XXXIVe Journées d'Études sur la Parole – JEP 2022
13-17 June 2022, Île de Noirmoutier, France

Retours acoustiques de la production de parole : caractérisation des différences informationnelles entre le son aérien et le son par conduction osseuse

Pierre Baraduc[1] Coriandre Vilain[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
[prenom].[nom]@grenoble-inp.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 le son émis par le conduit auditif pendant la production de parole comme indicateur du signal interne. Après enregistrement des signaux internes et aériens, une méthode de conversion de voix nous permet d'évaluer leurs différences informationnelles. Comme anticipé, les voyelles et consonnes nasales corrèlent avec plus d'intensité et d'information osseuse ; de manière moins attendue, on observe également plus d'information osseuse pour les consonnes occlusives et fricatives. De façon globale, la somme des retours acoustiques aérien et osseux amène une lisibilité supérieure des trajectoires formantiques qui pourrait faciliter le contrôle de la production de parole.

ABSTRACT

Acoustic feedbacks during speech production : informational differences between aerial and bone-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 the sound emitted by the eardrum and ear canal during speech as a proxy for internal sound. After recording aerial and internal sound, we made use of a voice conversion method to evaluate their informational differences. As expected, nasal vowels and consonants correlate with stronger internal signal ; more surprisingly, we observe also more internal information during stop and fricative consonants. Overall, 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.

10.21437/JEP.2022-104

1 Introduction

Durant la production de parole, les tissus mous péri-oraux ainsi que les os sous-jacents vibrent, conduisant un signal acoustique interne jusqu'à la cochlée de manière directe (vibration du rocher) et indirecte (vibration des osselets, vibration du tympan). En grande partie à cause de la difficulté à objectiver ce signal, on connaît peu de choses sur cette partie du retour acoustique, alors qu'il pourrait jouer un rôle complémentaire important dans le contrôle moteur de la parole. L'essentiel de nos connaissances sur la conduction osseuse provient d'études sur l'animal (p. ex. Tonndorf 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 sons 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 parole aérienne et osseuse pour une petite sélection de voyelles ou consonnes voisées isolées. Si cette méthode comporte des limitations évidentes, elle peut être employée de manière bien plus extensive et il nous a paru important de chercher à l'employer pendant une production de parole naturelle. Nous en rapportons ici les données préliminaires.

Bien entendu, le signal acoustique péri-tympanique n'est pas identique à celui perçu par l'auditeur, et une correction, un peu délicate à réaliser, est nécessaire pour pouvoir comparer le retour acoustique aérien effectif et l'estimation du retour aérien qui serait perceptivement équivalent au signal interne. Afin de contourner cette difficulté, nous avons préféré utiliser ici une méthode de conversion de voix (Phung *et al.*, 2012) pour mettre en évidence, indépendamment de différences perceptives entre retour aérien et interne, si un signal porte une information absente de l'autre. Plus précisément, si l'on convertit du signal aérien en signal interne (en réalité, signal acoustique péri-tympanique, mais dans la suite de l'article, pour simplifier l'exposé, nous utiliserons cet abus de langage), on peut comparer les spectrogrammes de ce signal prédit avec celui du signal réel. La divergence entre les deux signale l'existence d'une information portée uniquement par le signal interne.

2 Méthodes

2.1 Sujets

Six sujets francophones du laboratoire (4 femmes, 2 hommes, 22-49 ans) dont les deux auteurs, ont participé sans compensation à cette expérience.

2.2 Dispositif expérimental

L'enregistrement du signal acoustique péri-tympanique nécessite d'isoler acoustiquement le conduit auditif du sujet du signal aérien de la parole. Nous avons donc construit une "boite à oreille", sorte d'oreillette de casque géante, qui permet d'accéder confortablement au conduit auditif du sujet et écrante d'au moins 30 dB sur l'ensemble du spectre sonore tout en ne produisant pas l'effet d'occlusion que généreraient de simples bouchons d'oreille.

