system. Ultrasound frames were captured at 30 fps, while OBS Studio (v30.2) simultaneously recorded the on-screen ultrasound video along with an audio monitoring channel sampled at 44.1 kHz. Pocket-sized ultrasound systems have previously demonstrated validity and reliability

for swallowing assessment, supporting their suitability for non-invasive and repeatable evaluation of swallowing biomechanics (Winiker, Hammond, Thomas, Dimmock & Huckabee, 2022).

An example of the setup, including a participant equipped with the sensors, is shown in Figure 2.1.

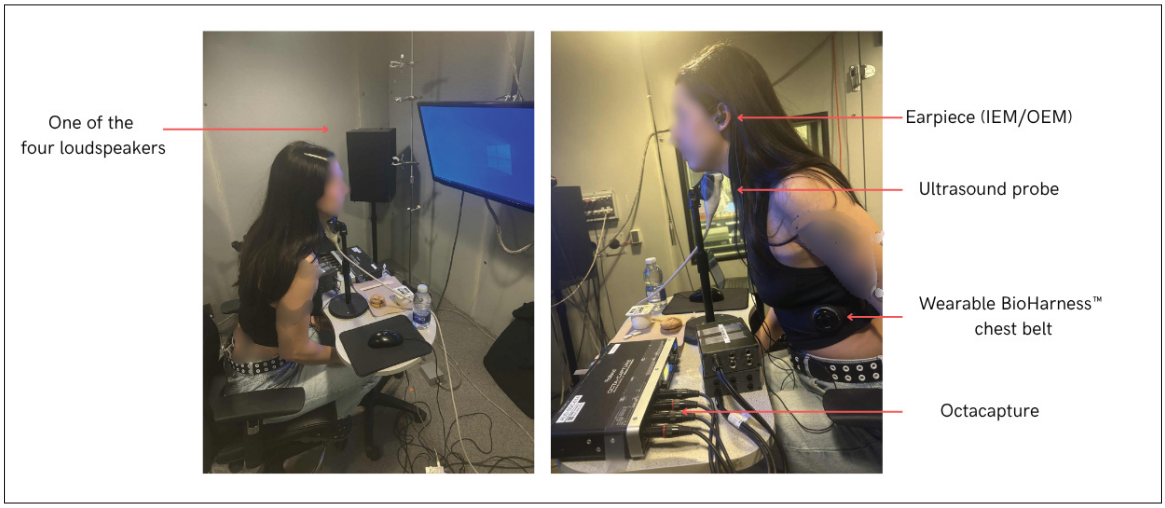


Figure 2.1 An example showing the setup and the placement of the different sensors during the recording

2.3.1.2 Procedure

The protocol was repeated three times under different auditory conditions : quiet, babble and factory noise with identical procedures per session, as illustrated in Figure 2.2. Longitudinal recordings were not included in this study. Prior physiological studies report no significant test–retest differences in swallowing timing measures across sessions in healthy individuals (Lof & Robbins, 1990). Each session began with a fitting test to ensure proper placement : a good acoustical seal and, therefore, optimal signal quality. This fitting test was repeated as many times as necessary until the quality criteria defined in Section 2.3.1.3 were met, ensuring reliable data acquisition throughout the session. After the initial fitting validation, the participants proceeded to Sequence 1, as shown in Figure 2.2, which included various tasks such as tongue clicking, blinking, scalp touching, drinking water, clicking teeth, eating biscuits or clearing the throat. Upon completion of this sequence, participants were given a pause period before undergoing a

new fitting test, again ensuring correct positioning and signal integrity. This was followed by Sequence 2, which involved actions such as teeth grinding, drinking water, eating yogurt, rubbing the eyes, artificially yawning, and chewing gum. After a second pause, Sequence 3 followed, where participants performed head movements, squats, and walking in place, interspersed with fit tests between activities to confirm ongoing fitting adequacy, particularly after physical activities that could affect device positioning. In this sequence, no ultrasound was used. Sequence 3 was included to capture motion-related artifacts and physiological variability in heart rate and respiration. Alternating between sequences 1 and 2, participants also carried out specific inter-sequence events including dry swallowing, a voluntary 10 s apnea, and a return to regular breathing. At the end of each session, a final fitting test was performed prior to concluding the recording. This structured approach ensured that the device fitting was carefully controlled and maintained at every step of the experiment. Before each activity, participants pressed the Start button in the MATLAB interface to mark the onset. Immediately after completing the activity, they pressed an End button to mark the offset. Both markers were generated and time-stamped by the same MATLAB process that controlled signal acquisition, sharing the acquisition clock. This provided an initial coarse, time-synchronized annotation of the recordings, including the activity class.

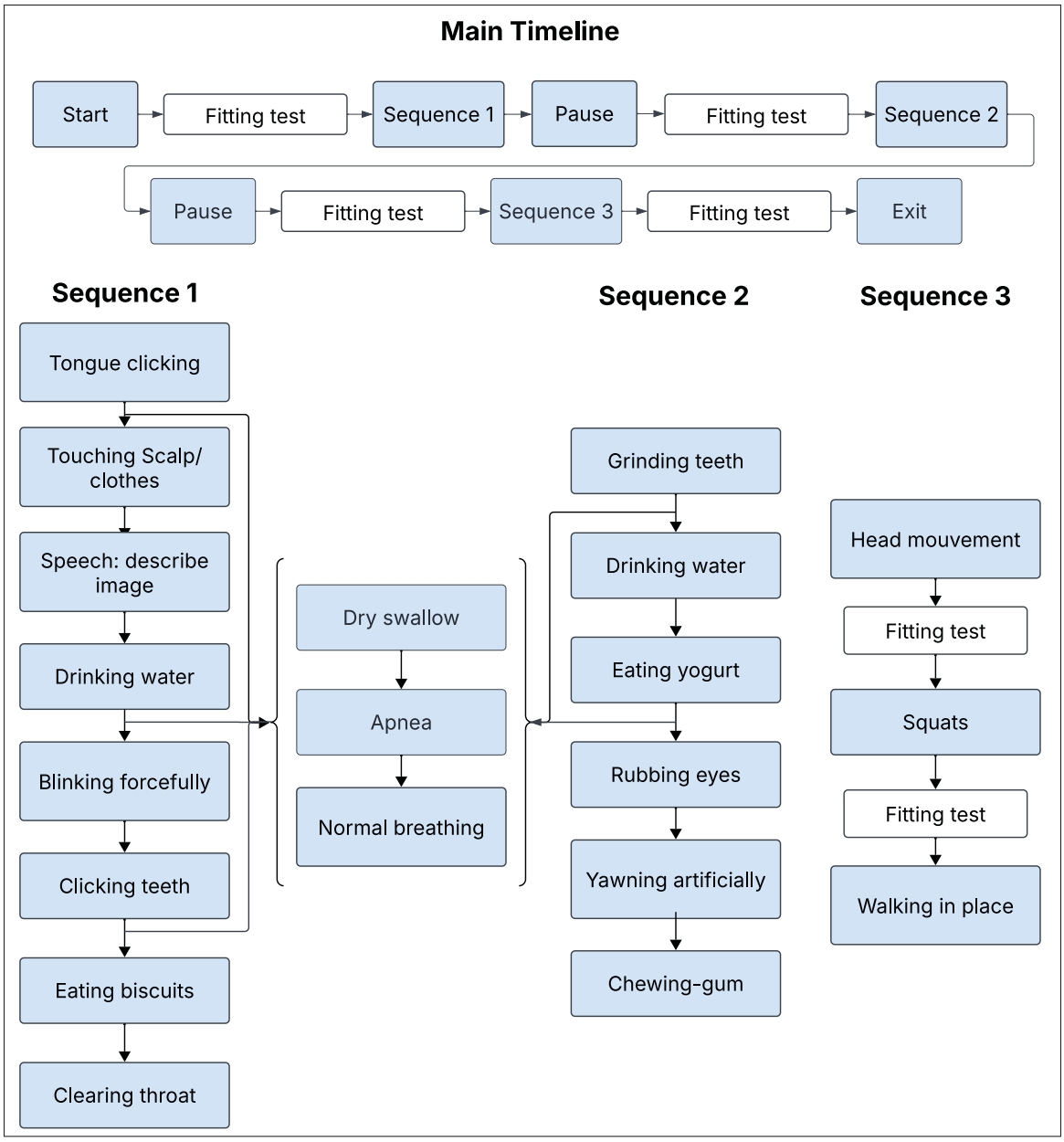


Figure 2.2 Sequence diagram of the experimental protocol

2.3.1.3 Acoustic Seal Validation

To verify an adequate acoustic seal, pink noise at 85 dBA was played through four loudspeakers placed in the audiometric room. The custom earpiece fit was considered acceptable if the attenuation between the OEM and the IEM signals was at least 8 dB at 250 Hz (Bernier & Voix,

2013). The foam tips of the earpieces were adjusted as needed, and different sizes were used depending on each participant's ear canal dimensions to ensure proper insertion and sealing.

2.3.1.4 Multimodal Synchronization

Data were acquired using three different sensing systems : custom earpieces for in-ear audio signals, a chest belt for physiological measurements including ECG, respiration, and acceleration, and an ultrasound system for imaging tongue movement. As these modalities operated independently, a two-step synchronization protocol was required to align the recordings.

First Synchronization Step : Earpiece and Ultrasound System

The initial synchronization step aligned the earpiece and ultrasound data streams using an audio cue. Specifically, at the beginning of the recording protocol, a 500 ms beep was played, which was simultaneously recorded by the external microphone (capturing the ultrasound display) and by the earpiece's microphone channels. This beep, recorded by both the external microphone and the OEM, was subsequently identified in the recordings through autocorrelation analysis between the external microphone signal and the ultrasound video audio track, as well as between the OEM signal and the earpiece audio recording. We then shifted the delayed signal by the estimated offset to align the beep onsets. Given the audio sampling rates of 48 kHz and 44.1 kHz, the temporal resolution of this alignment is limited to one sample period, corresponding to a synchronization uncertainty of approximately ±0.02 ms. Figure 2.3 illustrates the detection of the synchronization beep in both data streams.

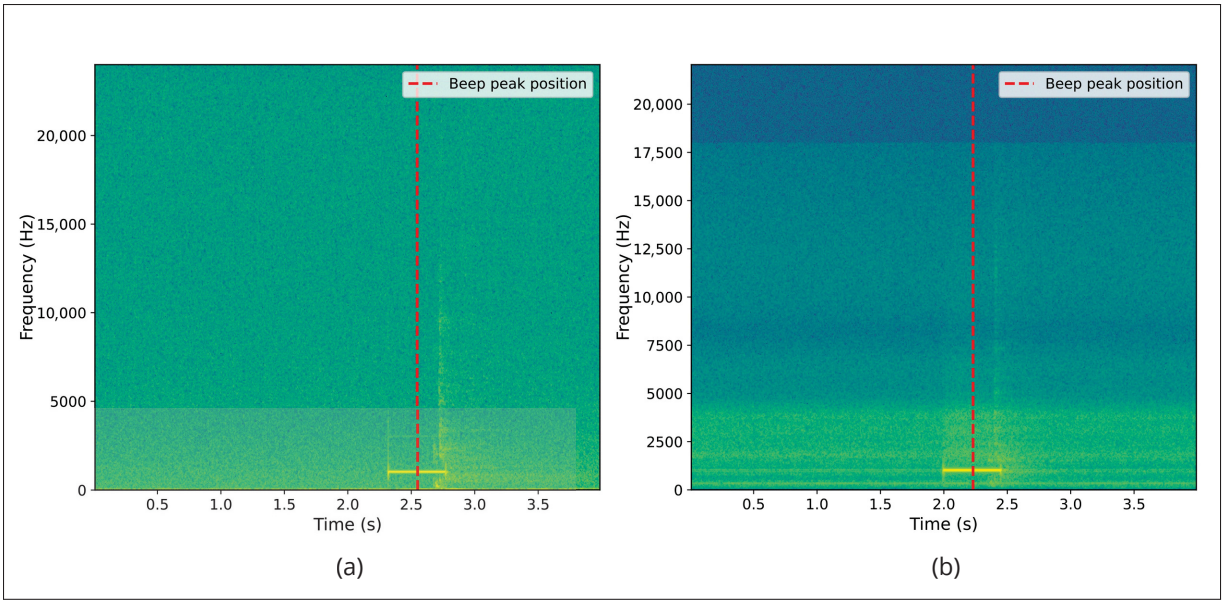


Figure 2.3 First synchronization marker alignment. (**a**) Spectrogram of the beep recorded from right ear IEM microphone. (**b**) Spectrogram of the beep recorded from the external microphone, synchronized with ultrasound via OBS

Second Synchronization Step : Earpiece and Chest Belt via Respiration

The second synchronization step aligned the audio from the earpiece and the data from the chest belt using the respiration signal as shown in Figure 2.4. Participants were instructed to hold their breath for 10 s at the start of the protocol, followed by an immediate exhalation after the synchronization beep. This maneuver allowed the identification of the exhalation event in both the OEM signals and the respiration signal. The respiration signal exhibited a plateau corresponding to breath-holding. This was followed by a marked change indicating the onset of exhalation, characterized by a progressive decrease in lung volume.

Simultaneously, the sound of exhalation was visible in the earpiece audio spectrogram. Since the earpiece and ultrasound data streams had already been synchronized in the first step, aligning the earpiece and chest belt signals via respiration allowed for the synchronization of all three data streams. Since the respiration signal was sampled at 25 Hz, the synchronization uncertainty

is limited to one sample period, corresponding to ±40 ms. This represents the maximum cross-modal synchronization error.

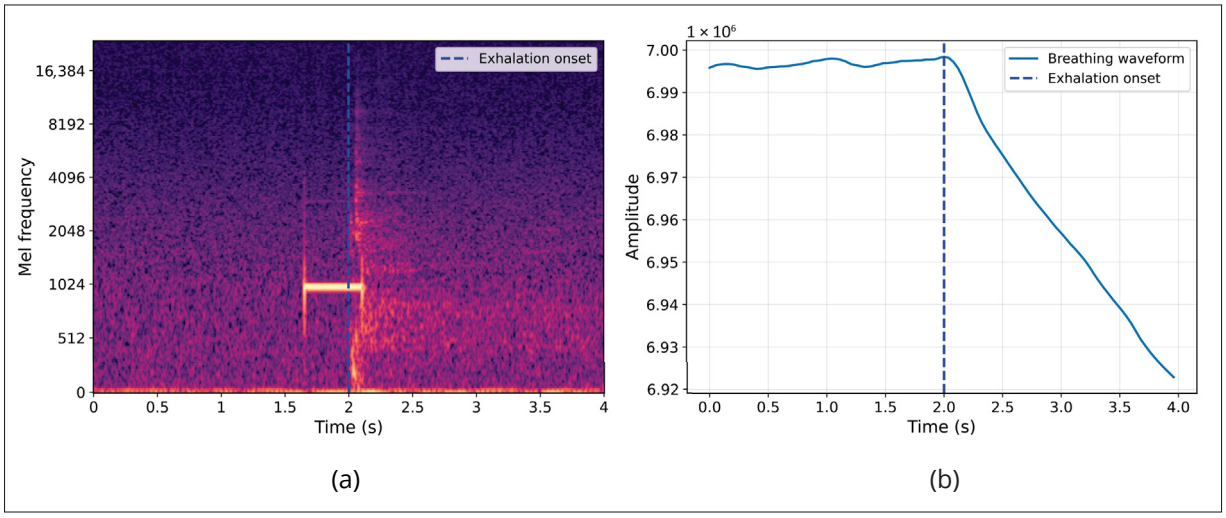


Figure 2.4 Second synchronization marker alignment. (**a**) Spectrogram of the exhalation sound recorded by right ear IEM microphone. (**b**) Breathing waveform recorded by BioHarness 3.0 device

2.3.1.5 Labeling

Labeling targeted four types of swallow : saliva, water, yogurt, and biscuits. Within the start and end markers defined in Section 2.3.1.2, we reviewed each 30 s activity window using synchronized ultrasound video and in-ear audio. Ultrasound recordings were examined frame by frame to identify candidate events that show the characteristic propulsion/retraction pattern of the tongue associated with swallowing, which were then confirmed when a temporally aligned swallow sound was observed in the IEM recording. Ultrasound was recorded at 30 fps (33 ms per frame). To avoid onset and offset ambiguity at the frame level, each swallow label was padded by one frame (±33 ms), so the entire event is reliably included in the annotated segment. For each confirmed event, onset and offset times were annotated. All annotations were performed by a single rater. A subset of labels was independently checked by an expert.

2.3.2 Swallowing Binary Classification

The study initially enrolled 34 healthy young adults. Prior to analysis, several participants were excluded because of missing or unreliable swallow annotations due to incomplete ultrasound labeling. The final dataset used for classification included 26 participants and 8202 annotated acoustic segments.

For binary swallowing classification, any segment containing a swallow regardless of bolus type (saliva, water, yogurt, or biscuit during mastication) was assigned to the positive class (label = 1). All remaining segments were assigned to the negative class (label = 0), which contains quiet respiration or silence, saliva mix sounds, and physiological noise.

2.3.2.1 Preprocessing

All analyses used in-ear recordings captured at 48 kHz. Recordings were normalized and split into fixed 4 s windows via zero-padding or truncation as needed to match the maximum swallow duration observed and to have a consistent input size. Each segment was resampled to 8 kHz to match the effective bandwidth of in-ear recordings imposed by the occlusion effect (Brummund *et al.*, 2014) (Nyquist preserves content up to 4 kHz).

