Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Retour auditif interne de la production de parole : mesures préliminaires de la vibration osseuse par accélérométrie et comparaison au son aérien

Raphaël Vancheri Coriandre Vilain
Nathalie Henrich Bernardoni Pierre Baraduc

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Grenoble, F-38000, France
[prenom].[nom]@gipsa-lab.fr

RÉSUMÉ

Lorsqu'on parle, le retour auditif se décompose en une voie aérienne et une voie interne ou 'par conduction osseuse'. Un locuteur entend les deux composantes, contrairement au récepteur. Alors que la moitié du signal cochléaire est interne, on connaît mal l'information qu'il véhicule et comment elle impacte le contrôle moteur oral. Dans cette étude, nous considérons deux indicateurs du signal auditif interne pendant la production de parole, la vibration des dents de la mâchoire supérieure et le son enregistré près du tympan. Une méthode de conversion de voix nous permet d'évaluer les différences informationnelles entre voix aérienne et voix "osseuse" interne. Comme observé précédemment par la simple méthode d'enregistrement péritympanique, la somme des retours acoustiques aérien et interne amène une lisibilité supérieure des trajectoires formantiques qui pourrait faciliter le contrôle de la production de parole.

ABSTRACT

Internal acoustic feedback during speech production : preliminary measure of bone vibration during speech with an accelerometer and comparison to aerial-conducted sound.

During speech, the auditory feedback involves both an aerial component picked up by the external ear, and an internal vibration : the 'bone conduction' component. While a speaker hears both components, a listener only hears the aerial part. Although half of the cochlear signal comes from internal conduction, the information it conveys, and how it impacts oral motor control, is still unclear. In this study, we considered two proxies of the internal auditory signal : the vibration of the upper teeth and the sound emitted next to the eardrum. A voice conversion method allows to evaluate the informational differences between aerial and internal bone voice. As preceedingly observed with the peritympanic recording method, the summation of internal and aerial feedback leads to clearer formantic trajectories, which may facilitate speech motor control.

MOTS-CLÉS : conduction osseuse, production de parole, perception, contrôle moteur.

KEYWORDS: bone conduction, speech production, perception, motor control.

1 Introduction

Durant la production de parole, les tissus mous péri-oraux ainsi que les os sous-jacents transmettent un signal acoustique interne jusqu'à la cochlée de manière directe (vibration du rocher) et indirecte

(vibration des osselets, vibration du tympan). Ce signal interne est éminemment complexe à objectiver du fait de la quasi impossibilité de mesurer directement les vibrations acoustiques au niveau de la cochlée. L'essentiel de nos connaissances sur la conduction interne (communément mais improprement appelée conduction osseuse) provient d'études sur l'animal (p. ex. Tonndorf *et al.* 1966) ou sur des cadavres (Eeg-Olofsson *et al.*, 2008), qui ont permis de quantifier la conduction des vibrations dans les différentes structures physiologiques (Stenfelt & Goode, 2005), mais bien évidemment pas pendant une vibration laryngée. Pörschmann (2000), à l'aide d'une méthode de masquage, a étudié le premier le spectre sonore de la parole interne, et mis en évidence une différence entre phones voisés et non-voisés. Plus récemment, Reinfeldt *et al.* (2010) ont utilisé l'enregistrement direct des vibrations du conduit auditif comme un indicateur du signal par conduction osseuse, et observé les différences de spectre sonore entre une petite sélection de voyelles ou consonnes voisées isolées. Nous avons précédemment utilisé cette méthode pour décrire le contenu spectral de signaux aériens et péri-tympaniques pendant la production de parole continue (Baraduc & Vilain, 2022). Toutefois ces travaux n'ont mesuré que la vibration des tissus mous ; une partie du signal interne restait inconnue. Dans cet article, nous décrivons une méthode pour dépasser cette limite, et les résultats préliminaires obtenus sur 3 sujets pilotes.