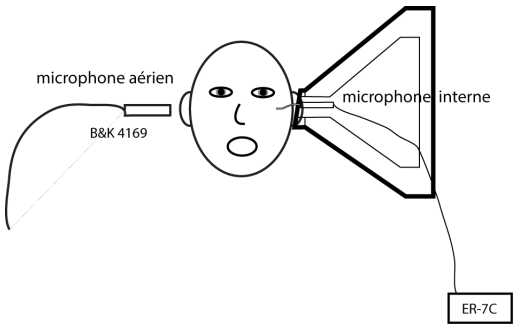

FIGURE 1 – Dispositif expérimental.

Le dispositif expérimental est illustré Figure 1. Le sujet, assis confortablement, collait l'oreille gauche à cette boîte. Nous avons enregistré le son au fond du conduit auditif grâce à un microphone capillaire (Etymotic ER-7C), et le son aérien à la pinna de l'oreille contralatérale (Bruel & Kjaer 4169). Ces deux signaux étaient échantillonnés à 48 kHz par une interface audio RME Fireface 800, commandée via Matlab (Psychophysics toolbox).

2.3 Protocole

Après une étape d'équilibration perceptive qui n'a pas d'intérêt pour les résultats présentés ici, les sujets devaient lire à voix haute les 100 premières phrases du corpus FHarvard (Aubanel *et al.*, 2020). Chaque liste de 10 phrases de ce corpus original (qui comprend 70 listes) a un contenu phonémique équilibré entre les mots.

2.4 Analyse des données

Les enregistrements aériens et internes ont été débruités par une méthode de soustraction spectrale. La segmentation phonétique a été réalisée sous *Praat* à l'aide du module EasyAlign (Goldman, 2011), puis vérifiée et corrigée manuellement si besoin. Les spectrogrammes ont été calculés avec Matlab. Pour la conversion de voix, nous avons suivi une méthode développée par Toda & Shikano (2005) dans son implémentation par T. Hueber (Hueber & Bailly, 2016). Les signaux ont été transformés en séries de coefficients mel-cepstraux via SPTK. Pour chaque conversion (aérien $\rightarrow$ interne ou réciproque), un modèle par mélange de gaussiennes a ensuite été entraîné sur ces matrices de coefficients, obtenus sur une base de données de 80 phrases. Enfin, une régression via ce mélange de gaussiennes a permis la conversion entre coefficients mel-cepstraux pour chaque pas de temps, qui ont été enfin retransformés en enveloppes spectrales.

Pour chaque échantillon de temps, nous avons considéré la somme des valeurs absolues des différences entre enveloppes spectrales effective et prédite par la conversion comme mesure de la divergence entre signal effectif et converti, c'est-à-dire une mesure de l'information présente dans le premier signal que la conversion du deuxième ne parvient pas à reproduire. Pour une phrase donnée, afin d'aligner temporellement les signaux provenant de locuteurs différents, nous avons calculé le temps médian de production de chaque phone, et représenté les mesures de chaque sujet dans cette base de temps

grâce à une interpolation temporelle (’time warping’). Ceci nous a permis de calculer une moyenne inter-sujets des différences informationnelles. Ce procédé n’est pas parfait : pour pouvoir réaliser cet alignement, les schwas et silences produits par certains sujets ont dû être ignorés (et considérés comme parties du phone précédent).

3 Résultats

3.1 Comparaison des signaux bruts

La comparaison visuelle des spectrogrammes des signaux aérien et interne permet d’emblée de situer les différences entre ces deux composantes du retour acoustique. On constate, comme déjà rapporté par Reinfeldt *et al.* (2010), que la conduction osseuse renforce les voyelles et consonnes nasales qui donnent lieu à une forte résonance acoustique dans les fosses nasales, proches de l’oreille. Toutefois, d’autres différences apparaissent : le deuxième formant des voyelles fermées est plus marqué, et certains formants deviennent visibles dans le signal interne pendant les consonnes occlusives ou fricatives. Si cet effet est plus marqué pendant le voisement, il est également très net sur des consonnes non voisées (p.ex. le /k/, /s/ ou /t/) comme on peut le constater sur la Figure 2 ci-dessous.