2.3.2.2 Classification

The choice of YAMNet was motivated by its strong generalization ability for environmental and non-speech acoustic events. This pre-trained model provides robust frame-level embeddings that capture both spectral and temporal features relevant to short impulsive body sounds such as swallows. Moreover, previous work by So *et al.* (2025) demonstrated that YAMNet embeddings used as a feature extraction stage achieved high performance for swallow sound detection, confirming its suitability for this task. We therefore adopted the same backbone configuration and adapted it to our specific frequency range. Figure 2.5 shows the employed architecture. Each preprocessed segment was fed to the YAMNet backbone adapted for low-rate audio (8 kHz). YAMNet produces a sequence of $T \times 1024$ frame-level embeddings, where T is the number of

frames obtained by sliding YAMNet's 0.96 s analysis window with 50 % overlap, as used during its AudioSet pretraining (Gemmeke *et al.*, 2017). We then summarized these embeddings by computing their temporal mean and standard deviation, yielding two 1024-dimensional summary vectors. In parallel, we computed the ZCR on the 8 kHz waveform using a 0.96 s analysis window with 50 % overlap, and summarized ZCR by its mean and standard deviation. As described by (So *et al.*, 2025), we concatenated both features to form a fixed-length 2050-dimensional descriptor :

$$\mathbf{x} = \left[\mu_{\mathrm{e}},\ \sigma_{\mathrm{e}},\ \mu_{\mathrm{ZCR}},\ \sigma_{\mathrm{ZCR}}\right] \in \mathbb{R}^{2050}.$$

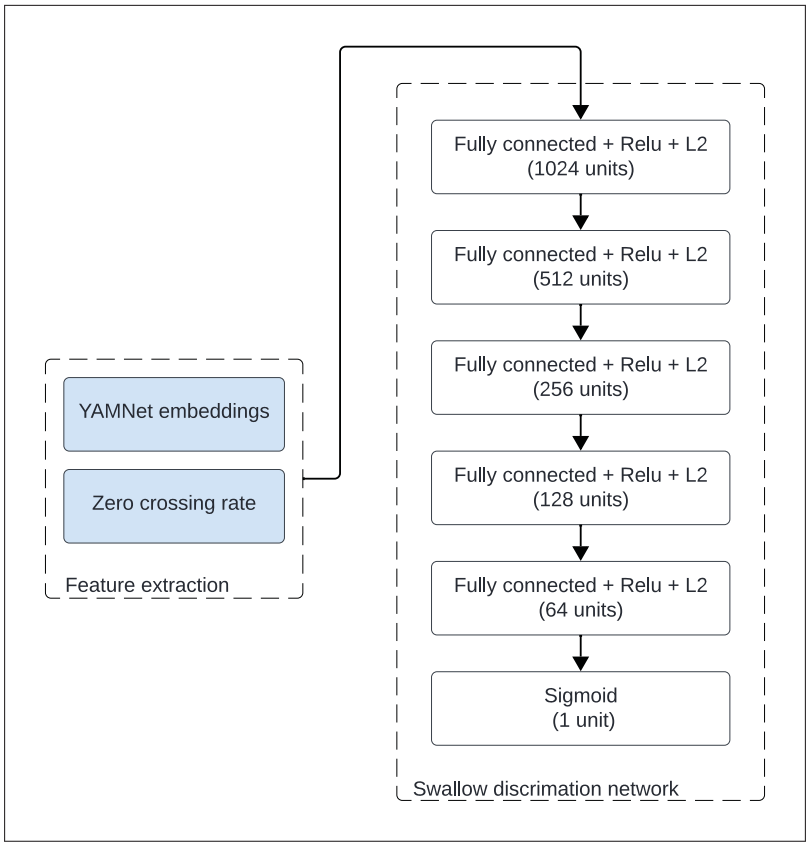


Figure 2.5 Fully connected network for swallow classification

The dataset was imbalanced, with 60 % non-swallow and 40 % swallow segments. Within the swallow class, 46 % were water, 34 % saliva, 15 % yogurt, and 5 % biscuit. We addressed class imbalance only during training using dynamic batch balancing : each mini-batch was

constructed to contain exactly half swallow and half non-swallow samples, with the swallow class further stratified so that saliva, water, yogurt and biscuit each represented 25 %. Importantly, the validation and test sets were left untouched, preserving their natural class distributions.

In the transfer learning approach used in this study, the softmax layer of YAMNet was then replaced with a fully connected neural network FCNN. The FCNN comprised five hidden layers, each followed by batch normalization, ReLU activation, and dropout regularization that were chosen empirically, with values ranging from 0.3 to 0.4. The output layer consisted of a single sigmoid unit that estimated the probability of a swallow event.

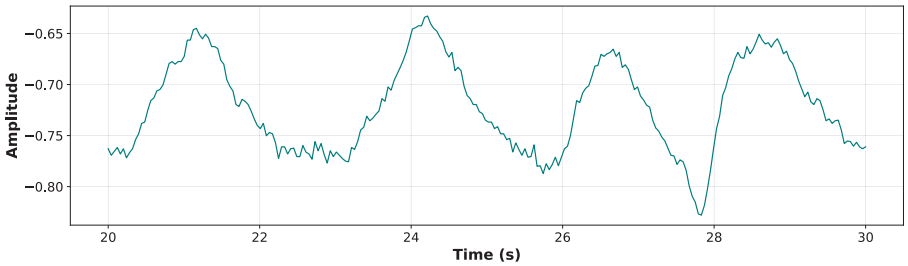
Before training, all features were standardized using statistics from the training set only to avoid information leakage. Optimization used the AdamW optimizer with a learning rate of 1×10^{-3} and a weight decay of 1×10^{-5}. Training was regularized with early stopping based on validation loss, saving the best weights, and a Reduce on Plateau schedule that halved the learning rate whenever validation performance did not show further improvement. Finally, to ensure a subject-independent evaluation, we adopted a 5-fold cross-validation where folds were defined by participant IDs. In each fold, approximately 80 % of participants (21 out of 26) were allocated to the training set and 20 % (5 out of 26) to the test set. Within the training set, 15 % of participants were further reserved for validation.

2.4 Results

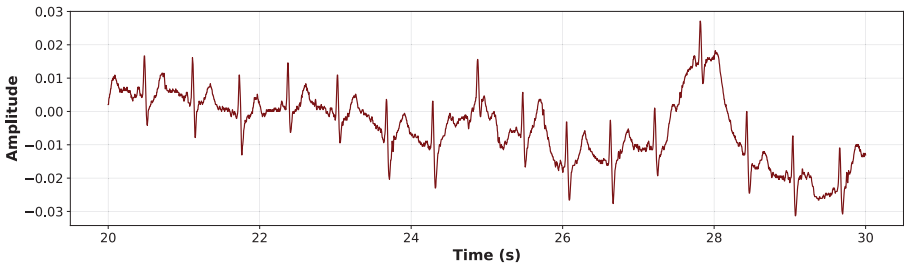
2.4.1 Dataset Overview

This section presents representative excerpts from the collected dataset to illustrate the nature and quality of the acquired signals. Figure 2.6 presents three synchronized physiological signals recorded over a 10 s window. The top panel, Figure 2.6a, shows the respiratory waveform, characterized by smooth, periodic oscillations corresponding to inhalation and exhalation cycles. This signal reflects a regular breathing pattern, captured via a chest belt sensor. The middle panel, Figure 2.6b, displays the ECG signal, where sharp and repeated peaks representing QRS

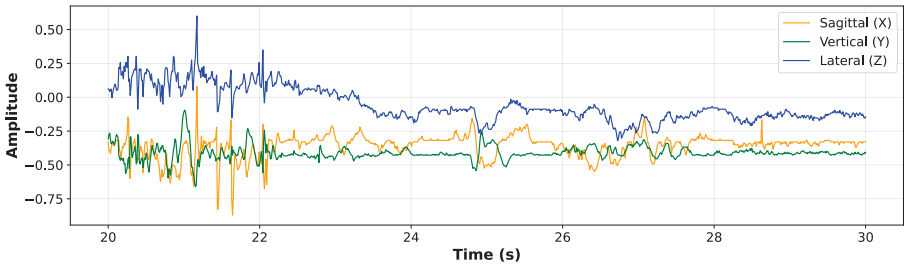
waves indicate a stable heart rhythm. Slight amplitude modulations may suggest movement artifacts. The bottom panel, Figure 2.6c, shows triaxial accelerometer data along the sagittal (X), vertical (Y), and lateral (Z) axes. Variations in these signals reflect subtle movements of the body or head, potentially related to breathing, swallowing, or postural adjustments.



a) Breathing



b) ECG



c) Accelerometer signals

Figure 2.6 Three synchronized physiological signals recorded with the Zephyr chest belt, over a 10 s window for participant 27

As shown in Figure 2.7, the different panels illustrate several stages of the swallowing process captured by submental ultrasound imaging.

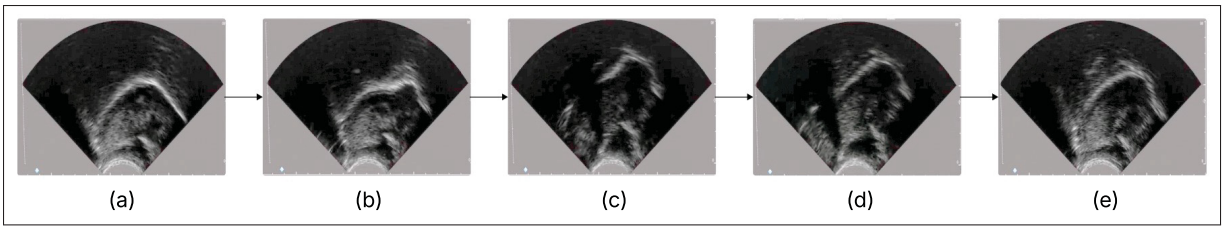


Figure 2.7 Ultrasound imaging of tongue during swallowing stages. (**a**) Resting position of the tongue prior to swallowing. (**b**) Tongue curving upward and backward, initiating bolus propulsion toward the pharynx. (**c**) Tongue retraction and floor-of-mouth muscle engagement as the bolus advances posteriorly. (**d**) Continued posterior progression of the bolus. (**e**) Repositioning of the tongue following swallowing

The observed tongue motion pattern is distinctive of swallowing. While tongue movements also occur during speech, their ultrasound characteristics differ and speech events are identifiable in the microphone recordings.

Figure 2.8 shows twelve representative 4 s in-ear audio spectrograms recorded during different verbal and non-verbal events, illustrating the acoustic diversity captured in the dataset.

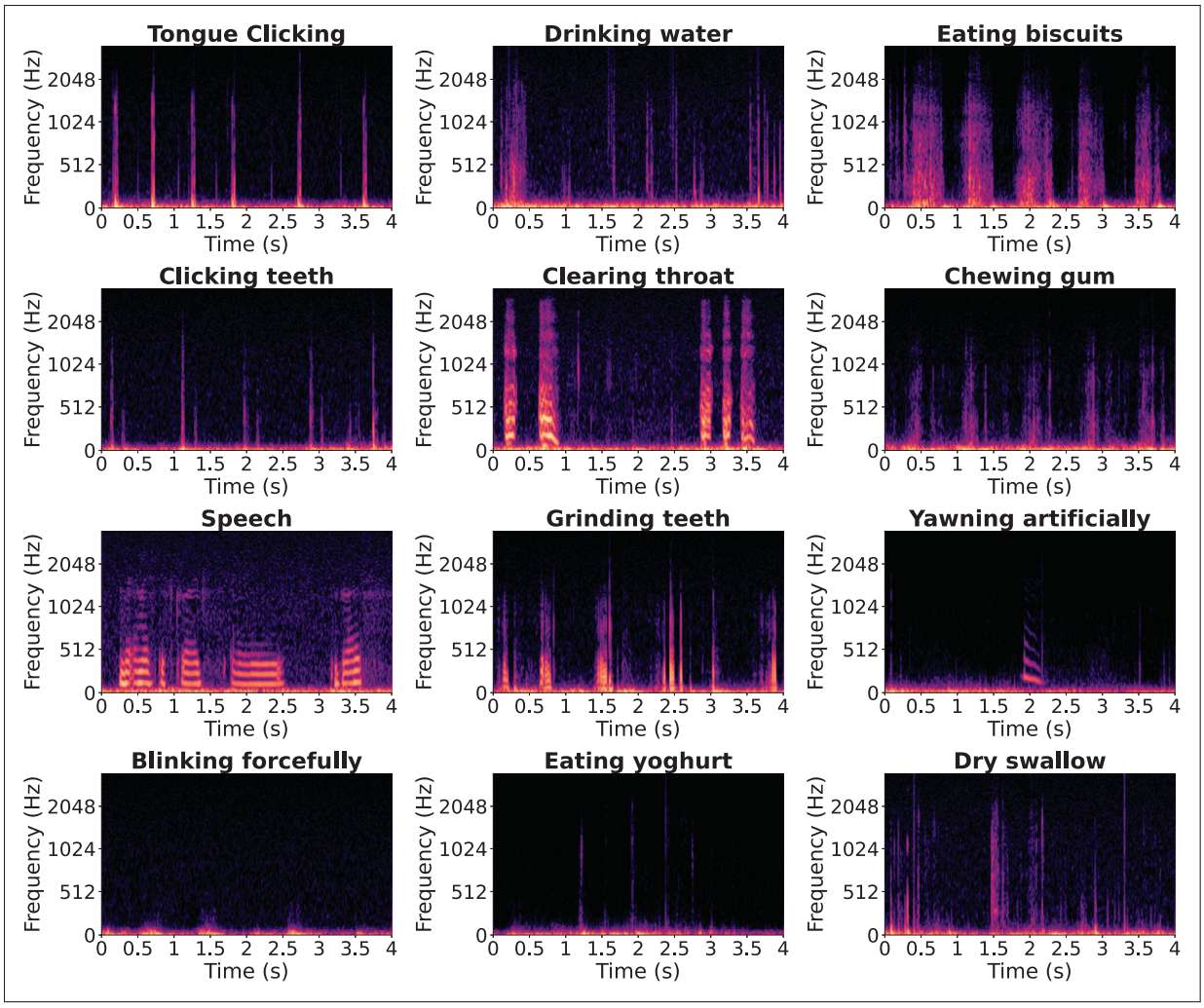
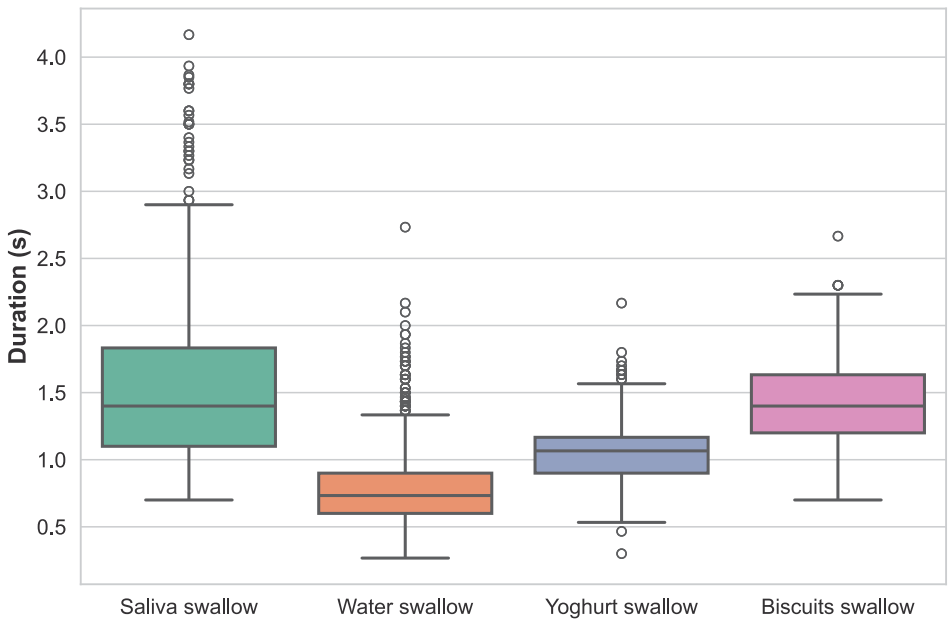
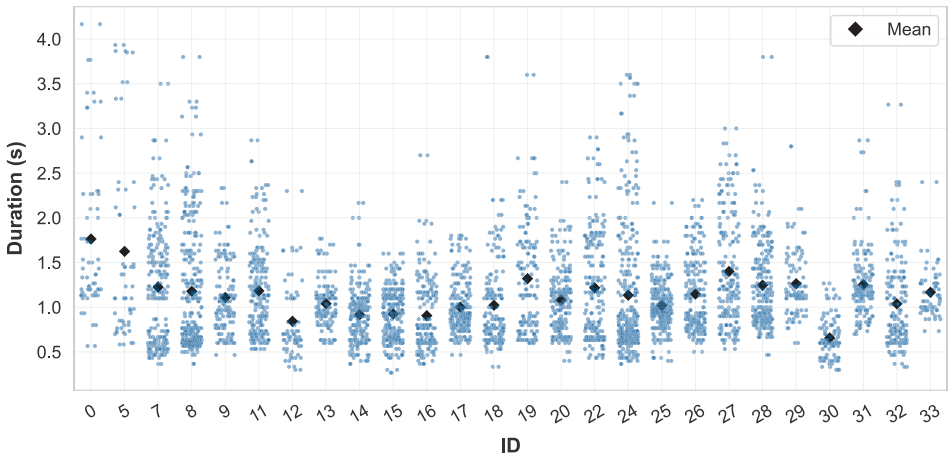


Figure 2.8 Twelve spectrograms of 4 s in-ear audio segments recorded during different verbal and non-verbal events

The swallow dataset consists of 1632 dry saliva swallows, 2208 water swallows, 740 yogurt swallows, and 228 biscuit swallows, highlighting an imbalance between textures. The distribution of swallowing durations by texture, as shown in Figure 2.9a, shows clear differences. A linear mixed-effects model with texture as a fixed effect and a random intercept per participant confirmed a significant influence of texture on swallowing duration ($p < 0.001$). On average, saliva swallows were the longest ($M = 1.55 \pm 0.59$ s), followed by yogurt ($M = 1.04 \pm 0.24$ s) and biscuits ($M = 1.43 \pm 0.35$ s), while water swallows were markedly shorter ($M = 0.79 \pm 0.28$ s).



a) Swallowing duration by texture



b) Intra-individual variability in swallowing duration

Figure 2.9 Analysis of swallowing durations across textures and intra-individual variability

In addition to texture effects, inter-subject variability was quantified using the variance components of the mixed-effects model with a random intercept per participant. The estimated

between-subject variance (0.0375) reflects systematic differences in average swallowing duration across individuals, whereas the within-subject residual variance (0.1229) captures intra-individual variability and unexplained fluctuations. These components correspond to an ICC of 0.234, indicating that approximately 23 % of total swallowing-duration variability is attributable to inter-individual differences.