2 Matériel et méthodes

2.1 Sujets

Trois sujets francophones du laboratoire (1 femme, 2 hommes, 24-49 ans) parmi les auteurs, ont participé sans compensation à cette expérience en tant que sujets pilotes.

2.2 Dispositif expérimental

Les sujets ont été équipés d'un capteur acoustique intra-auriculaire (microphone capillaire Etymotic ER-7C) permettant d'enregistrer le signal sonore péritympanique. Ce signal est isolé acoustiquement de l'environnement extérieur par une boite fabriquée au laboratoire, placée contre l'oreille du sujet et écrantant les signaux extérieurs d'au moins 30 dB sur l'ensemble du spectre sonore ; son volume est suffisant pour éviter les résonances (effet d'occlusion) que généreraient de simples bouchons d'oreille. Les sujets sont également équipés d'un accéléromètre miniature (PCB Piezotronics 352A73 conditionné par une carte d'acquisition Data Translation) fixé par une résine photopolymérisable sur une dent de la mâchoire supérieure (incisive, canine ou prémolaire en fonction de configuration de la surface dentaire). Cet accéléromètre mesure les vibrations locales des os, après leur passage à travers le ligament alvéolo-dentaire et la dentine. Trois microphones aériens (BK 4189) sont également utilisés pour mesurer les signaux acoustiques : à 50 cm de la bouche du sujet (dans l'axe), à 2 cm de l'oreille droite (dans l'espace extérieur), à 2 cm de l'oreille gauche (dans la boîte isolante). L'acquisition des données est effectuée sous Matlab à l'aide de la PsychoPhysics Toolbox (pour les données audio) et d'une toolbox DataTranslation pour les données d'accélérométrie.

2.3 Protocole

Après une étape d'équilibration perceptive ne se rapportant pas aux résultats présentés ici, les sujets devaient lire à voix haute les 100 premières phrases du corpus FHarvard (Aubanel *et al.*, 2020), présentées sur un moniteur placé en face d'eux.

2.4 Analyse des données

Les signaux ont été synchronisés par intercorrélation (avec pour signal de référence l'enregistrement du microphone placé face au sujet), puis nous avons édité parallèlement tous les signaux afin d'en enlever les répétitions et erreurs. Les signaux d'enregistrement intra-auriculaire et d'accélérométrie ont été amplifiés (de 15 dB et 40 dB respectivement) ; ce dernier a préalablement été filtré passe-haut à 5 Hz afin d'en retirer la composante continue. Enfin, les deux signaux ont également été débruités par soustraction spectrale (-15 dB, par Audacity). Les enregistrements ont été ensuite segmentés sous Praat au moyen d'un outil d'alignement phonétique automatique, EasyAlign (Goldman, 2010) ; les résultats ont été manuellement vérifiés et édités au besoin. Afin de mieux cerner les différences en terme d'information portée par les signaux aérien et osseux, nous avons cherché à déterminer dans quelle mesure ils étaient interconvertibles. Cette conversion de voix a été réalisée par régression par mélange de gaussiennes à partir de coefficients mel-cepstraux (trames de 5 ms), grâce à un code partagé par T. Hueber (Hueber & Bailly, 2016), après sous-échantillonage à 16 kHz. Les différences spectrales en dB ont été calculées dans la bande 0–5 kHz. L'enveloppe spectrale de la parole est aisément calculée à partir des coefficients mel-cepstraux ; la resynthèse peut s'effectuer avec un filtre MLSA. Ces calculs ont été réalisés avec SPTK 4.0 (`github.com/sp-nitech/SPTK/releases`). La visualisation des résultats a été obtenue sous Matlab, en utilisant en particulier la fonction `spectrogram`.