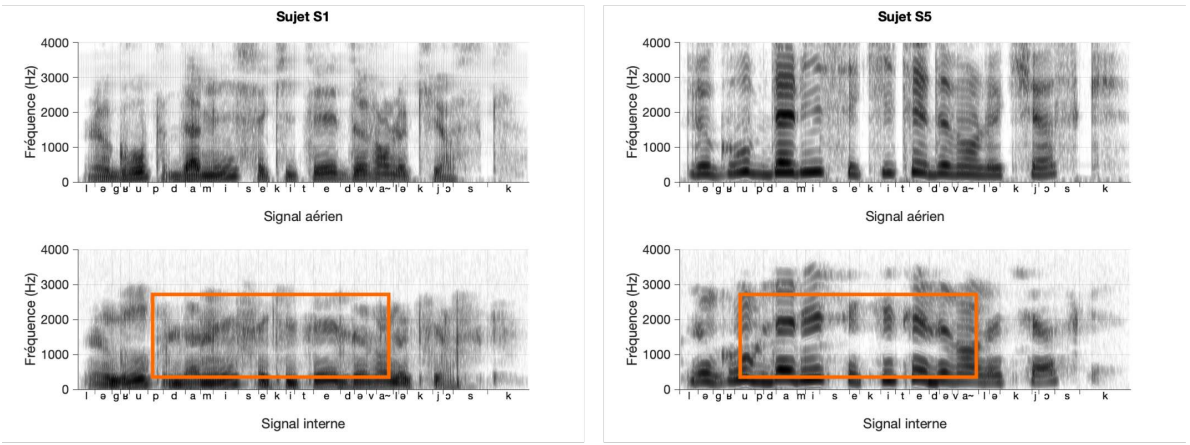

FIGURE 2 – Spectrogrammes de parole chez deux sujets, pour la phrase *"le groupe d’amis s’est quitté devant le kiosque"*. En haut, les spectrogrammes du signal aérien ; en bas les spectrogrammes du signal interne. La trajectoire de F2 est lisible dans le signal interne même pendant la succession de consonnes non voisées de *s’est quitté* (encadré orange).

Bien entendu, des éléments importants sont absents du signal interne ; les hautes fréquences au-delà de 3,5 kHz sont absentes de nos enregistrements, ce qu’on peut attribuer au moins en partie à l’absorption par les tissus mous du conduit vocal (d’autres explications complémentaires sont abordées dans la Discussion). Les voyelles ouvertes sont bien plus marquées dans le signal aérien, et l’essentiel des bruits aéro-acoustiques des plosives et fricatives n’est appréciable que dans ce signal. Pour aller au-delà d’une simple description de surface de ces différences, qui sont également influencées par l’acoustique du conduit auditif externe, nous avons cherché à les quantifier, c’est-à-dire à mesurer

l'information portée exclusivement par l'un ou l'autre signal.

A cette fin, nous avons utilisé la conversion de voix afin de mettre en évidence les parties de l'information sonore d'un signal qui ne sont pas déductibles de l'autre de manière déterministe. En convertissant le signal aérien en signal interne, et en comparant cette approximation au signal interne réel, on peut ainsi apprécier la "valeur ajoutée" du signal interne (les parties du signal interne qui ne sont pas prédictibles connaissant le signal aérien). Réciproquement, on peut convertir le signal interne en signal aérien pour mesurer l'information exclusive au signal aérien — qui reste celui qu'un locuteur cherche à transmettre. Il faut noter que la conversion de voix étant une méthode d'association statistique (régression), elle peut permettre de reconstituer des aspects d'un signal qui sont absents de l'autre, si ces signatures spectrales sont *toujours* associées. Cette méthode révèle donc des différences plus fondamentales que les différences de surface qu'on peut voir sur la Figure 2.