Substantial intra-individual variability was observed as shown in Figure 2.9b. Even within the same participant, swallowing durations ranged from less than 0.5 s to more than 3 s.

2.4.2 Binary Classification Results

In this application, minimizing both false negatives (FNs) and false positives (FPs) is equally important. Therefore, the F1-score is used as the main evaluation metric, as it provides a balance between precision and recall. The model was evaluated using 5-fold cross-validation. Mean performance, summarized in Table 2.2, was strong and consistent across folds (Accuracy 0.874 ± 0.013, F1-score 0.875 ± 0.013). The classifier demonstrated excellent discriminative ability, achieving an AUROC of 0.942 ± 0.010 and an AUPRC of 0.931 ± 0.014. Small fold-to-fold variations were observed. The lowest AUROC/AUPRC (Fold 5) can be attributed to comparatively lower segment counts and higher intra-subject variability in swallowing duration in the corresponding participant split, reducing training diversity.

Tableau 2.2 Five-fold cross-validation metrics. Best values per metric are in bold

Fold	**Accuracy**	**F1**	**AUROC**	**AUPRC**
Fold 1	0.880	0.878	0.949	0.943
Fold 2	**0.889**	**0.895**	**0.949**	**0.944**
Fold 3	0.873	0.873	0.943	0.932
Fold 4	0.872	0.864	0.945	0.927
Fold 5	0.854	0.864	0.924	0.909
Mean ± SD	**0.874 ± 0.013**	**0.875 ± 0.013**	**0.942 ± 0.010**	**0.931 ± 0.014**

The confusion matrix (Figure 2.10) shows balanced behavior across classes. Averaged over folds, the classifier correctly identified 86.6 % of non-swallows, with a corresponding false-positive rate of 13.4 %. For swallows, the average true-positive rate was 88.1 %, with 11.9 % missed swallows (FNs).

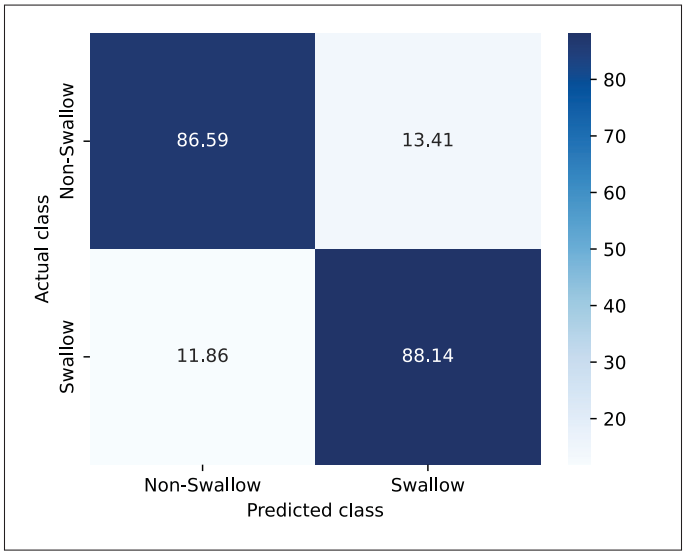


Figure 2.10 Average confusion matrix across five folds

To assess the influence of acoustic conditions, performance was also evaluated separately for each environment (quiet, babble, and factory noise), as summarized in Table 2.3. Overall, the classifier demonstrated stable performance across noise conditions. Performance slightly decreased in the babble noise condition (AUROC 0.930 ± 0.01, F1-score 0.850 ± 0.014), which may be attributed to the spectral overlap between human babble noise and swallowing sounds. The limited impact of background noise on classification performance may be explained by the acoustic seal created by the ear tips, which passively attenuates external noise while preserving internally generated physiological signals such as swallowing sounds.

Tableau 2.3 AUROC and F1-score across noise conditions

	Quiet		**Babble**		**Factory**	
	AUROC	**F1**	**AUROC**	**F1**	**AUROC**	**F1**
Mean	0.952	0.890	0.930	0.850	0.944	0.884
SD	±0.010	±0.013	±0.010	±0.014	±0.011	±0.011

2.5 Discussion and Limitations

Post-classification analyses showed consistent performance (F1 = 0.875) in all folds, indicating that in-ear audio carries sufficient information for swallow/non-swallow discrimination between individuals.

Across folds, the majority of false positives came from saliva mix segments that were labeled as non-swallow but are acoustically similar to saliva swallows, especially to the oral preparatory phase, when the tongue contracts to propel the bolus against the soft palate.

FNs were dominated by saliva swallows (50.5 %), followed by water (19.1 %), yogurt (18.9 %) and biscuits (11.4 %). FNs were most often aligned with atypical or prolonged swallowing patterns. As detailed in Section 2.3.1.2, participants were asked to swallow their saliva repeatedly over a 30 s window. This repetitive task progressively dries the oral cavity, reducing the

available saliva. In practice, participants produced roughly five swallows within that window, the fourth and fifth swallows tended to deviate from typical patterns showing longer durations and more fragmented acoustics because repeated, effortful dry swallowing is harder to execute. These atypical dynamics diverge from the normal swallow signature and, therefore, were more frequently missed by the model. Such task-induced modulation of swallowing timing and dynamics in healthy individuals has been previously reported, including differences between cued and non-cued swallowing conditions (Nagy *et al.*, 2013).

Despite the careful design and implementation of our data acquisition protocol, several limitations must be recognized. Although the dataset includes multimodal recordings from a chest belt (ECG, respiration, and acceleration), these signals were not always available with optimal quality. In particular, ECG recordings were occasionally missing or degraded.

Another limitation is the use of fixed 4 s segments for classification, which may include extended low-activity portions in non-swallow samples and reduce temporal precision. Future work will compare shorter and sliding-window segmentations.

Label validity and reproducibility is also a limitation. Annotation timing is bounded by the 30 fps ultrasound resolution, and labels were generated by a single rater with partial expert review. In the absence of reliability assessment (Cohen's kappa) and a label validation study, residual label uncertainty cannot be excluded.

It is also important to note that the dataset consists exclusively of healthy young adults aged 20 to 29 years. As a result, it may not adequately represent populations with different physiological characteristics or medical conditions, such as older adults or individuals with swallowing problems. The primary objective of this study was to establish proof of concept in a healthy population. Future work will extend the dataset to elderly individuals and patients with swallowing disorders, such as dysphagia or Parkinson's disease, to evaluate clinical generalizability.

The classification experiment was designed as a baseline proof of concept. We intentionally used a lightweight FCNN on top of frozen YAMNet embeddings and did not attempt to

optimize performance with more complex architectures (CRNN, Transformers). Preliminary experiments with from-scratch models (CNNs trained on log-Mel spectrograms and per-channel energy normalization PCEN representations) showed lower cross-subject generalization than the transfer-learning approach. Although ultrasound and chest-belt signals were collected, they were not used as auxiliary classifier inputs in this study because our objective was to evaluate a practical, non-invasive swallow detection approach based only on in-ear audio. Future work will investigate multimodal learning and fusion, and more complex models.

2.6 Conclusions

In this work, we present the first comprehensive multimodal dataset dedicated to swallowing and related non-verbal events, recorded using IEM and synchronized physiological signals. It addresses the current lack of publicly available resources for training and validating swallowing detection algorithms. Due to the wide range of tasks performed in both quiet and noisy environments, this dataset is well suited for developing robust and real-world systems across various research and clinical projects. As an initial proof of concept, an FCNN trained on YAMNet+ZCR embeddings achieved a mean F1-score of 0.875, supporting the feasibility of in-ear audio-based swallow detection and motivating future multimodal and clinical validation.

CHAPITRE 3

SWALLOWING DETECTION FROM BILATERAL IN-EAR AUDIO

Elyes Ben Cheikh[1] , Yassine Mrabet[1,2] , Catherine Laporte[1] , Rachel E. Bouserhal[1,2]

[1] Department of Electrical Engineering, École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Centre for Interdisciplinary Research in Music Media and Technology,
527 Rue Sherbrooke O, Montréal, Québec, Canada H3A 1E3

Article soumis pour publication dans la revue *«Applied Acoustics»*

3.1 Abstract

Swallowing is a vital physiological function whose frequency and temporal dynamics reflect underlying neuromuscular health. Abnormal swallowing patterns are clinically associated with dysphagia and sialorrhea. Yet, standard assessments rely on invasive instrumentation and brief in-clinic recordings, which are incompatible with continuous monitoring in everyday settings. To address this gap, this paper presents a method for detecting swallow events in bilateral in-ear audio recordings captured from a wearable device. Raw recordings are resampled to 4 kHz. Time-frequency features are extracted as Per-Channel Energy Normalization PCEN spectrograms computed over 1.4 s sliding windows with 0.14 s hop. A CRNN with shared bilateral encoding and adaptive confidence-based fusion produces window-level swallow probabilities, which are converted into discrete events via a user-adaptive hysteresis detector calibrated from a short baseline recorded at the start of each session. The non-swallow class comprises speech, normal breathing, chewing, throat clearing, teeth and tongue clicking, and physical movement. Evaluated on 34 participants using subject-disjoint five-fold cross-validation, window-level classification achieves a mean F1-score of 0.892 ± 0.006, and event-level detection achieves a mean F1-score of 0.835 ± 0.012 with precision of 0.779 ± 0.014 and recall of 0.900 ± 0.020 across quiet, factory, and babble noise conditions.

3.2 Introduction

Swallowing is a complex physiological process that transports food, liquids, and saliva from the oral cavity to the stomach. It involves the coordinated activity of multiple anatomical structures across three phases oral, pharyngeal, and esophageal requiring precise coordination between voluntary and reflexive muscle actions to safely propel the bolus through the aerodigestive tract (Matsuo & Palmer, 2008; Jean, 2001). Beyond nutrition and hydration, swallowing also contributes to airway protection by regularly clearing saliva and secretions from the pharynx (Dodds, Stewart & Logemann, 1990).

In healthy individuals, spontaneous saliva swallowing rates reported in the literature vary widely, from approximately 0.14 to 6.7 swallows per minute during wakefulness (Lear *et al.*, 1965; Rudney, Ji & Larson, 1995). Monitoring swallowing longitudinally may therefore offer valuable insights into swallowing function. In particular, reduced SSF has been identified as a sensitive indicator of dysphagia across several clinical populations. In acute stroke, spontaneous swallow frequency analysis achieved 96 % sensitivity in identifying dysphagia, outperforming nurse-administered clinical screening on the same cohort (Crary, Carnaby & Sia, 2014). Accumulated oropharyngeal secretions resulting from infrequent swallowing have also been associated with increased aspiration risk (Murray, Langmore, Ginsberg & Dostie, 1996). However, normative SSF values remain poorly established, and measurement heterogeneity across studies prevents the definition of clinically actionable thresholds (Bulmer *et al.*, 2021).

Tracking swallowing patterns over time may help reveal abnormal behavior. Reduced swallowing frequency can impair saliva clearance and lead to pharyngeal secretion accumulation, a phenomenon frequently observed in Parkinson's disease where drooling (sialorrhea) is often linked to insufficient swallowing rather than excessive saliva production (Chou *et al.*, 2007). Similarly, longitudinal monitoring may help reveal early signs of dysphagia by detecting changes in swallowing frequency or duration. Dysphagia can lead to aspiration pneumonia, malnutrition, dehydration, and reduced quality of life (González-Fernández, Ottenstein, Atanelov & Christian,

2013; Sura, Madhavan, Carnaby & Crary, 2012), and is common after stroke, where it is associated with increased risk of aspiration pneumonia and mortality (Martino *et al.*, 2005).

Clinical assessment of swallowing primarily relies on VFSS and FEES. These gold-standard techniques allow direct visualization of swallowing physiology and detection of aspiration or pharyngeal residue (Logemann, 1998; Castro *et al.*, 2025). However, they are performed in specialized settings, provide only brief observations under controlled conditions, and are unsuitable for continuous monitoring in everyday life. Beyond these gold-standard evaluations, spontaneous swallowing frequency has been proposed as a non-invasive indicator of dysphagia risk (Crary *et al.*, 2014; Bulmer *et al.*, 2021). However, a systematic review of 19 studies revealed that no standardized method exists for measuring SSF (Bulmer *et al.*, 2021). This lack of standardization underscores the need for robust, automated swallowing detection systems that can provide consistent and reproducible measurements across sessions and individuals.

To enable more continuous monitoring, several sensing approaches have been proposed using sensors placed on the neck, including microphones, accelerometers, surface electromyography, or strain sensors to capture acoustic, vibratory, or muscular signals associated with swallowing (So *et al.*, 2023). Although promising in controlled settings, these approaches present practical limitations : signal quality is sensitive to sensor placement and anatomical variability, and neck-mounted sensors are susceptible to motion artifacts from neck movements, clothing friction, and postural changes. From a usability perspective, surface electrodes and adhesive sensors can cause skin irritation and may detach during daily activities, reducing practicality for long-term use.

Consumer-grade wearables, including smartwatches and ear-worn devices, are increasingly deployed for continuous physiological monitoring, enabling long-term data collection with minimal disruption to daily activities (Pantelopoulos & Bourbakis, 2010; Heikenfeld *et al.*, 2018). However, wearable monitoring systems must satisfy several practical constraints to be effective in real-world settings. Sensors must be comfortable to wear, socially acceptable, energy efficient, and robust to environmental noise and body motion. These challenges are particularly

relevant for swallowing monitoring. Unlike other physiological signals such as heart rate or respiration, swallowing events are brief, irregular, and relatively sparse.

In-ear sensing has recently emerged as a promising alternative for unobtrusive physiological monitoring. Ear-worn devices, often referred to as hearables, provide a discreet and socially acceptable platform for continuous sensing (Martin & Voix, 2018; Röddiger *et al.*, 2025). Self-reported headphone use averages approximately 1.3 h/day in women and 1.9 h/day in men, suggesting meaningful daily wear time for sensing applications (Shrimal & Nandurkar, 2021). When the ear canal is partially or fully occluded by an earbud, internally generated sounds transmitted through bone and soft tissue are amplified by the occlusion effect, improving the SNR for subtle physiological events. IEMs can capture internal body sounds even under substantial ambient noise (Martin & Voix, 2018; Chabot *et al.*, 2021), making ear-based sensing well-suited for monitoring physiological signals in everyday environments.

Automatic approaches to swallowing analysis were first developed mainly from cervical recordings, relying on microphones or accelerometers placed over the anterior neck. Detection was typically formulated as a supervised classification problem based on hand-crafted descriptors and shallow models. Kimura et al. combined cepstral features with SVM and ensemble classifiers on a clinical throat-sound database (Kimura *et al.*, 2024), while Gravellier et al. addressed pharyngolaryngeal activity detection in wearable cervical sensing (Gravellier *et al.*, 2024). Related work on eating and swallowing sounds has explored multichannel neck and peri-auricular sensing with MFCC-based representations and sequence models (Nakamura *et al.*, 2021). More recent studies have shifted toward learned representations and deep models, including transfer learning from general-purpose audio embeddings, multimodal wearable systems, and deep or ensemble architectures for throat-related event analysis (So *et al.*, 2025; Jayatilake *et al.*, 2026; Song *et al.*, 2025; Shin *et al.*, 2024). In-ear swallowing studies remain limited in scope : existing work has mainly validated the ear canal as a sensing site and characterized intra-aural swallow acoustics (Yoshimoto *et al.*, 2022; Haji & Yamaguchi, 2025; Asakura & Miwa, 2025). Ear-mounted sensing has shown promise for related tasks such as eating-episode detection (Bi *et al.*, 2018), but robust event-level swallow detection from in-ear audio in realistic daily conditions remains

largely unaddressed. This gap is particularly important because IEMs are exposed to nearby physiological and acoustic interferers, including breathing, speech, mastication, and other nonverbal body sounds (Mehrban *et al.*, 2024; Papapanagiotou *et al.*, 2017b; Chabot *et al.*, 2021).