3 Résultats

3.1 Observations de surface : signal aérien, osseux, et des tissus mous

Le signal de l'accéléromètre reflète-t-il des propriétés de la conduction interne des vibrations de parole ? En quoi diffère-t-il du l'enregistrement péritympanique, censé refléter surtout la vibration des tissus mous pendant la vocalisation ? La figure 1 donne un exemple de spectrogrammes des signaux de parole, selon qu'ils sont aériens ou tirés de l'accéléromètre (pour simplifier, nous utiliserons le raccourci abusif de "signal osseux" dans la suite des résultats), ou du signal aérien. Comme observé précédemment avec l'enregistrement péritympanique (Baraduc & Vilain, 2022), le signal interne permet de mieux suivre certaines transitions formantiques. Toutefois, alors que le signal des tissus mous était particulièrement différent pendant les fricatives, on peut déjà remarquer sur ces exemples que ce n'est pas le cas pour le signal osseux. Une comparaison directe des deux méthodes d'estimation du retour auditif interne n'est malheureusement pas possible sur ce jeu de données, les enregistrements péritympaniques ont été particulièrement décevants pour deux sujets (cf. infra).

Une deuxième observation a trait à un aspect technique important de l'estimation du retour par conduction interne. Les tissus mous ayant tendance à absorber les vibrations de haute fréquence, l'enregistrement péritympanique est délicat au-dessus de 4 kHz. Ceci n'est pas le cas avec l'accéléromètre, dont la plage de mesure s'étend jusqu'à 40 kHz. En fait, la densité spectrale de la parole osseuse n'est

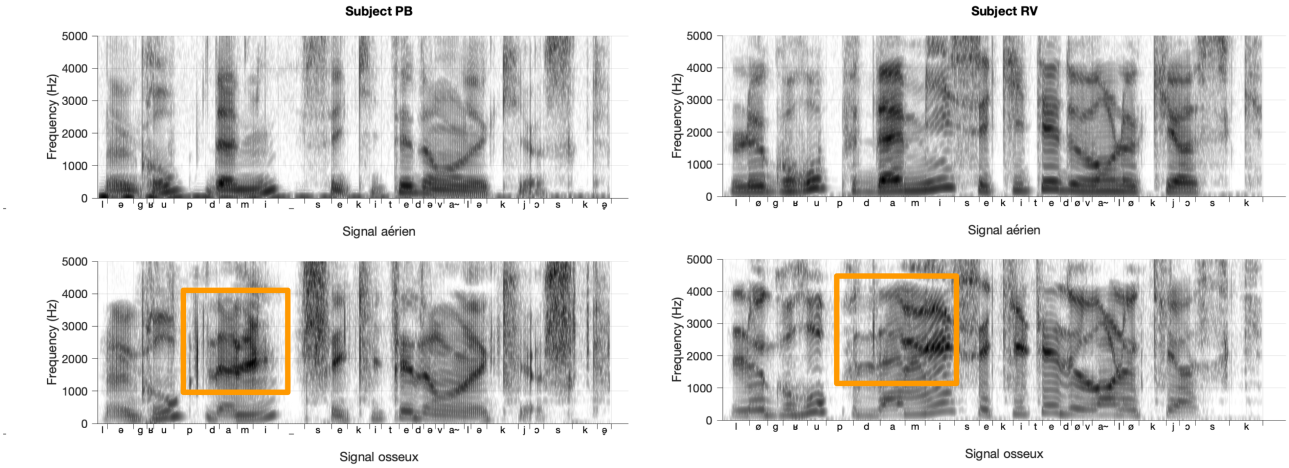

FIGURE 1 – Spectrogrammes de parole chez deux sujets, pour la phrase *"le groupe d'amis s'est quitté devant le kiosque"*. En haut, les spectrogrammes du signal aérien ; en bas les spectrogrammes du signal interne. Comme vu précédemment avec les enregistrements péritympaniques, la trajectoire de F2 est particulièrement lisible dans le signal interne pendant *d'amis* (encadré orange).