3.2 Conversion de voix

Nous donnons un exemple de l'application de cette technique de conversion à notre question centrale, la mise en évidence des différences entre composantes du retour acoustique de la parole. La Figure 3 illustre le résultat d'une comparaison entre la prédiction du signal aérien à partir du signal interne (panneau du haut), ou réciproquement (panneau du bas), sur une phrase du corpus. L'amplitude des différences entre prédiction et signaux mesurés est figurée par une échelle de couleurs. Dans cet exemple, le signal aérien comporte une information difficile à reconstruire sur la voyelle ouverte /a/, ou le haut et le bas du spectre de la fricative /s/. Réciproquement, le signal interne comporte une information spécifique sur les nasales /õ/ ou /n/. On voit aussi que la séquence /øsəmɛ/ comporte une information sur F2 et F3 difficile à reconstruire à partir du signal aérien.

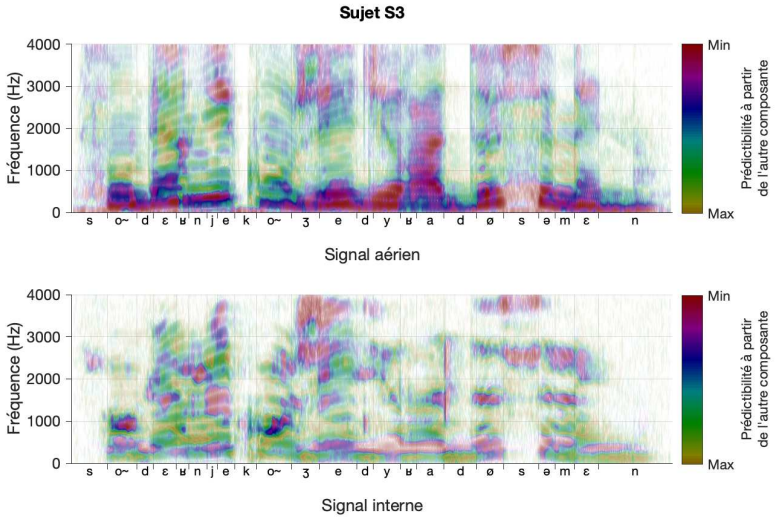

FIGURE 3 – Spectrogrammes de la phrase *"son dernier congé dura deux semaines"*, colorisés selon la facilité avec laquelle on peut les reconstruire à partir du signal de l'autre composante.

Cette analyse permet de révéler quelle parties du spectre d'une composante lui semblent spécifiques. Pour comparer les deux composantes de manière plus systématique, entre phrases et entre locuteurs,

nous avons cherché à mesurer de manière globale l'information spécifique à une composante acoustique. Pour cela, nous avons considéré la différence absolue cumulée entre spectres prédits et réels dans la bande 0–4 kHz. La suite de l'article se fonde sur cette mesure.

3.3 Information propre au signal aérien

L'information propre au signal aérien est donc quantifiée par la différence globale entre spectres du signal aérien réel et de celui obtenu par conversion depuis le signal interne. Cette différence est illustrée Figure 4 sur une sélection de phrases du corpus, pour 5 locuteurs. On constate tout d'abord une homogénéité assez claire entre locuteurs. De manière générale, les voyelles ouvertes comme /a/ ou /ɛ/ comportent une information en partie absente du signal interne, ce qui semble logique. Les fricatives comme /f/, /ʒ/ ou /s/, dont le spectre est très étalé, sont apparemment aussi difficiles à prédire à partir du signal interne.