Despite growing interest in hearable sensing, no prior study has demonstrated end-to-end swallowing event detection from in-ear audio. Existing work has either validated the ear canal as a sensing site without proposing a detection system (Yoshimoto *et al.*, 2022; Haji & Yamaguchi, 2025; Asakura & Miwa, 2025), relied on indirect heuristics such as inter-chewing interval analysis rather than acoustic modeling of the swallowing signal itself, or focused on cervical rather than in-ear recordings (Kimura *et al.*, 2024; Gravellier *et al.*, 2024). The present work addresses this gap directly. In this work, we propose a complete framework for detecting swallow events from bilateral in-ear audio recordings at a low sampling rate of 4 kHz. The framework relies on a PCEN-based feature representation to improve robustness to environmental noise and employs a dual-ear CRNN that integrates convolutional feature extraction, a BiLSTM for temporal modeling, attention-based pooling, and adaptive confidence-based fusion of both ears. Finally, a hysteresis-based post-processing strategy converts window-level predictions into discrete swallow events using a calibrated baseline. The proposed framework is evaluated on a database of 34 healthy participants using subject-disjoint five-fold cross-validation, under three ambient conditions (quiet, factory, babble) and across three bolus types (dry, water, yogurt). Performance is reported at both the window level (classification) and the event level (detection), and systematic ablation, feature, architecture, and fusion studies are conducted to isolate the contribution of each design choice.

3.3 Materials and methods

3.3.1 Dataset

This study builds on a multimodal in-ear audio and physiological dataset previously introduced in (Cheikh, Mrabet, Laporte & Bouserhal, 2026). The data were collected from 34 healthy

adults (20 males, 14 females), aged 20–29 years, using custom IEMs mounted in earpieces that tightly sealed the ear canal. Each participant completed a structured protocol designed to elicit swallowing events and other common non-verbal activities.

Swallowing tasks comprised voluntary swallows of three bolus types : liquid (still water), semi-solid (plain yogurt), and sequences of repeated dry swallows. To capture likely confounders, the protocol also included tongue clicks, teeth clicking/grinding, chewing gum, throat clearing, forceful blinks, physical movements, and periods of quiet breathing or speech. Recordings were acquired under both quiet and noisy (babble and factory) ambient conditions.

Swallow events were manually annotated using synchronized ultrasound recordings of the tongue as ground-truth reference. Ultrasound recordings were examined frame by frame to identify candidate events showing the characteristic propulsion–retraction pattern of the tongue associated with swallowing, which were then confirmed when a temporally aligned swallow sound was observed in the bilateral IEM recording. Each confirmed event was marked with an onset and offset timestamp on the IEM signal, and these annotations serve as the reference intervals for both window-level labeling and event-level evaluation throughout this work.

For the experiments reported in this work, a balanced subset of the full database was constructed to ensure sufficient representation of both swallow and non-swallow activity, and to prevent any single non-swallow category from dominating the negative class. All annotated swallow events were retained, regardless of bolus type (dry, water, yogurt) or noise condition. In contrast, non-swallow segments were subsampled within each category : speech, breathing, chewing, throat clearing, tongue and teeth clicking, and physical movement, so that the categories contribute approximately equal amounts of data to the negative class. The resulting sliding-window corpus contains approximately 40 % positive (swallow) and 60 % negative (non-swallow) windows, with the negative class itself balanced across its constituent acoustic sub-classes. This design preserves the acoustic diversity of the original database while avoiding the extreme class imbalance that would result from sampling continuous recordings uniformly, given the sparse nature of swallowing events.

3.3.2 Acoustic seal verification and interaural asymmetry

At data acquisition time, adequate acoustic sealing was verified using two complementary measures. First, fit check segments consisting of a short pink-noise exposure at 85 dBA were recorded simultaneously by the IEM and OEM at the beginning of each recording sequence. The custom earpiece fit was considered acceptable if the OEM-to-IEM attenuation reached at least 8 dB at 250 Hz (Bernier & Voix, 2013). Six fit check panels were performed across the session (Cheikh *et al.*, 2026).

Second, to track seal quality within each sequence, the interaural RMS asymmetry was monitored continuously. Because recordings were acquired inside a controlled audiometric environment, any interaural difference in IEM signal energy must arise from differences in canal occlusion. The normalized asymmetry index

$$A_i = \frac{\mathrm{RMS}_L - \mathrm{RMS}_R}{\mathrm{RMS}_L + \mathrm{RMS}_R} \tag{3.1}$$

serves as a time-varying indicator of relative occlusion quality, with positive values indicating a tighter left-ear seal.

Fig. 3.1(a) shows that the right channel consistently exhibits higher RMS values ($\mathrm{RMS}_R \approx 0.007$–$0.011$ vs. $\mathrm{RMS}_L \approx 0.005$–$0.008$), indicating that the left ear is, on average, more tightly sealed. However, Fig. 3.1(b) reveals that this dominance pattern is neither uniform across participants nor stable across panels : 47 % of participants exhibit at least one reversal of dominant ear across the six panels. This variability is consistent with progressive earbud displacement, perspiration-induced changes in foam tip compliance, and jaw movements affecting canal occlusion.

These findings demonstrate that the relative reliability of the left and right IEM channels is a session-dependent, time-varying property that cannot be inferred from a static fit check measurement. This directly motivates the adaptive fusion strategy described in Section 3.3.4 :

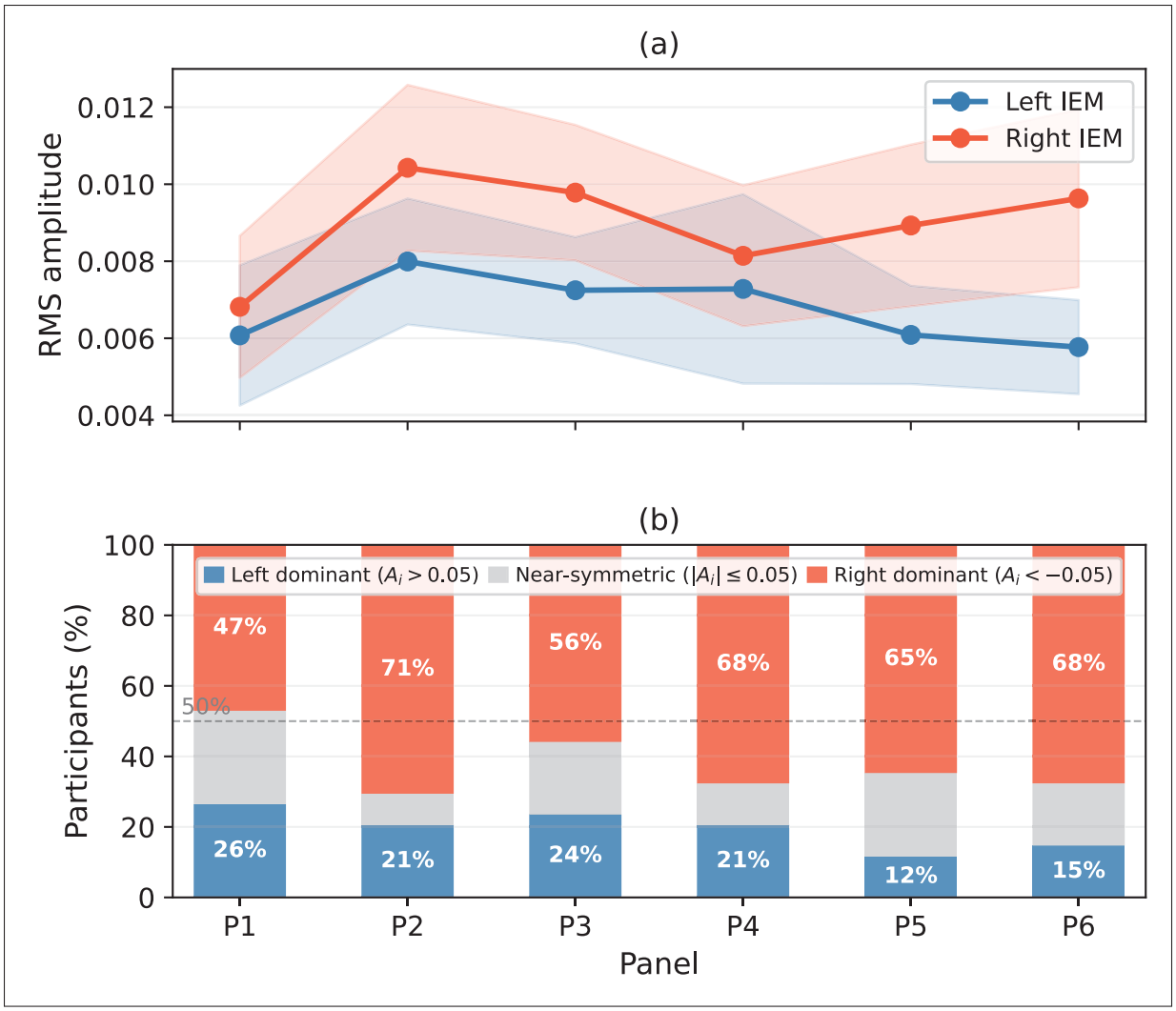


Figure 3.1 Interaural seal asymmetry across six fit check panels. (a) Mean IEM RMS amplitude for left and right channels. (b) Proportion of participants per dominant ear

rather than assigning a fixed weight to one ear, the model dynamically estimates which channel provides the cleaner swallowing evidence at each analysis window.

3.3.3 Preprocessing and feature extraction

All audio signals were resampled to 4 kHz prior to feature extraction. This sampling rate reduces storage and processing requirements while preserving the spectral content most relevant to swallowing sounds, which is mainly concentrated at low frequencies due to the occlusion

effect (Brummund *et al.*, 2014). Stereo information is preserved by applying the resampling independently to each channel.

The continuous recordings were segmented using a fixed-length sliding window of duration $T_w = 1.4\,\text{s}$ with 90 % overlap, resulting in a hop length of $T_h = 0.14\,\text{s}$. Each window is labeled as swallowing if it overlaps at least 20 % with any annotated swallow interval. Window duration was selected via grid search over the validation set, maximizing window-level F1-score while ensuring coverage of the longest observed swallow duration.

Amplitude normalization was applied to reduce variability introduced by differences in recording conditions. This normalization reduces inter-session and inter-participant variability.

Time-frequency features were computed as Mel spectrograms independently for each channel, with $n_{\text{mels}} = 64$ Mel bands, $n_{\text{fft}} = 256$, hop length 40, and frequency range $f_{\text{min}} = 20\,\text{Hz}$ to $f_{\text{max}} = 1800\,\text{Hz}$.

Following Mel spectrogram computation, PCEN was applied independently to each stereo channel. PCEN is a nonlinear transformation that performs adaptive gain normalization and dynamic-range compression in the time-frequency domain (Lostanlen *et al.*, 2019). Let $S_{f,t}$ denote the Mel spectral energy at frequency bin f and time frame t. A smoothed temporal envelope was first computed using an exponential moving average :

$$M_{f,t} = (1 - s)\, M_{f,t-1} + s\, S_{f,t}. \tag{3.2}$$

The PCEN transformation is defined as :

$$\text{PCEN}_{f,t} = \left(\frac{S_{f,t}}{(\epsilon + M_{f,t})^{\alpha}} + \delta \right)^{r} - \delta^{r}, \tag{3.3}$$

with parameters $s = 0.025$, $\alpha = 0.75$, $\delta = 2.0$, $r = 0.5$, and $\epsilon = 10^{-6}$, selected empirically on the validation set. The resulting stereo PCEN features form a tensor of shape $(2, 64, T)$.

PCEN provides several advantages over conventional logarithmic compression. The adaptive normalization compensates for slow variations in signal energy, making the representation robust to channel differences and microphone placement variability. The nonlinear compression enhances transient events while suppressing stationary background noise, properties that are particularly beneficial for detecting brief, low-amplitude swallowing events under low-SNR conditions.

3.3.4 Dual-ear CRNN architecture

The classifier operates on stereo PCEN feature tensors of shape $(2, 64, T)$. Rather than concatenating both channels at the input, each ear is processed independently through a *shared* encoder. This design preserves ear-specific acoustic observations while encouraging representations that generalize across channels. The complete architecture is illustrated in Fig. 3.2.

The single-channel encoder (Fig. 3.3) consists of two convolutional blocks followed by bidirectional temporal modeling. Each convolutional block applies a 3×3 convolution with batch normalization and ReLU activation, followed by max-pooling along the frequency axis only, progressively compressing the frequency dimension while preserving temporal resolution. The resulting feature maps are reshaped into frame-level embeddings and fed into a BiLSTM layer with hidden size 128 in each direction, enabling exploitation of both past and future context within the analysis window.

A temporal attention module aggregates the BiLSTM output sequence into a fixed-dimensional embedding for each ear. Let $\mathbf{h}_t^{(e)} \in \mathbb{R}^{256}$ denote the BiLSTM output at frame t for ear $e \in \{L, R\}$. The attention weights are computed as :

$$a_t^{(e)} = \frac{\exp\left(\mathbf{w}^\top \tanh\left(\mathbf{W}\,\mathbf{h}_t^{(e)} + \mathbf{b}\right)\right)}{\sum_{t'} \exp\left(\mathbf{w}^\top \tanh\left(\mathbf{W}\,\mathbf{h}_{t'}^{(e)} + \mathbf{b}\right)\right)}, \tag{3.4}$$

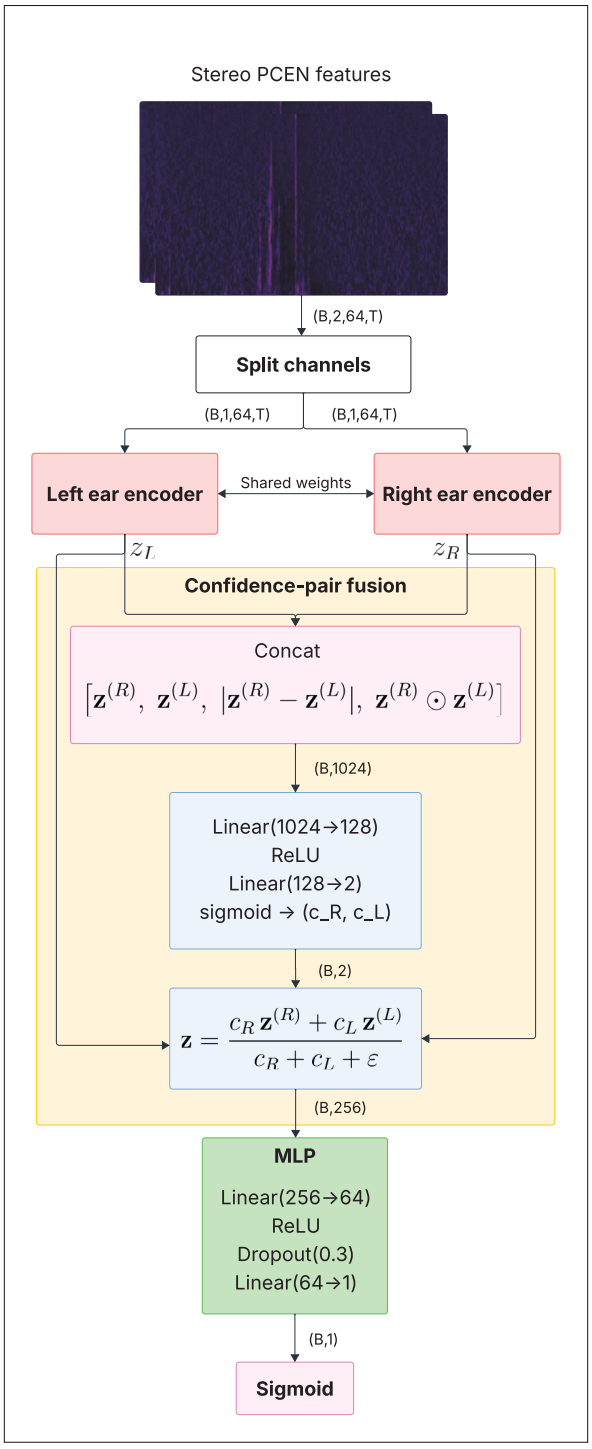


Figure 3.2 Proposed dual-ear CRNN architecture with shared encoder, BiLSTM temporal modeling, attention pooling, and adaptive confidence-based fusion

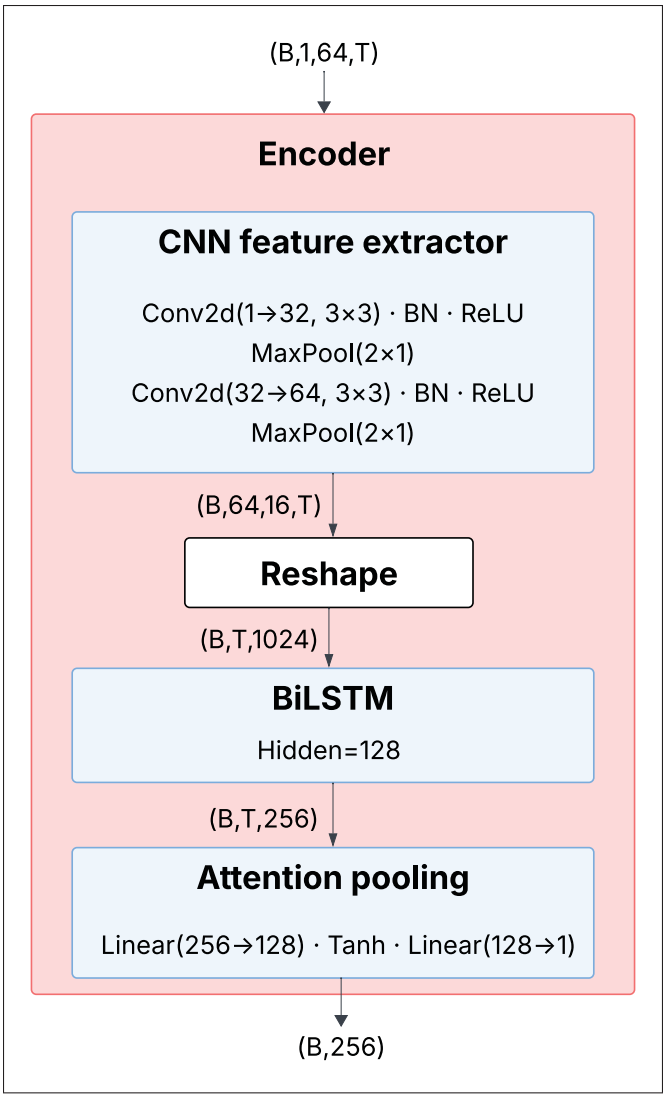


Figure 3.3 Shared encoder architecture : convolutional blocks, BiLSTM, and attention pooling

where **W**, **b**, and **w** are learned parameters shared across both ears. The ear-level embedding is the weighted sum :

$$\mathbf{z}^{(e)} = \sum_t a_t^{(e)} \mathbf{h}_t^{(e)}, \qquad \sum_t a_t^{(e)} = 1. \tag{3.5}$$

By assigning higher weights to the most informative time frames, this mechanism down-weights silence, stationary background noise, and irrelevant physiological activity.