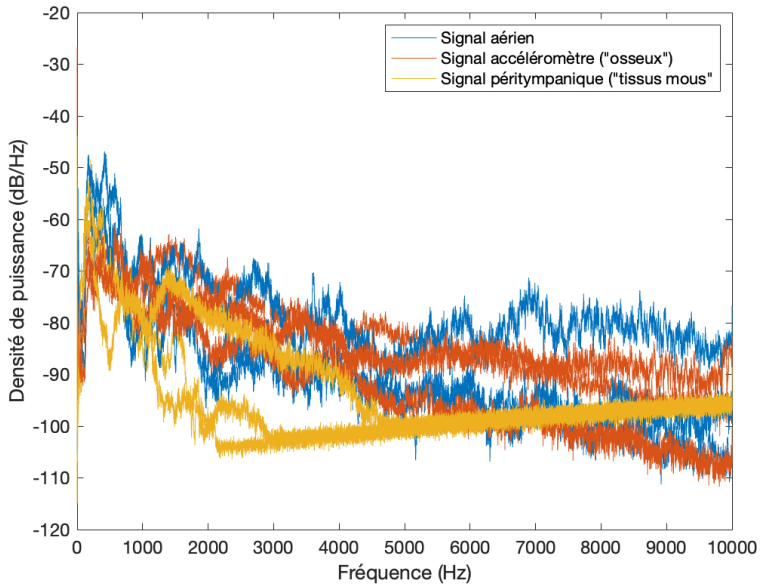

FIGURE 2 – Densité de puissance du signal de parole, pour les 3 sujets de l'expérience pilote. En bleu, me signal aérien ; en rouge le signal de l'accéléromètre ("osseux") ; en jaune le signal péritympanique ("tissus mous")

pas différente de la parole aérienne, comme on le voit figure 2. Par ailleurs, on remarque que pour un sujet, la densité de puissance péritympanique n'est pas très différente des autres signaux aériens ou osseux, mais est tronquée à 4,5 kHz par le plancher de bruit du microphone. L'amortissement des vibrations hautes fréquences par les tissus mous ne semble pas, ou au moins pas toujours, être responsable de cette difficulté à estimer le retour auditif interne de la parole dans le haut du spectre. Quoi qu'il en soit, en pratique, le signal de l'accéléromètre permet une bien meilleure estimation de la conduction interne des hautes fréquences, comme on peut l'observer également sur la figure 1.

3.2 Analyse de différences informationnelles

Afin de quantifier si le signal "osseux" porte une information redondante avec le signal aérien, nous avons utilisé une méthode de conversion de voix (cf. §2.4), comme dans nos travaux précédents. Brièvement, le principe en est le suivant : les différences spectrales entre signal "osseux" et signal aérien converti en "osseux" sont révélatrices de l'information *spécifique* portée par le signal osseux, puisqu'elles correspondent à des parties du spectre qu'on peut difficilement prédire à partir de la voix aérienne. L'argument est symétrique et l'information spécifique portée par le signal aérien peut être estimée par la conversion de voix réciproque "osseux"→ aérien.

La figure 3 permet de comparer des spectrogrammes de parole aérienne et "osseuse", avec une mise en évidence par la couleur des parties du spectre difficilement convertibles, donc spécifiques de chaque signal. Cet exemple peut être comparé à une figure similaire de (Baraduc & Vilain, 2022). On remarque que les transitions formantiques sont plus lisibles dans le signal "osseux", ce qui est particulièrement clair dans l'extrait illustré *"son dernier c[ongé]"*, pour lequel les /o~/, /d/, et la séquence /nj/ ont effectivement une mesure de spécificité osseuse élevée.