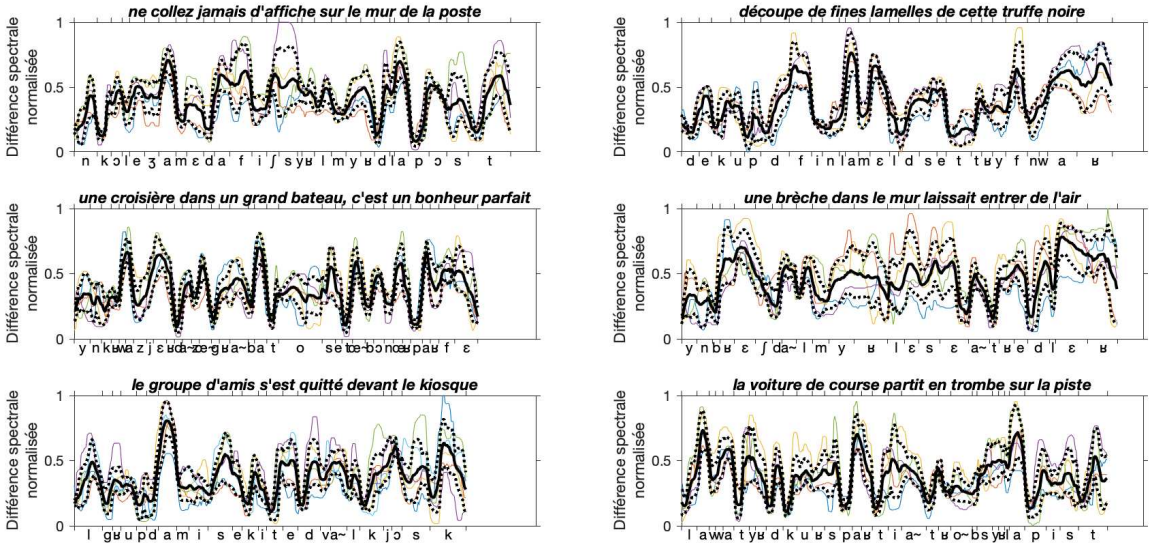

FIGURE 4 – Information aérienne non reconstructible d'après le signal interne. Après alignement temporel, on a représenté la mesure pour chaque sujet (traces colorées), ainsi que la moyenne (trait noir) et le 20ème et 80ème percentiles (pointillé). On voit que le signal aérien comporte une autre information que le signal interne pour les voyelles ouvertes et les fricatives, de manière générale.

Les plosives non voisées /t/, /p/ ou /k/ ne portent une information aérienne spécifique que dans dans la partie terminale de leur bruit aéro-acoustique (on le voit nettement en fin de phrase). De fait, on verra plus bas que la composante interne a tendance à être riche sur ces phones. Toutefois, une certaine variabilité est apparente sur ces consonnes, et particulièrement sur leurs équivalentes voisées, ce qui méritera une analyse plus approfondie, en s'attachant en particulier aux aspects de co-articulation qui peuvent expliquer ces changements dépendants du contexte.

3.4 Information propre au signal interne

La même analyse, conduite en miroir, montre tout d'abord deux caractéristiques principales : 1) une amplitude apparente des différences plus faible ; 2) une variabilité inter-sujets plus grande. Le premier point s'explique en revenant à la Figure 3 : le signal interne, en tout cas celui auquel nous avons accès, est plus limité en bande passante ; par ailleurs, ils apparaît que les différences informationnelles sont souvent limitées à des formants particuliers. La différence globale est donc comparativement plus faible. La variabilité inter-sujets a plusieurs origines possibles sur lesquelles nous reviendrons dans la discussion.