The left and right embeddings $\mathbf{z}^{(L)}$ and $\mathbf{z}^{(R)}$ are combined through a learned late-fusion module. The fusion block receives as input the concatenation :

$$\mathbf{u} = \left[\mathbf{z}^{(R)},\ \mathbf{z}^{(L)},\ |\mathbf{z}^{(R)} - \mathbf{z}^{(L)}|,\ \mathbf{z}^{(R)} \odot \mathbf{z}^{(L)}\right], \tag{3.6}$$

where $\odot$ denotes element-wise multiplication. This representation captures complementary information (direct embeddings), differential information (absolute difference), and joint structure (element-wise product). A two-layer MLP maps $\mathbf{u}$ to two scalar confidence scores via sigmoid activations :

$$(c_R,\ c_L) = \sigma\big(\mathrm{MLP}_{\mathrm{conf}}(\mathbf{u})\big), \qquad c_R,\ c_L \in (0, 1). \tag{3.7}$$

The fused embedding is the confidence-weighted average :

$$\mathbf{z} = \frac{c_R\, \mathbf{z}^{(R)} + c_L\, \mathbf{z}^{(L)}}{c_R + c_L + \varepsilon}, \tag{3.8}$$

where $\varepsilon = 10^{-6}$. This formulation allows the model to rely more heavily on the ear providing the clearest swallowing evidence in each window. The fused embedding is passed through a classification head consisting of a linear layer ($256 \to 64$), ReLU activation, dropout ($p = 0.3$), and a final linear layer ($64 \to 1$), producing the window-level swallow logit.

To prevent the model from collapsing into a single channel, auxiliary linear classifiers are attached to $\mathbf{z}^{(L)}$ and $\mathbf{z}^{(R)}$ during training and discarded at inference. Additionally, modality dropout is applied at the embedding level : with probability $p = 0.30$, one embedding is set to zero before fusion, forcing the confidence-fusion module to operate reliably even when one channel is absent or degraded. The complete network contains 1.38 million trainable parameters including auxiliary heads.

3.3.5 Training procedure

The model minimizes a composite binary cross-entropy loss :

$$\mathcal{L} = \mathcal{L}_{\text{BCE}}(\hat{y},\, y) + \alpha\Big[\mathcal{L}_{\text{BCE}}\big(\hat{y}^{(R)},\, y\big) + \mathcal{L}_{\text{BCE}}\big(\hat{y}^{(L)},\, y\big)\Big], \tag{3.9}$$

where $\hat{y}$, $\hat{y}^{(R)}$, and $\hat{y}^{(L)}$ are logits from the main and auxiliary heads, $y \in \{0, 1\}$ is the ground-truth label, and $\alpha = 0.30$. The auxiliary terms ensure that each ear embedding remains individually discriminative. To compensate for class imbalance (approximately 60 % non-swallow, 40 % swallow), each BCE term applies positive-class up-weighting proportional to the negative-to-positive ratio within the training fold.

The network is optimized with Adam (learning rate 10^{-3}, weight decay 10^{-4}). A ReduceLROnPlateau scheduler halves the learning rate when validation AUROC shows no improvement over 3 consecutive epochs. Training runs for a maximum of 30 epochs, with early stopping triggered after 6 epochs without improvement exceeding 10^{-4}.

Three data augmentation strategies are applied. SpecAugment is applied to the PCEN feature maps : each sample undergoes a random temporal shift of up to ±6 frames, followed by 2 time masks and 2 frequency masks with maximum widths of 12 frames and 10 bins, respectively. This encourages the model to rely on distributed spectro-temporal patterns rather than fixed-position cues, which is important given the temporal uncertainty in swallow onset within labeled windows. At the input level, one ear channel is randomly zeroed with probability 0.10, simulating complete sensor loss. Finally, embedding-level modality dropout ($p = 0.30$) is applied before fusion, providing complementary regularization at two stages of the pipeline.

3.3.6 Event-level detection

Window-level classification provides a swallow probability for each 1.4 s segment, whereas the final objective is the detection of discrete swallow events in continuous recordings. Inference is performed using a sliding-window procedure with the same hop size as for training, producing

a dense probability sequence $p(t) \in [0, 1]$ over time. However, directly applying a threshold to the raw probability sequence is often unstable in practice. The output of the classifier may exhibit short-term fluctuations due to motion artifacts or the sliding-window inference procedure itself, which produces overlapping predictions that can vary from one window to the next. For these reasons, a smoothing and adaptive thresholding strategy is applied before event extraction.

The raw posterior sequence is temporally smoothed using an EMA :

$$\tilde{p}(t) = \alpha\, p(t) + (1 - \alpha)\, \tilde{p}(t - 1), \tag{3.10}$$

where $\alpha \in (0, 1]$ controls the trade-off between responsiveness and stability.

To account for inter-subject and inter-session variability, detection thresholds are calibrated to each user and session using a short baseline segment recorded at the start of the session, prior to any swallowing activity. In real-world deployment, this calibration could be performed automatically during the first minute following the hearable insertion. In the present study, a non-swallow baseline segment of at least 45 s, recorded between the first fit check panel and the beginning of the first session, is used to characterize the participant-specific non-event posterior distribution. This segment contains spontaneous breathing, minor postural adjustments, and ambient exposure, but no swallowing activity. These baseline data are excluded from model training, validation, and testing. Let $\mathcal{N}$ denote the set of smoothed probabilities from this baseline. Two adaptive thresholds are estimated from the empirical quantiles of $\mathcal{N}$:

$$\tau_{\text{low}} = \min\left(Q_{0.85}(\mathcal{N}),\, 0.85\right), \qquad \tau_{\text{high}} = \min\left(Q_{0.90}(\mathcal{N}),\, 0.95\right). \tag{3.11}$$

The lower quantile ($Q_{0.85}$) characterises the majority of non-event posteriors and sets the hysteresis floor. The cap at 0.85 prevents the floor from rising above the detection threshold. The upper quantile ($Q_{0.90}$) anchors the onset criterion : choosing a quantile strictly below the cap (0.95) ensures that the threshold adapts to the participant's baseline in the vast majority of sessions, with the cap acting as a true safety ceiling only when the baseline is severely elevated.

This asymmetric design was validated by a grid search over $(q_{\text{low}}, q_{\text{high}}, \text{cap}_{\text{low}}, \text{cap}_{\text{high}})$ across all held-out participants. A swallow event is initiated when $\tilde{p}(t)$ exceeds τ_{high} and remains active until it falls below τ_{low}. Compared with single-threshold detection, hysteresis reduces false alarms and limits fragmentation near the decision boundary. The raw segments are refined through post-processing : segments shorter than 0.3 s are discarded, adjacent segments separated by less than 0.2 s are merged, and segments longer than 3.0 s are split at the minimum local score when a clear internal valley is present (Fig. 3.4).

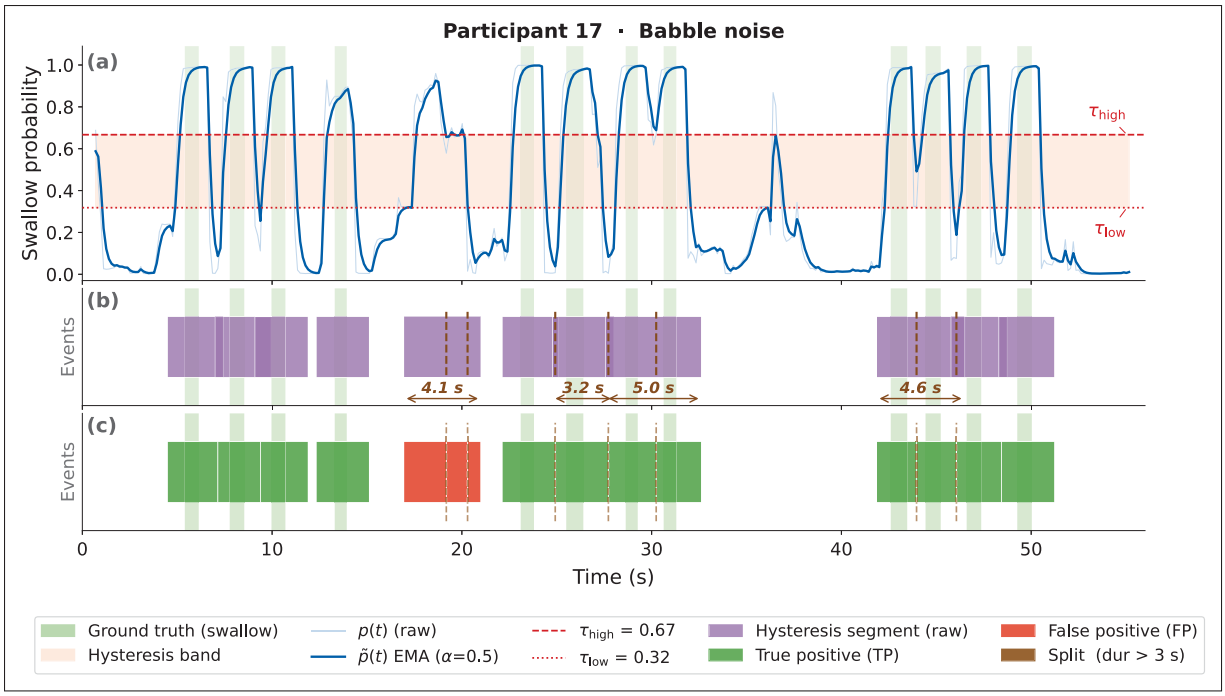


Figure 3.4 Swallow event detection pipeline for Participant 17 under babble noise. (a) Smoothed posterior $\tilde{p}(t)$ with adaptive hysteresis thresholds. (b) Hysteresis output before post-processing. (c) Final detected events after the splitting step

3.3.7 Evaluation protocol

To ensure subject-independent evaluation of the proposed method, participant-disjoint five-fold cross-validation was performed, with folds defined by participant ID. In each fold, approximately 80 % of participants (27/34) were used for training and 20 % (7/34) for testing. Within the training set, 15 % of participants (6/27) were further held out for validation. Training, validation,

and test sets were strictly disjoint at the participant level, ensuring that no recording from a test participant was seen during training or model selection.

Event-level evaluation uses an overlap-based criterion with a minimum IoU threshold of 30 %. A predicted segment is counted as a true positive if it overlaps at least one reference swallow event by this criterion. Predicted segments without sufficient overlap are counted as false positives, and unmatched reference events as false negatives. Event-level precision, recall, and F1-score are computed from these counts. The matching rule is illustrated in Fig. 3.5.

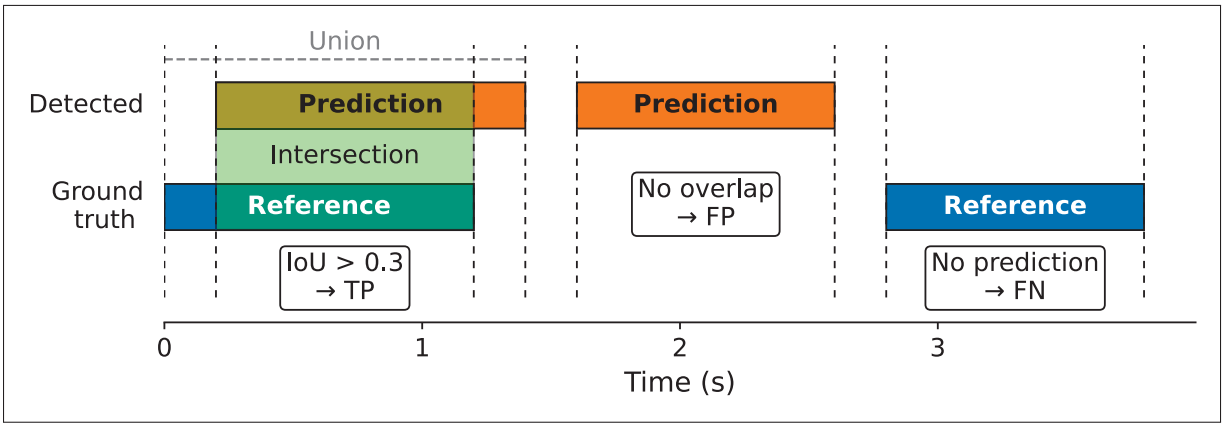


Figure 3.5 Illustration of the event-level IoU evaluation criterion

3.4 Results

3.4.1 Overall performance

The proposed dual-ear CRNN achieved consistent performance across all five participant-disjoint folds. At the window level, mean classification F1-score is 0.892 ± 0.006, with precision of 0.879 ± 0.005, recall of 0.906 ± 0.016, and AUROC of 0.914 ± 0.004. The low standard deviations confirm stable generalization to unseen participants. Recall exceeded precision, which constitutes a favorable operating point for longitudinal swallowing monitoring : missed swallows are embedded within long continuous recordings and cannot be recovered, whereas false detections are explicit, time-stamped, and can be reviewed or filtered post hoc.

At the event level, mean detection F1-score is 0.835 ± 0.012, with precision of 0.779 ± 0.014 and recall of 0.900 ± 0.020. Recall remains above 0.870 across all folds, with three folds exceeding 0.90, indicating that the system reliably captured the large majority of swallowing events regardless of participant identity. The elevated recall relative to precision reflects the intended behavior of the adaptive hysteresis detector, which favors detection coverage over false-alarm suppression.

Table 3.1 reports performance stratified by noise condition for both the classification and detection tasks. For classification, quiet conditions yield the highest F1-score (0.905 ± 0.002), followed by factory (0.893 ± 0.008) and babble (0.877 ± 0.010). The primary effect of ambient noise is a reduction in recall rather than an increase in false positives. For detection, F1-scores remain comparable across all three conditions ($\Delta \leq 0.021$), confirming robustness to diverse acoustic environments.

Tableau 3.1 Classification and detection performance by noise condition (mean ± SD over 5 folds)

Task	Condition	Precision	Recall	F1
Classif.	Babble	0.878±0.007	0.876±0.022	0.877±0.010
	Factory	0.880±0.004	0.907±0.019	0.893±0.008
	Quiet	0.878±0.010	**0.934±0.012**	**0.905±0.002**
Detect.	Babble	**0.797±0.019**	0.899±0.018	**0.845±0.011**
	Factory	0.773±0.013	0.882±0.024	0.824±0.014
	Quiet	0.769±0.011	**0.921±0.024**	0.838±0.011

Figure 3.6 illustrates the per-participant precision recall trade-off for event-level detection. The proposed system generalizes well across participants. For the vast majority of participants, the system achieves an F1-score above 0.80, and most often yields results in the high-recall region (recall $\geq$ 0.90), confirming that the model reliably detects swallowing events regardless of participant identity.

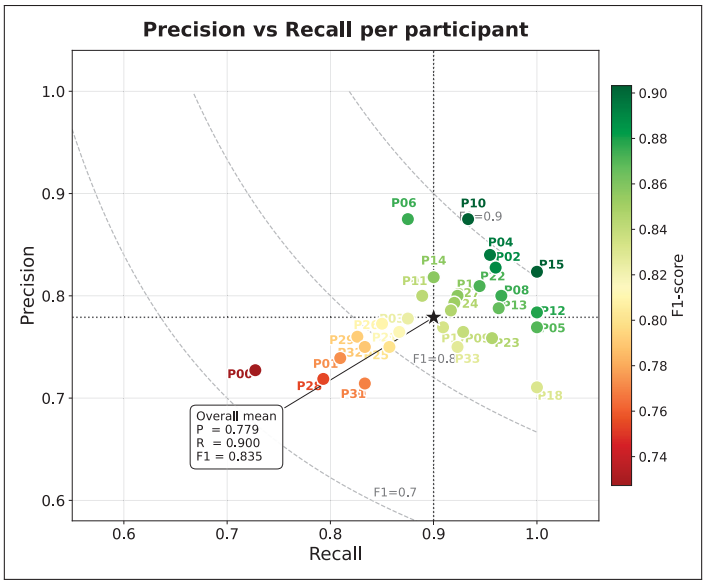


Figure 3.6 Per-participant precision vs. recall for event-level detection. Each marker represents one participant, coloured by F1-score (green : high, red : low). Dashed lines indicate overall mean precision and recall. The black star denotes the system-level mean results. Iso-F1 contours at 0.7, 0.8, and 0.9 are shown as dashed curves

3.4.2 Feature representation study

The choice of input representation has a pronounced impact on performance (Table 3.2). PCEN consistently outperforms log-Mel spectrograms across both the classification and detection tasks. In classification, PCEN improves recall by 12.6 percentage points (0.906 vs. 0.780) and F1-score by 8.1 points (0.892 vs. 0.811). The advantage is equally pronounced for detection, where PCEN achieves recall of 0.900 compared to 0.752 for log-Mel, yielding a detection F1-score of 0.835 versus 0.755. The considerably larger variance in recall scores associated with log-Mel features across folds particularly in recall (±0.057 for classification, ±0.043 for detection) further underscores its sensitivity to inter-participant variability, whereas PCEN produces substantially more stable estimates.