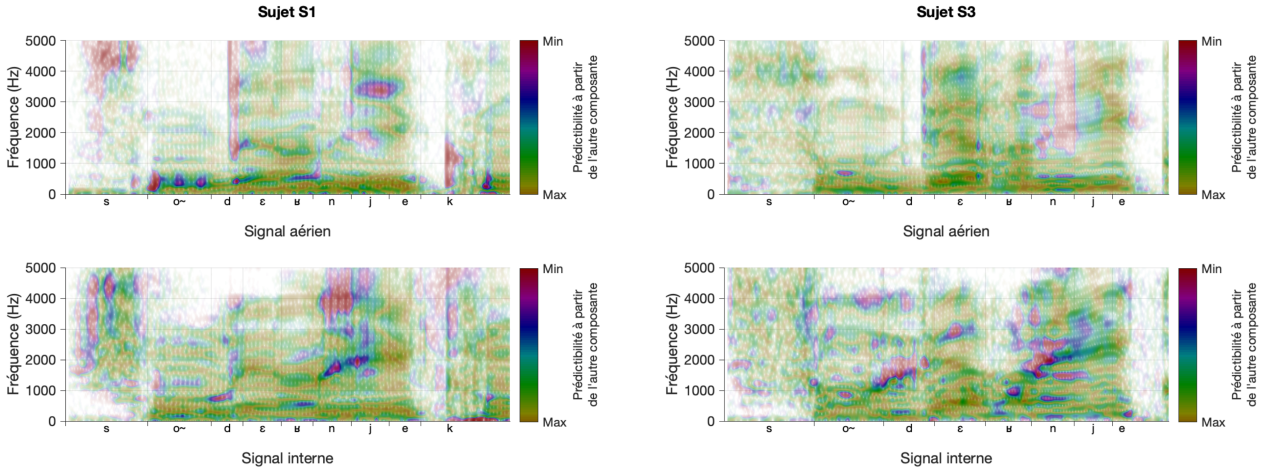

FIGURE 3 – Spectrogrammes du signal aérien (en haut) et du signal "osseux" (en bas), correspondant au début de la prononciation de la phrase *"Son dernier congé dura deux semaines"*. Les couleurs indiquent les différences statistiquement non interconvertibles (selon le code couleur représenté à droite). On remarque que les voyelles et consonnes nasales, les occlusives voisées ou même le /ʁ/ ont une signature osseuse particulière (et plus claire).