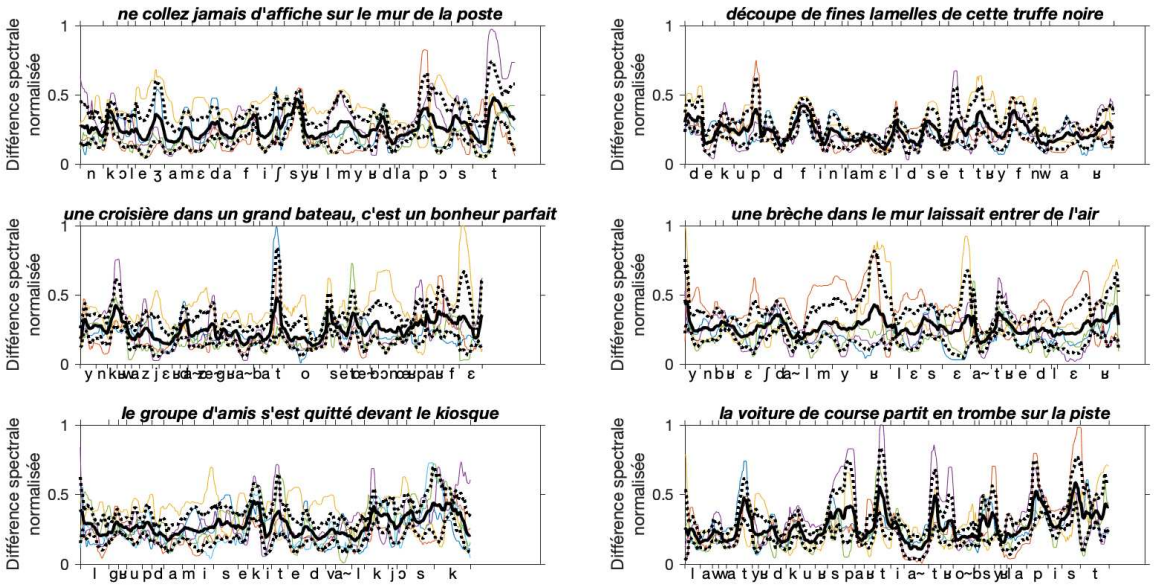

FIGURE 5 – Information interne non reconstructible d'après le signal aérien. On voit que le signal interne comporte une autre information que le signal interne pour les voyelles et consonnes nasales, ainsi que pour les occlusives et les fricatives, de manière générale.

Comme attendu, notre mesure d'information est plus importante sur les voyelles et consonnes nasales. Si cet effet n'est pas très marqué, c'est que la différence entre signaux sur ces phones est souvent limitée au formant nasal (cf. Figure 3). En revanche, comme mentionné plus haut, on remarque des maxima d'information sur les plosives et fricatives, ce qui était initialement moins attendu. En fait, ces consonnes rayonnent également de l'énergie à l'intérieur du conduit vocal, dont les résonances sont alors perceptibles dans le signal interne. On le voit par exemple clairement sur la Figure 2, où la transition formantique /ise/ n'est lisible que dans le /s/ interne. Dans ces instants, les deux composantes du retour auditif sont complémentaires : le bruit aérien de turbulence large bande est absent du signal interne, tandis que la transition formantique correspondant à la cavité arrière est absente du signal aérien.

4 Discussion

Ces résultats sont, à notre connaissance, la première tentative de description des différences entre retour acoustique aérien de la parole naturelle, et retour par conduction interne ou 'osseuse'. Reinfeldt *et al.* (2010) n'avaient décrit que des différences spectrales, et ce, pendant la production d'une série limitée de phones isolés. Par ailleurs, nous avons introduit une méthode de quantification des différences informationnelles entre signaux basée sur une méthode statistique de conversion de signaux de parole. Les résultats obtenus permettent de mieux apprécier l'apport du retour acoustique par conduction osseuse sur la perception et la production de sa propre parole.

Toutefois, nos méthodes ont plusieurs limitations qu'il faut garder à l'esprit. Tout d'abord, le signal 'interne' enregistré n'est pas *stricto sensu* la partie du signal cochléaire provenant de la conduction interne : il ne reflète que la voie externe (vibration du conduit auditif et du tympan), et ne permet pas d'appréhender l'effet de la conduction via les osselets ou par la vibration du rocher. Le filtrage passe-bas de notre signal par les tissus mous ainsi que l'effet d'amplification aux fréquences de résonance du conduit auditif sont probablement moins marqués dans le signal interne réel.