Tableau 3.2 Feature comparison (mean ± SD over 5 folds). Best values in bold

Feature	Task	Precision	Recall	F1
Log-Mel	Classif.	0.843±0.010	0.780±0.057	0.811±0.030
	Detect.	0.759±0.009	0.752±0.043	0.755±0.026
PCEN	Classif.	**0.879±0.005**	**0.906±0.016**	**0.892±0.006**
	Detect.	**0.779±0.014**	**0.900±0.020**	**0.835±0.012**

3.4.3 Architecture comparison

The choice of the BiLSTM-based temporal backbone described in Section 3.3.4 is grounded in a comparative study of four candidate architectures : a unidirectional LSTM, a BiLSTM, a BiGRU, and a Transformer encoder. The results of this comparison are presented in Table 3.3. The BiLSTM achieves the highest F1-score in both classification (0.892 ± 0.006) and detection (0.835 ± 0.012). The unidirectional LSTM shows markedly lower recall in both tasks (detection F1 = 0.767), demonstrating that access to future context within the analysis window is critical for reliable offline detection. The Transformer achieves the highest detection precision (0.785) but yields one of the lowest recall scores among the compared architectures (0.779) and exhibits the largest variance in recall (±0.149), indicating instability under cross-participant evaluation, likely attributable to the limited dataset size relative to the capacity of global self-attention. The BiGRU represents a competitive but consistently inferior alternative (detection F1 = 0.815).

3.4.4 Ablation study

Table 3.4 isolates the contribution of individual architectural components within the proposed method. Removing the BiLSTM yields the most severe degradation (detection F1 drops from 0.835 to 0.748), confirming the necessity of recurrent modeling for capturing swallowing temporal dynamics. Removing attention pooling reduces detection F1 to 0.812 and classification F1 from 0.892 to 0.825, confirming that selective temporal weighting is essential for both discrimination and localization. Removing the CNN front-end has a smaller but measurable impact (classification F1 drops to 0.847), confirming that convolutional features complement

Tableau 3.3 Architecture comparison (mean ± SD over 5 folds). Best values in bold

Backbone	Task	Precision	Recall	F1
LSTM	Classif.	0.865±0.025	0.785±0.070	0.823±0.028
	Detect.	0.755±0.009	0.779±0.082	0.767±0.039
BiLSTM	Classif.	**0.879±0.005**	**0.906±0.016**	**0.892±0.006**
	Detect.	0.779±0.014	**0.900±0.020**	**0.835±0.012**
BiGRU	Classif.	0.872±0.013	0.802±0.037	0.835±0.015
	Detect.	0.766±0.018	0.872±0.041	0.815±0.019
Transformer	Classif.	0.868±0.013	0.790±0.045	0.827±0.022
	Detect.	**0.785±0.038**	0.779±0.149	0.782±0.077

Tableau 3.4 Ablation study (mean ± SD over 5 folds). Best values in bold

Configuration	Task	Precision	Recall	F1
Full model	Classif.	**0.879±0.005**	**0.906±0.016**	**0.892±0.006**
	Detect.	0.779±0.014	**0.900±0.020**	**0.835±0.012**
Without attention	Classif.	0.852±0.013	0.801±0.042	0.825±0.018
	Detect.	0.771±0.011	0.859±0.059	0.812±0.022
Without BiLSTM	Classif.	0.861±0.016	0.789±0.053	0.823±0.025
	Detect.	0.723±0.006	0.776±0.039	0.748±0.021
Without CNN	Classif.	0.868±0.014	0.828±0.039	0.847±0.016
	Detect.	**0.782±0.015**	0.888±0.032	0.831±0.019

recurrent temporal modeling. The full architecture achieves the best performance across all metrics, confirming that each component provides a non-redundant contribution.

3.4.5 Fusion study

Table 3.5 demonstrates that bilateral ear processing is the dominant factor in achieving high detection recall. Single-ear configurations yield F1-scores of 0.781 (right) and 0.811 (left), compared to 0.835 for the full dual-ear system. The recall gain of the dual-ear system (0.900)

over both monaural baselines (0.795 right, 0.863 left) indicates that the two ears capture complementary swallowing evidence that is inconsistently lateralized. Fixed fusion strategies mean (0.796) and max (0.805) consistently underperform compared to the learned adaptive fusion, confirming that symmetric or selection-based rules are insufficient to exploit the time-varying interaural asymmetry documented in Section 3.3.2.

Tableau 3.5 Fusion study for event-level detection (mean ± SD over 5 folds). Best values in bold

Configuration	Precision	Recall	F1
Right-ear only	0.770 ±0.016	0.795 ±0.049	0.781±0.023
Left-ear only	0.763 ±0.006	0.863 ±0.044	0.811 ±0.017
Mean fusion	0.781 ±0.015	0.811 ±0.034	0.796 ±0.021
Max fusion	0.769 ±0.026	0.845 ±0.037	0.805 ±0.026
Proposed fusion	**0.779 ±0.014**	**0.900 ±0.020**	**0.835 ±0.012**

3.4.6 Bolus type analysis

Event-level detection was also evaluated per bolus type. Water swallows achieve the highest recall (0.968±0.015) and F1-score (0.861±0.015), reflecting the acoustically consistent nature of liquid bolus transit as shown in Table 3.6. Yogurt swallows yield the highest precision (0.806 ± 0.024), suggesting that the semi-solid bolus produces distinctive in-ear signatures that limit false positives, albeit with slightly lower recall. Dry swallows exhibit the lowest F1-score (0.801 ± 0.009) and the largest precision recall asymmetry, consistent with their greater duration variability. These findings align with prior reports that bolus viscosity significantly influences swallowing acoustics (Asakura & Miwa, 2025). The F1-score range across bolus types ($\Delta = 0.060$) remains narrow, confirming that the PCEN-based representation and participant-adaptive thresholding generalize across bolus conditions.

Tableau 3.6 Detection performance across bolus types (mean ± standard deviation over 5 folds). Best values in bold

Bolus type	Precision	Recall	F1-score
Dry	0.766 ± 0.017	0.839 ± 0.026	0.801 ± 0.009
Water	0.776 ± 0.021	**0.968 ± 0.015**	**0.861 ± 0.015**
Yogurt	**0.806 ± 0.024**	0.906 ± 0.036	0.852 ± 0.021

3.5 Discussion

Compared to the transfer-learning baseline established in our prior work (Cheikh *et al.*, 2026), which fine-tuned YAMNet on the same corpus, the proposed CRNN achieves comparable window-level classification performance (F1-score 0.892 vs. 0.875). The primary advantage of the present system lies at the event level : the dual-ear CRNN with adaptive post-processing achieves a detection F1-score of 0.835, substantially outperforming the 0.78 obtained with YAMNet. This gap highlights that strong frame-level discrimination does not automatically translate into reliable event-level detection, and that the temporal modeling provided by the BiLSTM and the adaptive hysteresis post-processing are key to bridging this gap. To better understand the contribution of each channel to this overall performance, we further analyze single-channel results. The fusion study, presented in Table 3.5, shows that on average, for our dataset, using the left ear signals alone outperforms use of the right ear signals in single-channel detection (F1-score = 0.811 vs. 0.781), with a particularly marked recall advantage (0.863 vs. 0.795). This performance gap is a direct consequence of the occlusion effect. When an earbud seals the ear canal, internally generated low-frequency body sounds are amplified within the closed cavity, while environmental noise is attenuated, raising the SNR for swallowing-related signals. A tighter seal therefore allows for a more detectable swallowing signature, not only because the swallow sounds are amplified, but because the residual noise in the ear is lower.

The seal study in Section 3.3.2 provides the empirical link between this physical mechanism and the detection asymmetry : the left channel exhibited lower IEM RMS amplitudes ($\mathrm{RMS}_L \approx 0.005$–$0.008$ vs. $\mathrm{RMS}_R \approx 0.007$–$0.011$) and a mean attenuation advantage of +0.66 dB at 250 Hz,

confirming that the left earpiece achieved tighter occlusion on average. A better-sealed ear suppresses more ambient noise, preserves more of the low-frequency swallowing energy, and consequently yields higher classifier confidence, which translates directly into the higher recall observed for the left-only configuration.

Dry swallows are associated with the lowest recall and F1-score among bolus types. This may be explained by physiological factors. As detailed in (Cheikh *et al.*, 2026), the repeated dry swallow task progressively dries the oral cavity, and the fourth and fifth swallows in each window tend to show longer durations and more fragmented acoustic patterns that diverge from the normal swallow signature the model was trained on. Such task-induced modulation has been previously reported in healthy individuals (Nagy *et al.*, 2013).

The detector's high-recall operating point (0.900 vs. precision 0.779) reflects an explicit clinical priority. This distinction is reinforced from a practical standpoint : false positives correspond to explicit detections with known temporal locations and can therefore be reviewed or filtered. In contrast, missed events remain embedded within long continuous recordings without temporal markers, making retrospective identification impractical. As a result, false negatives constitute an irrecoverable source of error. Furthermore, given that spontaneous swallowing occurs frequently throughout the day even in the absence of food intake, a reduced recall would lead to several hundred missed events per day, resulting in an important underestimation of the frequency of daily swallowing.

Several limitations should be acknowledged, which we group into limitations of the proposed method and limitations of the experimental study. The first category concerns the proposed method itself. The detector operates offline rather than in real time, as a direct consequence of the bidirectional recurrence in the BiLSTM encoder and of the post-processing stage, which both require access to future windows to produce a decision. However, this is not a critical limitation in our target context, as immediate feedback is not the primary objective : in applications such as dietary monitoring, clinical assessment, or longitudinal behavior analysis, data are summarized and analyzed retrospectively by the patient or clinician without requiring instant detection. A

real-time adaptation, replacing the BiLSTM with a unidirectional LSTM and reformulating the post-processing to operate without future scores, is technically feasible, but the unidirectional LSTM already shows markedly lower recall (F1-score = 0.767 vs. 0.835), and this gap would likely widen in less controlled conditions. A second methodological consideration is that the detection threshold calibration relies on a short swallow-free baseline segment recorded at the start of each session. We view this requirement as a practical feature rather than a limitation : participant and session-adaptive calibration is what allows the detector to absorb the inter-session variability in seal quality documented in Section 3.3.2, and in a real-world hearable deployment, the calibration window would naturally coincide with the first minute following device insertion, before any intentional swallow or intake. If enforcing a strictly silent baseline becomes impractical in unconstrained ambulatory monitoring, an automated non-swallow segment identification procedure could be integrated upstream of the detector without modifying its core design.

The second category concerns the experimental study. First, each participant was evaluated in a single session, and no test-retest reliability analysis was conducted across days or weeks. Prior work on oropharyngeal swallowing has shown that no statistically significant durational differences emerge across repeated trials and testing dates in healthy adults (Lof & Robbins, 1990), suggesting that detection performance would likely remain stable across sessions for individuals within the same age range and health status (Cheikh *et al.*, 2026). Nevertheless, explicit cross-session validation remains an important direction for further validation, particularly for longitudinal monitoring. Second, the cohort in this study is exclusively composed of healthy adults aged 20–29, and the acoustic manifestations of dysphagic or parkinsonian swallowing in the ear canal remain entirely uncharacterized, a gap that cannot be bridged by data augmentation or architecture changes, but only through new clinical data collection. Third, confounders were elicited sequentially rather than spontaneously, and compound events (e.g., eating while walking) could produce false detections in free-living conditions that are not fully captured by the present evaluation protocol. A further consideration concerns hardware generalizability. The earpieces used here were fitted with industrial-grade foam eartips available in three sizes (small, medium, large), providing consistent tight occlusion that is central to the system's operating principle.

Consumer-grade hearables rely on silicone tips with unreliable sealing and a generally lower attenuation. The loss or degradation of the acoustic seal has a twofold impact. First, it reduces the occlusion effect, so the physiological signals of interest captured by the IEM become weaker. Second, it decreases passive attenuation of external sounds, thereby degrading the signal-to-noise ratio in everyday environments. In practice, this means that models trained on tightly occluded recordings are more likely to transfer to hearables operating under active noise cancellation, where a stable seal is maintained. By contrast, performance may degrade in transparency or awareness modes, which deliberately reintroduce ambient sound and effectively reduce the benefits of occlusion.

3.6 Conclusions

This paper presented, to the best of our knowledge, the first offline framework for swallow event detection from bilateral in-ear audio based on learned acoustic representations. The proposed system combines PCEN-based feature extraction, a dual-ear CRNN with shared encoding, bidirectional temporal modeling, attention pooling, and adaptive confidence-based fusion, followed by a hysteresis detector calibrated from a short 45 s baseline breathing segment. Evaluated on 34 participants under quiet, factory, and babble noise conditions using subject-disjoint five-fold cross-validation, the system achieved a window-level classification F1-score of 0.892 ± 0.006 and an event-level detection F1-score of 0.835 ± 0.012, with recall consistently exceeding 0.870 across all folds. Ablation studies confirm the non-redundant contributions of each architectural component, while the fusion study demonstrates that adaptive bilateral integration outperforms both single-ear and fixed-fusion baselines, consistent with the time-varying interaural seal asymmetry observed across participants. The system's high-recall operating point is well suited to longitudinal swallowing monitoring, where missed events constitute an irrecoverable source of error in long continuous recordings. These results suggest that in-ear audio captured under occlusion provides a viable and unobtrusive modality for swallowing assessment, with potential deployment in hearable devices for monitoring swallowing behavior across diverse acoustic environments. Future work will focus on validating the system in clinical populations

presenting dysphagia or sialorrhea, extending evaluation to unconstrained daily recordings, and investigating causal reformulations for real-time deployment.

CHAPITRE 4

CONCLUSIONS ET PERSPECTIVES

4.1 Contributions et retour sur les objectifs de recherche

Ce mémoire visait quatre sous-objectifs complémentaires, dont l'atteinte constitue autant de contributions scientifiques.

1. **Créer et valider une base de données multimodale :** La base de données collectée auprès de 34 adultes sains, couvrant trois types de bolus, six catégories de confondants et trois conditions acoustiques, constitue la première ressource publique d'audio intra-auriculaire bilatéral annoté pour la déglutition, avec une vérité terrain fondée sur l'imagerie échographique linguale. Elle répond au manque structurel identifié en introduction et fournit un protocole reproductible permettant la comparaison de méthodes futures.
2. **Établir une preuve de concept pour la classification de la déglutition :** Un classifieur FCNN entraîné sur des représentations YAMNet complétées par le ZCR atteint un score F1 de 0,875 ± 0,013 en validation croisée à cinq plis disjoints au niveau sujet, confirmant que l'audio intra-auriculaire porte une information suffisante pour discriminer la déglutition des autres événements physiologiques.
3. **Développer un algorithme de détection bilatérale :** Le premier cadre de détection de bout en bout de la déglutition à partir d'audio intra-auriculaire est proposé. Il associe une représentation PCEN, une architecture CRNN à double encodeur partagé avec modélisation temporelle bidirectionnelle et agrégation par attention, un module de fusion adaptative par confiance apprise entre les deux oreilles, et un détecteur à hystérésis calibré par participant. La fusion adaptative est motivée par la mise en évidence empirique que la dominance acoustique relative des deux canaux IEM est une propriété variable dans le temps, justifiant le rejet des stratégies fixes.
4. **Évaluer le système proposé :** L'évaluation sur les 34 participants donne un score F1 de classification de 0,892 ± 0,006 et un score F1 de détection de 0,835 ± 0,012, avec un rappel maintenu au-dessus de 0,870 dans toutes les conditions acoustiques. Les études d'ablation confirment la contribution non redondante de chaque composant. La comparaison des

stratégies de fusion démontre la supériorité de la fusion adaptative bilatérale, avec un gain de rappel de 3,7 points par rapport à la meilleure configuration monoaurale. La représentation PCEN surpasse les spectrogrammes Mel logarithmiques en rappel (0,906 contre 0,780) et en stabilité inter-participants.

4.2 Limites de l'étude

Malgré les performances encourageantes obtenues, plusieurs limites méthodologiques et opérationnelles doivent être reconnues afin de situer correctement la portée des résultats.

4.2.1 Taille et diversité du jeu de données

La cohorte étudiée comprend 34 adultes sains âgés de 20 à 29 ans, recrutés dans un environnement académique contrôlé. Si cet effectif est suffisant pour établir une preuve de concept robuste, confirmée par la faiblesse des écarts-types inter-plis, il demeure restreint au regard des exigences d'un déploiement clinique généralisable. Plusieurs dimensions de la diversité restent ainsi non couvertes.