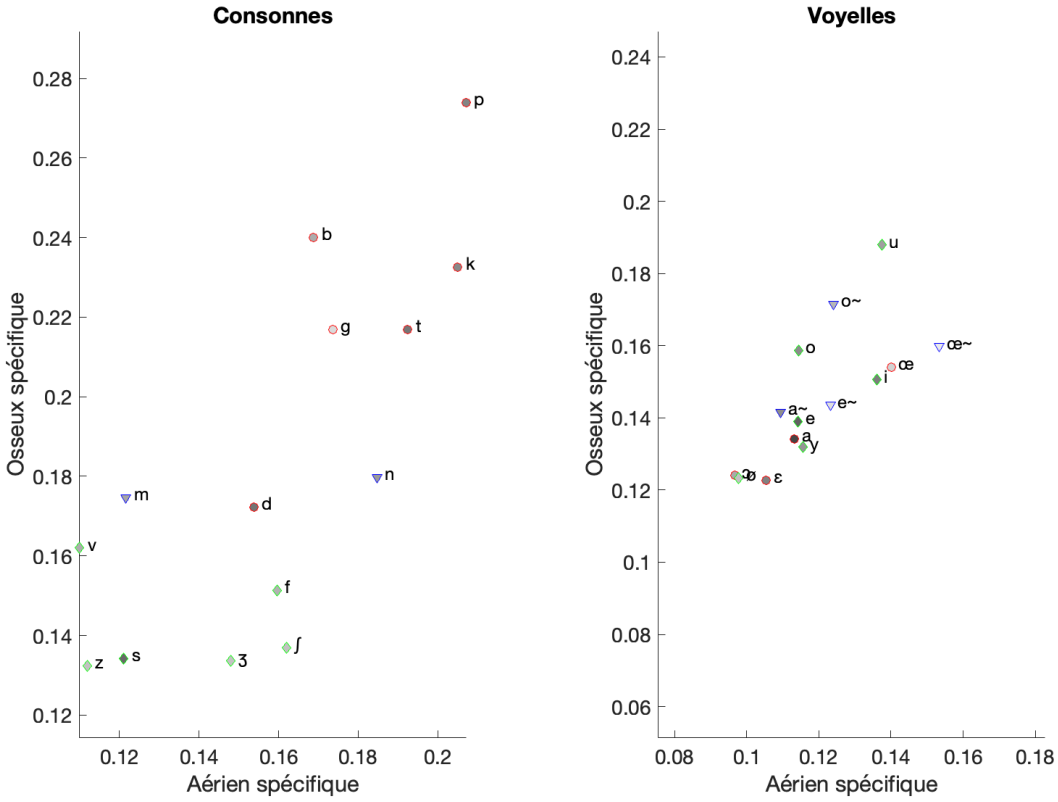

FIGURE 4 – Distribution des mesures de spécificité du signal aérien (axe horizontal) et du signal "osseux" (axe vertical) pour une sélection de phonèmes, consonnes (panneau de gauche) ou voyelles (panneau de droite). La spécificité est définie comme la distance maximale entre enveloppes spectrales du signal réel et du signal prédit par conversion depuis l'autre source. On remarque que les occlusives sont bien différenciées, les voyelles et consonnes nasales fermées également. Les fricatives portent peu d'information spécifique à la voie osseuse, en moyenne.

Pour donner un aperçu global, nous avons segmenté les signaux de parole et quantifié pour chaque phonème l'information spécifique portée par le signal aérien ou "osseux", en considérant comme mesure de spécificité la différence maximale d'enveloppe spectrale entre signal réel et signal prédit (converti), pendant la durée du phonème. Ces valeurs sont variables selon le contexte de coarticulation et le sujet, et ici, le nombre de participants étant limité, nous avons représenté la moyenne inter-sujets sur tout le jeu de données dans le graphe de la figure 4.

Les résultats confortent l'idée que le signal "osseux" est souvent différemment informatif du signal aérien. Les occlusives sont particulièrement différentes et différemment informatives dans les deux signaux. Une étude plus détaillée révèle que les occlusives non voisées portent souvent des traces de formants visibles dans le bruit de plosion, voire avant ; ceci est plus variable dans les consonnes voisées, dont certains exemples sont particulièrement clairs quant à la transition formantique en cours, d'autres moins. Les voyelles fermées, particulièrement lorsqu'elles sont postérieures ou nasales, ont aussi une signature assez spécifique à chaque composante auditive. Enfin, on remarque que les fricatives portent peu d'information spécifique dans le signal osseux, à l'inverse de ce que nous avions constaté précédemment dans le signal péritympanique. L'origine de cette différence devra être élucidée, mais il est possible qu'elle soit liée à la position de l'accéléromètre par rapport à la cavité orale postérieure, la résonance de cette cavité excitée par le bruit aéro-acoustique étant nette dans les enregistrements péritympaniques.

4 Discussion

Ces premiers résultats prolongent notre précédente description des différences entre retour acoustique aérien de la parole naturelle, et retour par conduction interne. Ils sont un aperçu des développements méthodologiques en cours qui devraient nous permettre de mieux caractériser le retour auditif de sa propre voix, en associant deux techniques complémentaires d'évaluation de la vibration interne. Si cet article décrit des travaux préliminaires, les techniques mises en œuvre ont également des limites que nous rappelons ici.

Une première limite tient au placement des capteurs. Si la mesure de la vibration des tissus mous est a priori idéale près du tympan, la mesure de la vibration des os de la tête serait probablement meilleure à une position plus proche de l'os temporal ; ceci est toutefois difficilement compatible avec un positionnement simple de l'accéléromètre. Par ailleurs il faut rappeler que la mesure intègre l'élasticité du ligament odonto-alvéolaire, dont les caractéristiques mécaniques pourraient affecter le signal (le graphe de la figure 2 étant toutefois assez rassurant).