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 corpus d'entraînement (seulement 5 minutes de parole) ; enregistrer un corpus plus important contribuerait à conforter ces résultats. D'autre part, nous avons résumé les différences spectrales en considérant leur somme sur la bande 0–4 kHz, ce qui caractérise les différences brutes, dans un espace linéaire, et ne rend pas compte du caractère éventuellement crucial d'une information sonore particulière à l'intérieur de ce spectre. Dans la suite de ce travail, nous évaluerons d'autres mesures dérivées d'outils statistiques différents.

Enfin, nous avons noté une variabilité inter-sujets importante du retour par conduction osseuse, qui doit nous conduire *a minima* à augmenter nettement la taille de l'échantillon. Cette variabilité a été décrite dans la littérature (Saleeby *et al.*, 1976; Pollard *et al.*, 2017) et correspond en grande partie à des différences morphologiques entre locuteurs (mais à ce stade on ne voit pas clairement émerger de différence attribuable au sexe du locuteur). Toutefois, une part de cette variabilité est possiblement d'origine méthodologique et liée à des différences de placement de notre capillaire d'enregistrement ; cet aspect devra être amélioré.

Malgré ces limites, ces résultats préliminaires nous semblent très encourageants. Au-delà des aspects perceptifs, les particularités du retour acoustique par conduction osseuse pourraient être essentielles au contrôle moteur de la parole. Nos résultats, notamment sur les différences entre composante aérienne et osseuse lors des consonnes, permettent de réapprécier le rôle potentiel du retour acoustique sur le contrôle en temps réel des articulateurs, en complément des entrées proprioceptives et tactiles. En particulier, la conduction osseuse pourrait jouer un rôle plus important qu'estimé dans l'apprentissage des sons de parole chez l'enfant, et dans leurs troubles éventuels, ce qui pourrait suggérer des stratégies orthophoniques spécifiques. De même, l'enfant sourd porteur d'implants cochléaires pourrait tirer bénéfice de l'addition d'un retour acoustique "osseux" dans le signal fourni par les implants. Ces applications thérapeutiques potentielles motivent d'autant la poursuite de nos travaux.

Remerciements

Nous remercions Thomas Hueber pour ses précieux conseils et le partage de son code de conversion de voix. Financements : IDEX NeuroCog [ANR-15-IDEX-02] et ANR [ANR-21-CE37-0017].

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.

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.

GOLDMAN J.-P. (2011). EasyAlign : an automatic phonetic alignment tool under Praat. In *InterSpeech 2011*, Firenze.

HUEBER T. & BAILLY G. (2016). Statistical Conversion of Silent Articulation into Audible Speech using Full-Covariance HMM. *Computer Speech and Language*, **36**, 274–293.

PHUNG N. T., UNOKI M. & AKAGI M. (2012). A study on restoration of bone-conducted speech in noisy environments with LP-based model and Gaussian mixture model. *J Sig Process*, **16**(5), 409–417.

POLLARD K. A., TRAN P. K. & LETOWSKI T. (2017). Morphological differences affect speech transmission over bone conduction. *The Journal of the Acoustical Society of America*, **141**(2), 936–944.

PÖRSCHMANN C. (2000). Influences of bone conduction and air conduction on the sound of one's own voice. *Acta Acustica united with Acustica*, **86**, 1038–1045.

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.

SALEEBY G. C., ALLEN G. D., MAHAFFEY R. B. & WOOD T. J. (1976). Air conduction + bone conduction = your own voice ? *The Journal of the Acoustical Society of America*, **59**(S1), S15–S15.

STENFELT S. & GOODE R. L. (2005). Bone-conducted sound : physiological and clinical aspects. *Otology & neurotology*, **26**(6), 1245–1261.

TODA T. & SHIKANO K. (2005). NAM-to-speech conversion with Gaussian mixture models. In *InterSpeech*, p. 1957–1960, Lisbon.

TONNDORF J. (1966). Bone conduction. Studies in experimental animals. *Acta oto-laryngologica*, **Suppl 213**, 7–133.