Homogénéité démographique : Le groupe n'inclut que des jeunes adultes en bonne santé, excluant a fortiori les populations âgées, dont la physiologie de la déglutition se distingue par des durées accrues, une coordination pharyngée altérée et une musculature laryngée affaiblie. Or, ce sont précisément ces populations qui présentent le risque le plus élevé de dysphagie. Les signatures acoustiques de leurs déglutitions dans le canal auriculaire restent entièrement non caractérisées, un écart que ni l'augmentation de données ni des modifications architecturales ne peuvent combler en l'absence de données cliniques.

Séance unique par participant : Chaque participant n'a été enregistré que lors d'une seule séance. L'absence d'analyse inter-séances ne permet pas de quantifier la variabilité intra-individuelle à long terme, pourtant critique pour un dispositif de suivi longitudinal. Si la littérature rapporte une stabilité des durées de déglutition sur des séances répétées chez des

adultes sains de profil comparable, cette hypothèse n'a pas été vérifiée empiriquement dans le cadre de ce travail.

Matériel spécifique : Les enregistrements ont été acquis avec des écouteurs équipés d'embouts en mousse de qualité industrielle disponibles en trois tailles (petit, moyen, grand), assurant une occlusion étanche et reproductible qui constitue le principe central de fonctionnement du système. Les écouteurs grand public reposent généralement sur des embouts en silicone dont l'étanchéité est plus variable et l'atténuation passive globalement plus faible. La perte ou la dégradation du sceau acoustique entraîne deux conséquences. D'une part, l'effet d'occlusion s'atténue, et avec lui l'amplification sélective des sons conduits par voie osseuse. D'autre part, l'isolation passive aux sons externes diminue, ce qui dégrade le rapport signal-sur-bruit dans les environnements quotidiens. En pratique, les modèles entraînés sur des enregistrements à occlusion étanche sont plus susceptibles de transférer vers des écouteurs grand public fonctionnant sous réduction active du bruit. À l'inverse, les performances pourraient se dégrader en mode transparence ou en mode conscience, qui réintroduisent délibérément le son ambiant et annulent en partie les bénéfices de l'occlusion.

4.2.2 Contraintes d'annotation et de conception du protocole

Le protocole d'acquisition repose sur une élicitation séquentielle et structurée des événements : les déglutitions, les bruits articulatoires et les mouvements physiques ont été sollicités dans un ordre défini, dans un environnement audiométrique contrôlé. Cette approche présente plusieurs limites.

Congruence limitée avec les conditions naturelles : Les événements concurrents ayant été élicités de manière isolée et consécutive, le modèle n'a pas été exposé aux combinaisons d'événements simultanés caractéristiques de la vie quotidienne, telles qu'une conversation tenue pendant un repas, une déglutition immédiatement consécutive à une bouchée mâchée ou l'ingestion d'une boisson en marchant. La présence d'événements composites pourrait générer des faux positifs dans la vie quotidienne, non reproduits lors de l'entraînement.

Fiabilité de la vérité terrain : La vérité terrain repose sur l'accord entre l'audio bilatéral et l'imagerie ultrasonore acquise à 30 images par seconde, soit une résolution temporelle d'environ 33 ms. Comme discuté au chapitre 2, cette discrétisation peut induire un léger décalage en avance par rapport à l'instant acoustique réel, indépendamment du soin apporté à l'étiquetage. La règle d'étiquetage par recouvrement minimum à 20 % avec la fenêtre glissante, propage cette ambiguïté aux fenêtres de transition et peut pénaliser légèrement la précision événementielle. À cette imprécision temporelle s'ajoute le fait que l'ensemble des annotations a été produit par un seul opérateur, ce qui n'a pas permis d'estimer la fiabilité de l'étiquetage. Une double annotation indépendante sur un sous-ensemble représentatif des enregistrements et le calcul d'un coefficient kappa de Cohen, permettraient de quantifier ce biais potentiel et de consolider la fiabilité de l'évaluation.

Calibration du seuil de référence : Le détecteur à hystérésis adaptif requiert un segment de référence non-déglutition d'au moins 45 secondes au début de chaque séance. Cette hypothèse, vérifiable dans un protocole experimental, peut être difficile à garantir dans un contexte ambulatoire non supervisé. Sa violation entraînerait une estimation erronée des seuils adaptatifs τ_{low} et τ_{high}, dégradant potentiellement la précision événementielle de façon significative. Une procédure de calibration, capable d'estimer le seuil de référence à partir des segments les plus calmes d'une fenêtre temporelle glissante quel que soit le moment d'activation du dispositif, devra être développée pour un usage en vie quotidienne.

4.3 Pistes de recherche futures

Les limitations identifiées tracent naturellement les axes de développement prioritaires pour les travaux futurs.

4.3.1 Extension et enrichissement de la base de données

La consolidation clinique du système passe en premier lieu par une diversification du corpus. Trois directions sont prioritaires.

Validation sur des populations cliniques : La transposition du système à des participants présentant une dysphagie avérée ou une sialorrhée constitue l'étape la plus structurante. Au-delà de la simple acquisition de nouvelles données, il s'agira d'évaluer plusieurs scénarios de transfert : application directe du modèle pré-entraîné ou un fine-tuning sur un sous-ensemble clinique. Cette étape devra s'accompagner d'une validation avec un examen de référence.

Élargissement démographique et suivi longitudinal : La diversification des profils inclus, en particulier l'inclusion de participants âgés et de populations pédiatriques, devra s'accompagner d'un protocole multi-séances réparti sur plusieurs semaines ou mois afin de caractériser la dérive éventuelle des seuils adaptatifs et la variabilité intra-individuelle à long terme. Un tel protocole permettrait également de tester la robustesse du module de fusion bilatérale face aux variations causées par le vieillissement, la progression de la maladie, les prises de médicaments, la croissance et le positionnement de l'écouteur entre séances.

Enregistrements dans un contexte non contrôlé : Le passage d'un protocole structuré à des enregistrements non supervisés en environnement quotidien implique de gérer des durées d'acquisition bien plus longues, une plus grande variabilité acoustique et l'absence de vérité terrain précise. Ce contexte exige de développer des stratégies d'annotation semi-supervisée ou d'apprentissage actif pour étiqueter efficacement ces données.

4.3.2 Extension à d'autres événements non verbaux

Le cadre proposé se prête naturellement à l'extension vers d'autres classes d'événements non verbaux d'intérêt clinique captables par voie intra-auriculaire. La toux, le ronflement, et le hoquet présentent des caractéristiques spectro-temporelles distinctes susceptibles d'être discriminées par un classifieur partagé. Par ailleurs, la détection combinée de la mastication et de la déglutition permettrait d'inférer des épisodes alimentaires complets, ouvrant la voie à un suivi nutritionnel passif et continu, pertinent pour les patients en réhabilitation ou en soins gériatriques.

4.3.3 Adaptation à l'opération en temps réel

Le système actuel est conçu pour une analyse hors ligne, s'appuyant sur un BiLSTM bidirectionnel qui exploite le contexte futur de chaque fenêtre d'analyse. Bien que cette contrainte ne soit pas critique pour les applications cibles de surveillance longitudinale ou de bilan clinique, une reformulation causale est souhaitable pour des scénarios de biofeedback en temps réel ou d'alerte clinique immédiate.

La substitution du BiLSTM par un LSTM unidirectionnel constitue la voie la plus directe, mais entraîne une dégradation notable du rappel (0,767 contre 0,835) documentée dans l'étude comparative d'architectures. Plusieurs stratégies permettraient d'atténuer cette perte : l'utilisation d'un contexte futur limité (*lookahead*) compatible avec une latence acceptable ou l'intégration d'une mémoire récurrente à état persistant (Gated Recurrent Unit (GRU) ou LSTM avec cache glissant). La reformulation du post-traitement par hystérésis en mode en ligne, sans accès aux scores futurs, devra également être validée.

BIBLIOGRAPHIE

Alagiakrishnan, K., Bhanji, R. A. & Kurian, M. (2013). Evaluation and management of oropharyngeal dysphagia in different types of dementia : a systematic review. *Archives of Gerontology and Geriatrics*, 56(1), 1–9. doi : 10.1016/j.archger.2012.04.011.

Altman, K. W., Yu, G.-P. & Schaefer, S. D. (2010). Consequence of dysphagia in the hospitalized patient : impact on prognosis and hospital resources. *Archives of Otolaryngology–Head & Neck Surgery*, 136(8), 784–789. doi : 10.1001/archoto.2010.129.

Asakura, T. & Miwa, M. (2025). Effect of age and bolus characteristics on the acoustic parameters of ear and neck swallowing sounds. *Sensors and Actuators A : Physical*, 383, 116169. doi : 10.1016/j.sna.2024.116169.

Ashiga, H., Takei, E., Magara, J., Takeishi, R., Tsujimura, T., Nagoya, K. & Inoue, M. (2019). Effect of attention on chewing and swallowing behaviors in healthy humans. *Scientific Reports*, 9, 6013. doi : 10.1038/s41598-019-42422-4.

Bedri, A., Li, R., Haynes, M., Kosaraju, R. P., Grover, I., Prioleau, T., Beh, M. Y., Goel, M., Starner, T. & Abowd, G. (2017). EarBit : Using Wearable Sensors to Detect Eating Episodes in Unconstrained Environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 1(3), 1–20.

Benacchio, S., Doutres, O., Le Troter, A., Varoquaux, A., Wagnac, E., Callot, V. & Sgard, F. (2018). Estimation of the ear canal displacement field due to in-ear device insertion. *Hearing Research*, 365, 16–27. doi : 10.1016/j.heares.2018.04.011.

Bernier, A. & Voix, J. (2013). An active hearing protection device for musicians. *Proceedings of Meetings on Acoustics (ICA 2013, Montreal)*, 19, 040015. doi : 10.1121/1.4800066.

Bi, S., Wang, T., Tobias, N., Nordrum, J., Wang, S., Halvorsen, G., Sen, S., Peterson, R., Odame, K., Caine, K., Halter, R., Sorber, J. & Kotz, D. (2018). Auracle : Detecting Eating Episodes with an Ear-Mounted Sensor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3), 1–27. doi : 10.1145/3264902.

Bouserhal, R. E., Bernier, A. & Voix, J. (2019). An in-ear speech database in varying conditions of the audio-phonation loop. *The Journal of the Acoustical Society of America*, 145(2), 1069–1077. doi : 10.1121/1.5091777.

Branda, E. (2012). Deep canal fittings : advantages, challenges, and a new approach. *Hearing Review*, 19, 24–27.

Brummund, M. K., Sgard, F., Petit, Y. & Laville, F. (2014). Three-dimensional finite element modeling of the human external ear : Simulation study of the bone conduction occlusion effect. *The Journal of the Acoustical Society of America*, 135(3), 1433–1444. doi : 10.1121/1.4864484.

Bulmer, J. M., Ewers, C., Drinnan, M. J. & Ewan, V. C. (2021). Evaluation of Spontaneous Swallow Frequency in Healthy People and Those With, or at Risk of Developing, Dysphagia : A Review. *Gerontology and Geriatric Medicine*, 7, 1–13. doi : 10.1177/23337214211041801.

Butkow, K.-J., Dang, T., Ferlini, A., Ma, D. & Mascolo, C. (2021). hEARt : Motion-resilient Heart Rate Monitoring with In-ear Microphones.

Calcagno, P., Ruoppolo, G., Grasso, M. G., Vincentiis, M. D. & Paolucci, S. (2002). Dysphagia in multiple sclerosis – prevalence and prognostic factors. *Acta Neurologica Scandinavica*, 105(1), 40–43. doi : 10.1034/j.1600-0404.2002.10062.x.

Castro, M., Dedivitis, R., Matos, L., Baraúna, J., Kowalski, L., Moura, K. & Partezani, D. (2025). Endoscopic and videofluoroscopic evaluations of swallowing for dysphagia : A systematic review. *Brazilian Journal of Otorhinolaryngology*, 91, 101598. doi : 10.1016/j.bjorl.2025.101598.

Chabot, P., Bouserhal, R. E., Cardinal, P. & Voix, J. (2021). Detection and classification of human-produced nonverbal audio events. *Applied Acoustics*, 171, 107643.

Cheikh, E. B., Mrabet, Y., Laporte, C. & Bouserhal, R. E. (2026). A multimodal in-ear audio and physiological dataset for swallowing and non-verbal event classification. *Sensors*, 26(7), 2019. doi : 10.3390/s26072019.

Chen, X. & Kamavuako, E. N. (2024). Enhancing Intake Monitoring : Transfer Learning for Audio-Based Detection of Swallowing Events. *IEEE MELECON*.

Chou, K. L., Evatt, M., Hinson, V. & Kompoliti, K. (2007). Sialorrhea in Parkinson's disease : A review. *Movement Disorders*, 22(16), 2306–2313. doi : 10.1002/mds.21646.

Cichero, J. A. Y. & Murdoch, B. E. (2002). Acoustic signature of the normal swallow : Characterization by age, gender, and bolus volume. *Annals of Otology, Rhinology & Laryngology*, 111(7), 623–632.

Crary, M. A., Carnaby, G. D. & Sia, I. (2014). Spontaneous Swallow Frequency Compared with Clinical Screening in the Identification of Dysphagia in Acute Stroke. *Journal of Stroke and Cerebrovascular Diseases*, 23(8), 2047–2053. doi : 10.1016/j.jstrokecerebrovasdis.2014.03.008.

Cuevas, J. L., III, E. W. C., Richter, J. E., McCutcheon, M. & Taub, E. (1995). Spontaneous swallowing rate and emotional state. *Digestive Diseases and Sciences*, 40(2), 282–286. doi : 10.1007/BF02065410.

Dean, M. S. & Martin, F. N. (2000). Insert earphone depth and the occlusion effect. *American Journal of Audiology*.

Dodds, W. J., Stewart, E. T. & Logemann, J. A. (1990). Physiology and radiology of the normal oral and pharyngeal phases of swallowing. *American Journal of Roentgenology*, 154(5), 953–963. doi : 10.2214/ajr.154.5.2108569.

Dudik, J. M., Jestrović, I., Luan, B., Coyle, J. L. & Sejdić, E. (2015). A comparative analysis of swallowing accelerometry and sounds during saliva swallows. *BioMedical Engineering OnLine*, 14(3).

Firmin, H., Reilly, S. & Fourcin, A. (1997). Non-invasive monitoring of reflexive swallowing. *Speech, Hearing and Language : Work in Progress*.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M. & Ritter, M. (2017). Audio Set : An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780. doi : 10.1109/ICASSP.2017.7952261.

González-Fernández, M., Ottenstein, L., Atanelov, L. & Christian, A. B. (2013). Dysphagia after Stroke : an Overview. *Current Physical Medicine and Rehabilitation Reports*, 1(3), 187–196. doi : 10.1007/s40141-013-0017-y.

Goverdovsky, V., von Rosenberg, W., Nakamura, T., Looney, D., Sharp, D. J., Papavassiliou, C., Morrell, M. J. & Mandic, D. P. (2017). Hearables : Multimodal physiological in-ear sensing. *Scientific Reports*, 7(6948).

Gravellier, L., Le Coz, M., Farinas, J. & Pinquier, J. (2023). Détection automatique de la déglutition dans les signaux d'auscultation cervicale à haute résolution. *XXIXème Colloque Francophone de Traitement du Signal et des Images (GRETSI)*.

Gravellier, L., Coz, M. L., Farinas, J. & Pinquier, J. (2024). Detection of Pharyngolaryngeal Activities in Real-World Settings Using Wearable Sensors. *46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–5. doi : 10.1109/EMBC53108.2024.10782007.

Hadjileontiadis, L. J. (2005). Wavelet-based enhancement of lung and bowel sounds using fractal dimension thresholding – Part I : Methodology. *IEEE Transactions on Biomedical Engineering*, 52(6), 1143–1148.

Haji, T. & Yamaguchi, Y. (2025). Intra-aural swallowing sound analysis with simultaneous videofluoroscopy and cervical swallowing sound recording. *Auris Nasus Larynx*, 52, 20–26. doi : 10.1016/j.anl.2024.11.003.

Hamlet, S. L., Patterson, R. L., Fleming, S. M. & Jones, L. A. (1992). Sounds of Swallowing Following Total Laryngectomy. *Dysphagia*, 7, 160–165.

Hauret, J., Joubaud, T., Zimpfer, V. & Éric Bavu. (2023). Configurable EBEN : Extreme Bandwidth Extension Network to Enhance Body-Conducted Speech Capture. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 3499–3512. doi : 10.1109/TASLP.2023.3313433.

Hauret, J., Olivier, M., Joubaud, T., Langrenne, C., Poirée, S., Zimpfer, V. & Éric Bavu. (2024). Vibravox : A Dataset of French Speech Captured with Body-conduction Audio Sensors.

Heikenfeld, J., Jajack, A., Rogers, J., Gutruf, P., Tian, L., Pan, T., Li, R., Khine, M., Kim, J., Wang, J. & Kim, J.-U. (2018). Wearable sensors : modalities, challenges, and prospects. *Lab on a Chip*, 18(2), 217–248. doi : 10.1039/C7LC00914C.

Hiramatsu, T., Kataoka, H., Osaki, M. & Hagino, H. (2015). Effect of aging on oral and swallowing function after meal consumption. *Clinical Interventions in Aging*, 10, 229–235. doi : 10.2147/CIA.S75211.