Ensuite, notre mesure de différence informationnelle par la conversion de voix a également un certain nombre de limites. Tout d'abord, elle peut être biaisée par le modeste corpus d'entraînement (seulement 5 minutes de parole). Nous travaillons actuellement sur une mesure dérivée de la théorie de l'information, qui sera indépendante de la spécification d'un modèle. D'autre part, nous avons résumé dans la figure 4 les différences spectrales en considérant leur somme sur la bande 0–5 kHz, ce qui ne rend pas compte du caractère éventuellement crucial d'une information sonore particulière à l'intérieur de ce spectre.

Néanmoins, ces premiers résultats d'estimation directe de la vibration osseuse nous semblent très encourageants. Un des intérêts d'évaluer le retour interne par une méthode d'accélérométrie est de libérer le sujet d'un capteur situé dans l'oreille (et d'une boîte acoustiquement isolée), ce qui permet

de simplifier les expériences sur la parole et la voix, voire la pratique musicale au sens large. A court terme, nous chercherons toutefois à mieux caractériser les différences entre retour interne purement osseux et conduction via les tissus mous, et évaluer leur variabilité interindividuelle.

Remerciements

Les auteurs remercient Thomas Hueber pour sa disponibilité, ses conseils et le partage de son code Matlab de conversion de voix.

Références

AUBANEL V., BAYARD C., STRAUSS A. & SCHWARTZ J. (2020). The Fharvard corpus : A phonemically-balanced French sentence resource for audiology and intelligibility research. *Speech Communication*, **124**, 68–74. DOI : https ://doi.org/10.1016/j.specom.2020.07.004.

BARADUC P. & VILAIN C. (2022). Retours acoustiques de la production de parole : caractérisation des différences informationnelles entre le son aérien et le son par conduction osseuse. In *Proc. XXXIVe Journées d'Études sur la Parole – JEP 2022*, p. 980–988. DOI : 10.21437/JEP.2022-104.

EEG-OLOFSSON M., STENFELT S., TJELLSTRÖM A. & GRANSTRÖM G. (2008). Transmission of bone-conducted sound in the human skull measured by cochlear vibrations. *International Journal of Audiology*, **47**(12), 761–769. DOI : doi :10.1080/14992020802311216.

GOLDMAN J.-P. (2010). Easyalign : a friendly automatic phonetic alignment tool under praat. In *Proc. Interspeech 2011*, p. 3233–3236. DOI : 10.21437/Interspeech.2011-815.

HUEBER T. & BAILLY G. (2016). Statistical conversion of silent articulation into audible speech using full-covariance hmm. *Computer Speech Language*, **36**, 274–293. DOI : https ://doi.org/10.1016/j.csl.2015.03.005.

PÖRSCHMANN C. (2000). Influences of Bone Conduction and Air Conduction on the Sound of One's Own Voice. *Acta Acustica united with Acustica*, **86**(6).

REINFELDT S., ÖSTLI P., HÅKANSSON B. & STENFELT S. (2010). Hearing one's own voice during phoneme vocalization—Transmission by air and bone conduction. *The Journal of the Acoustical Society of America*, **128**(2), 751–762. DOI : 10.1121/1.3458855.

STENFELT S. & GOODE R. L. (2005). Bone-Conducted Sound : Physiological and Clinical Aspects. *Otology & Neurotology*, **26**(6), 1245–1261. DOI : 10.1097/01.mao.0000187236.10842.d5.

TONNDORF J., GREENFIELD E. C. & KAUFMAN R. S. (1966). The Relative Efficiency of Air and Bone Conduction in Cats. *Acta Oto-Laryngologica*, **61**(sup213), 105–123. DOI : 10.3109/00016486609120802.