Honda, T., Baba, T., Fujimoto, K., Goto, T., Nagao, K., Harada, M., Honda, E. & Ichikawa, T. (2016). Characterization of Swallowing Sound : Preliminary Investigation of Normal Subjects. *PLoS ONE*, 11(12), e0168187.

Hughes, T. A. T. & Wiles, C. M. (1996). Clinical measurement of swallowing in health and in neurogenic dysphagia. *QJM : An International Journal of Medicine*, 89(2), 109–116. doi : 10.1093/qjmed/89.2.109.

Humbert, I. A. & Robbins, J. (2008). Dysphagia in the Elderly. *Physical Medicine and Rehabilitation Clinics of North America*, 19(4), 853–866. doi : 10.1016/j.pmr.2008.06.002.

Jayatilake, D., Teramoto, Y., Ueno, T., Matsuo, K. & Suzuki, K. (2026). Validation of automated 5 mL thin liquid swallowing sound segmentation for estimating audio-derived pharyngeal clearance time. *Scientific Reports*, 16, 11908. doi : 10.1038/s41598-026-39699-7.

Jean, A. (2001). Brain Stem Control of Swallowing : Neuronal Network and Cellular Mechanisms. *Physiological Reviews*, 81(2), 929–969. doi : 10.1152/physrev.2001.81.2.929.

Jestrović, I., Dudik, J. M., Luan, B., Coyle, J. L. & Sejdić, E. (2013). The effects of increased fluid viscosity on swallowing sounds in healthy adults. *BioMedical Engineering OnLine*, 12(90).

Kang, Y. J. (2024). Biometric Vibration Signal Detection Devices for Swallowing Activity Monitoring. *Signals*, 5(3), 516–525. doi : 10.3390/signals5030028.

Kapila, Y. V., Dodds, W. J., Helm, W. J. & J., W. (1984). Relationship between swallow rate and salivary flow. *Digestive Diseases and Sciences*, 29(6), 528–533. doi : 10.1007/BF01296273.

Khalifa, Y., Coyle, J. L. & Sejdić, E. (2020). Non-invasive identification of swallows via deep learning in high resolution cervical auscultation recordings. *Scientific Reports*, 10(1), 8704.

Khlaifi, H., Badii, A., Istrate, D. & Demongeot, J. (2019). Automatic Detection and Recognition of Swallowing Sounds. *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC)*.

Kimura, S., Emoto, T., Suzuki, Y., Shinkai, M., Shibagaki, A. & Shichijo, F. (2024). Novel Approach Combining Shallow Learning and Ensemble Learning for the Automated Detection of Swallowing Sounds in a Clinical Database. *Sensors*, 24(10), 3057. doi : 10.3390/s24103057.

Konoike, Y., Naramura, K., Hasegawa, T., Tsukayama, I., Maruoka, S., Kawakami, T., Ishii, K., Yoshida, T., Hokari, M. & Suzuki-Yamamoto, T. (2025). Parameter analysis using swallowing sounds shows differences in bolus volume, bolus viscosity, sex, and age. *Scientific Reports*, 15, 30639.

Kurosu, A., Coyle, J. L., Dudik, J. M. & Sejdić, E. (2019). Detection of swallow kinematic events from acoustic high resolution cervical auscultation signals in patients with stroke. *Archives of Physical Medicine and Rehabilitation*.

Kyritsis, K., Tatli, C. L., Diou, C. & Delopoulos, A. (2017). Automated analysis of in-meal eating behavior using a commercial wristband IMU sensor. *Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2843–2846.

Kyritsis, K., Diou, C. & Delopoulos, A. (2020). A Data Driven End-to-End Approach for In-the-Wild Monitoring of Eating Behavior Using Smartwatches. *IEEE Journal of Biomedical and Health Informatics*, 25(1), 22–34.

Langmore, S. E., Schatz, K. & Olsen, N. (1988). Fiberoptic endoscopic examination of swallowing safety : A new procedure. *Dysphagia*, 2(4), 216–219. doi : 10.1007/BF02414429.

Lazareck, L. J. & Moussavi, Z. K. (2004). Automated algorithm for swallowing sound detection. *Proceedings of the IEEE Engineering in Medicine and Biology Society*.

Lear, C. S. C., Jr, J. B. F. & Moorrees, C. F. A. (1965). The frequency of deglutition in man. *Archives of Oral Biology*, 10(1), 83–99. doi : 10.1016/0003-9969(65)90060-9.

Liebich, S. & Jax, P. (2022). Occlusion Effect Cancellation in Headphones and Hearing Devices – The Sister of Active Noise Cancellation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 35–48. doi : 10.1109/TASLP.2021.3130957.

Liu, Y., Butkow, K.-J., Stuchbury-Wass, J., Pullin, A., Ma, D. & Mascolo, C. (2024). RespEar : Earable-Based Robust Respiratory Rate Monitoring.

Lof, G. L. & Robbins, J. (1990). Test-retest variability in normal swallowing. *Dysphagia*, 4(4), 236–242. doi : 10.1007/BF02407271.

Logemann, J. A. (1998). *Evaluation and Treatment of Swallowing Disorders* (éd. 2nd). Austin, TX : PRO-ED.

Loret, C. (2015). Using Sensory Properties of Food to Trigger Swallowing : A Review. *Critical Reviews in Food Science and Nutrition*, 55(1), 140–145. doi : 10.1080/10408398.2011.649810.

Lostanlen, V., Salamon, J., Cartwright, M., McFee, B., Farnsworth, A., Kelling, S. & Bello, J. P. (2019). Per-Channel Energy Normalization : Why and How. *IEEE Signal Processing Letters*, 26(1), 39–43. doi : 10.1109/LSP.2018.2878620.

Ma, D., Ferlini, A. & Mascolo, C. (2021). OESense : Employing Occlusion Effect for In-ear Human Sensing. *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '21)*, pp. 175–187. doi : 10.1145/3458864.3467680.

Maeda, K., Nagasaka, M., Nagano, A., Nagami, S., Hashimoto, K., Kamiya, M., Masuda, Y., Ozaki, K. & Kawamura, K. (2023). Ultrasonography for Eating and Swallowing Assessment : A Narrative Review of Integrated Insights for Noninvasive Clinical Practice. *Nutrients*, 15(16), 3560. doi : 10.3390/nu15163560.

Makeyev, O., Lopez-Meyer, P., Schuckers, S., Besio, W. & Sazonov, E. (2012). Automatic food intake detection based on swallowing sounds. *Biomedical Signal Processing and Control*, 7(6), 649–656.

Martin, A. & Voix, J. (2018). In-Ear Audio Wearable : Measurement of Heart and Breathing Rates for Health and Safety Monitoring. *IEEE Transactions on Biomedical Engineering*, 65(6), 1256–1263. doi : 10.1109/TBME.2017.2720463.

Martino, R., Foley, N., Bhogal, S., Diamant, N., Speechley, M. & Teasell, R. (2005). Dysphagia After Stroke : Incidence, Diagnosis, and Pulmonary Complications. *Stroke*, 36(12), 2756–2763. doi : 10.1161/01.STR.0000190056.76543.eb.

Matsuo, K. & Palmer, J. B. (2008). Anatomy and Physiology of Feeding and Swallowing : Normal and Abnormal. *Physical Medicine and Rehabilitation Clinics of North America*, 19(4), 691–707. doi : 10.1016/j.pmr.2008.06.001.

Mehrban, M. H. K., Voix, J. & Bouserhal, R. E. (2024). Classification of Breathing Phase and Path with In-Ear Microphones. *Sensors*, 24(20), 6679. doi : 10.3390/s24206679.

Mejía, J., Dillon, H. & Fisher, M. (2008). Active cancellation of occlusion : An electronic vent for hearing aids and hearing protectors. *The Journal of the Acoustical Society of America*, 124(1), 235–240. doi : 10.1121/1.2908279.

Merck, C., Maher, C., Mirtchouk, M., Zheng, M., Huang, Y. & Kleinberg, S. (2016). Multimodality Sensing for Eating Recognition. *Pervasive Health*.

Murray, J., Langmore, S. E., Ginsberg, S. & Dostie, A. (1996). The Significance of Accumulated Oropharyngeal Secretions and Swallowing Frequency in Predicting Aspiration. *Dysphagia*, 11(2), 99–103. doi : 10.1007/BF00417898.

Nagy, A., Leigh, C., Hori, S. F., Molfenter, S. M., Shariff, T. & Steele, C. M. (2013). Timing Differences Between Cued and Noncued Swallows in Healthy Young Adults. *Dysphagia*, 28(3), 428–434. doi : 10.1007/s00455-013-9456-y.

Nakamura, A., Saito, T., Ikeda, D., Ohta, K., Mineno, H. & Nishimura, M. (2021). Automatic Detection of Chewing and Swallowing Using Multichannel Sound Information. *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*, pp. 173–175. doi : 10.1109/LIFETECH52111.2021.9391873.

Namasivayam, A. M. & Steele, C. M. (2015). Malnutrition and Dysphagia in Long-Term Care : A Systematic Review. *Journal of Nutrition in Gerontology and Geriatrics*, 34(1), 1–21. doi : 10.1080/21551197.2014.1002656.

Nazari, G. & MacDermid, J. C. (2020). Reliability of Zephyr BioHarness Respiratory Rate at Rest, During the Modified Canadian Aerobic Fitness Test and Recovery. *Journal of Strength and Conditioning Research*, 34(1), 264–269. doi : 10.1519/JSC.0000000000003046.

Nguyen, N. P., Moltz, C. C., Frank, C., Vos, P., Smith, H. J., Karlsson, U. & Sallah, S. (2004). Dysphagia following chemoradiation for locally advanced head and neck cancer. *Annals of Oncology*, 15(3), 383–388. doi : 10.1093/annonc/mdh101.

Nielsen, C. & Darkner, S. (2011). The cartilage bone junction and its implication for deep canal hearing instrument fittings. *The Hearing Journal*, 64(3), 35–36.

Olubanjo, T. & Ghovanloo, M. (2014). Real-time swallowing detection based on tracheal acoustics. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Panni, L., Cosoli, G., Antognoli, L. & Scalise, L. (2024). Measurement of respiratory rate with cardiac belt : Metrological characterization. *Measurement : Sensors*, 34, 101244. doi : 10.1016/j.measen.2024.101244.

Pantelopoulos, A. & Bourbakis, N. G. (2010). A Survey on Wearable Sensor-Based Systems for Health Monitoring and Prognosis. *IEEE Transactions on Systems, Man, and Cybernetics, Part C : Applications and Reviews*, 40(1), 1–12. doi : 10.1109/TSMCC.2009.2032660.

Papapanagiotou, V., Diou, C. & Delopoulos, A. (2017a, 07). Chewing detection from an in-ear microphone using convolutional neural networks. 2017. doi : 10.1109/EMBC.2017.8037060.

Papapanagiotou, V., Diou, C., Zhou, L., van den Boer, J., Mars, M. & Delopoulos, A. (2017b). A Novel Chewing Detection System Based on PPG, Audio, and Accelerometry. *IEEE Journal of Biomedical and Health Informatics*, 21(3), 607–618. doi : 10.1109/JBHI.2016.2625271.

Papapanagiotou, V., Diou, C. & Delopoulos, A. (2021). Recognition of Food-Texture Attributes from Audio-Visual Signals using a Temporal Attention-based Deep Neural Network. *Proceedings of the ICPR Workshops*.

Päßler, S. & Fischer, W.-J. (2011). Acoustical method for objective food intake monitoring using a wearable sensor system. *Proceedings of the 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. doi : 10.4108/icst.pervasivehealth.2011.246029.

Päßler, S., Wolff, M. & Fischer, W.-J. (2012). Food intake monitoring : an acoustical approach to automated food intake activity detection and classification of consumed food. *Physiological Measurement*, 33(6), 1073–1093.

Röddiger, T., Küttner, M., Lepold, P., King, T., Moschina, D., Bagge, O., Paradiso, J. A., Clarke, C. & Beigl, M. (2025). OpenEarable 2.0 : Open-Source Earphone Platform for Physiological Ear Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(1), 1–33. doi : 10.1145/3712069.

Röddiger, T. et al. (2022). Sensing with Earables : A Systematic Literature Review and Taxonomy of Phenomena. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*.

Rudney, J. D., Ji, Z. & Larson, C. J. (1995). The prediction of saliva swallowing frequency in humans from estimates of salivary flow rate and the volume of saliva swallowed. *Archives of Oral Biology*, 40(6), 507–512. doi : 10.1016/0003-9969(95)00004-9.

Ruoppolo, G., Schettino, I., Frasca, V., Giacomelli, E., Prosperini, L., Cambieri, M. et al. (2013). Dysphagia in amyotrophic lateral sclerosis : prevalence and clinical findings. *Acta Neurologica Scandinavica*, 128(6), 397–401. doi : 10.1111/ane.12136.

Sazonov, E. S., Makeyev, O., Schuckers, S., Lopez-Meyer, P., Melanson, E. L. & Neuman, M. R. (2010). Automatic detection of swallowing events by acoustical means for applications of monitoring of ingestive behavior. *IEEE Transactions on Biomedical Engineering*, 57(3), 626–633.

Sejdić, E., Steele, C. M. & Chau, T. (2013). Classification of penetration-aspiration versus healthy swallows using dual-axis swallowing accelerometry signals in dysphagic subjects. *IEEE Transactions on Biomedical Engineering*, 60(7), 1859–1866.

Shin, B., Lee, S. H., Kwon, K., Lee, Y. J., Crispe, N., Ahn, S.-Y., Shelly, S., Sundholm, N., Tkaczuk, A., Yeo, M.-K., Choo, H. J. & Yeo, W.-H. (2024). Automatic Clinical Assessment of Swallowing Behavior and Diagnosis of Silent Aspiration Using Wireless Multimodal Wearable Electronics. *Advanced Science*, 11(34), 2404211. doi : 10.1002/advs.202404211.

Shirazi, S. S., Buchel, C., Daun, R., Lenton, L. & Moussavi, Z. (2012). Detection of swallows with silent aspiration using swallowing and breath sound analysis. *Medical & Biological Engineering & Computing*, 50, 1261–1268.

Shrimal, Y. & Nandurkar, A. (2021). Self-reported hearing status and audiometric thresholds among college students using headphones. *Journal of Otolaryngology-ENT Research*, 13(3), 60–68. doi : 10.15406/joentr.2021.13.00492.

So, B. P.-H., Chan, T. T.-C., Liu, L., Yip, C. C.-K., Lim, H.-J., Lam, W.-K., Wong, D. W.-C., Cheung, D. S. K. & Cheung, J. C.-W. (2023). Swallow Detection with Acoustics and Accelerometric-Based Wearable Technology : A Scoping Review. *International Journal of Environmental Research and Public Health*, 20(1), 170. doi : 10.3390/ijerph20010170.

So, S., Tadj, T., Schwerin, B., Chang, A. B. & Frakking, T. T. (2025). Use of Transfer Learning for the Automated Segmentation and Detection of Swallows via Digital Cervical Auscultation in Children. *Dysphagia*, 40, 1371–1380. doi : 10.1007/s00455-025-10833-3.

Song, Y., Yun, I., Giovanoli, S., Easthope, C. A. & Chung, Y. (2025). Multimodal deep ensemble classification system with wearable vibration sensor for detecting throat-related events. *npj Digital Medicine*, 8, 14. doi : 10.1038/s41746-024-01417-w.

Steele, C. M., Alsanei, W. A., Ayanikalath, S., Barbon, C. E. A., Chen, J., Cichero, J. A. Y., Coutts, K., Dantas, R. O., Duivestein, J., Giosa, L. et al. (2015). The influence of food texture and liquid consistency modification on swallowing physiology and function : a systematic review. *Dysphagia*, 30(1), 2–26.

Sura, L., Madhavan, A., Carnaby, G. & Crary, M. A. (2012). Dysphagia in the elderly : management and nutritional considerations. *Clinical Interventions in Aging*, 7, 287–298. doi : 10.2147/CIA.S23404.

Suttrup, I. & Warnecke, T. (2016). Dysphagia in Parkinson's Disease. *Dysphagia*, 31(1), 24–32. doi : 10.1007/s00455-015-9671-9.

Walker, W. & Bhatia, D. (2011). Swallow Sound Analysis for Automated Ingestion Detection. *Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA)*.

Wei, L., Pan, Y., Zhong, Y. & Huan, R. (2018). A Novel Approach for Swallow Detection by Fusing Throat Acceleration and PPG Signal. *IEEE Access*.

Winiker, K., Hammond, R., Thomas, P., Dimmock, A. & Huckabee, M.-L. (2022). Swallowing assessment in patients with dysphagia : Validity and reliability of a pocket-sized ultrasound system. *International Journal of Language & Communication Disorders*, 57(3), 539–551. doi : 10.1111/1460-6984.12703.

Yoshimoto, T., Taniguchi, K., Kurose, S. & Kimura, Y. (2022). Validation of Earphone-Type Sensors for Non-Invasive and Objective Swallowing Function Assessment. *Sensors*, 22(14), 5176. doi : 10.3390/s22145176.

Zhang, X., Wang, Y. et al. (2024). The EarSAVAS Dataset : Enabling Subject-Aware Vocal Activity Sensing on Earables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2), Article 83.

Zurbrügg, T., Stirnemann, A., Kuster, M. & Lissek, H. (2014). Investigations on the physical factors influencing the ear canal occlusion effect caused by hearing aids. *Acta Acustica united with Acustica*, 100(3), 527–536.