



HAL
open science

French Listening Tests for the Assessment of Intelligibility, Quality, and Identity of Body-Conducted Speech Enhancement

Thomas Joubaud, Julien Hauret, Véronique Zimpfer, Eric Bavu

► **To cite this version:**

Thomas Joubaud, Julien Hauret, Véronique Zimpfer, Eric Bavu. French Listening Tests for the Assessment of Intelligibility, Quality, and Identity of Body-Conducted Speech Enhancement. Interspeech 2025, ISCA - International Speech Communication Association, Aug 2025, Rotterdam, Netherlands. <hal-05098132>

HAL Id: hal-05098132

<https://hal.science/hal-05098132v1>

Submitted on 5 Jun 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

French Listening Tests for the Assessment of Intelligibility, Quality, and Identity of Body-Conducted Speech Enhancement

Thomas Joubaud¹, Julien Hauret^{1,2}, Véronique Zimpfer¹, Éric Bavu²

¹Acoustics and Protection of the Soldier, French-German Research Institute of Saint-Louis, France

²Laboratoire de Mécanique des Structures et des Systèmes Couplés, Conservatoire National des Arts et Métiers, France

{thomas.joubaud, veronique.zimpfer}@isl.eu, {julien.hauret, eric.bavu}@lecnam.net

Abstract

This study evaluates the Extreme Bandwidth Extension Network (EBEN) model on body-conduction sensors through listening tests. Using the Vibravox dataset, we assess intelligibility with a French Modified Rhyme Test, speech quality with a MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) protocol and speaker identity preservation with an A/B identification task. The experiments involved male and female speakers recorded with a forehead accelerometer, rigid in-ear and throat microphones. The results confirm that EBEN enhances both speech quality and intelligibility. It slightly degrades speaker identification performance when applied to female speakers' throat microphone recordings. The findings also demonstrate a correlation between Short-Time Objective Intelligibility (STOI) and perceived quality in body-conducted speech, while speaker verification using ECAPA2-TDNN aligns well with identification performance. No tested metric reliably predicts EBEN's effect on intelligibility.

Index Terms: body-conduction sensors, speech quality, intelligibility, speaker identity, MRT, MUSHRA, objective metric

1. Introduction

Remote voice communication in noisy environments requires effective speech capture and restitution. While integrating loudspeakers into hearing protection devices generally addresses the latter, capturing speech remains challenging. Body-conduction sensors, exploiting vocal vibrations through bones and soft tissues, provide a robust alternative to traditional microphones in environments above 75 dB(A) [1, 2, 3]. Despite their noise resilience, these sensors reduce intelligibility due to limited bandwidth. Prior research has focused on enhancing speech captured via bone-conduction transducers [4, 5], in-ear microphones [2, 6, 7], and throat microphones [8, 9]. Recent advances leverage deep neural networks for body-conducted speech enhancement [3, 4, 7, 10]. However, these approaches require extensive training data, but few publicly available datasets exist [11, 12, 13, 14]. The Vibravox dataset [14] addresses this gap by incorporating five body-conduction sensors. Using this dataset, researchers evaluated the Extreme Bandwidth Extension Network (EBEN) model [3] across three tasks: speech enhancement, speech-to-phoneme transcription, and speaker verification. Objective metrics [15, 16] confirm that EBEN improves speech quality, intelligibility, and transcription accuracy, but degrades speaker identity recognition.

Speech enhancement evaluation can rely on a wide range of metrics — 12 in [17], for instance. However, such metrics do not always align with human perception. While advancements like Audiobox [18] push audio evaluation boundaries, listening tests remain the gold standard. This study validates the obser-

vements from [14] through listening tests based on the Vibravox dataset. To keep test durations manageable, we focus on three body-conduction sensors: the forehead accelerometer (Knowles BU23173-000), the rigid in-ear (RIE) microphone [19], and the throat microphone. An airborne headset microphone serves as a reference. This study does not consider external noise because body-conducted sensors are inherently resilient to it. Given that their use is not necessarily limited to continuously noisy environments, our assessments with quiet recordings provide a relevant and representative setting. In the following sections, we evaluate speech intelligibility using a French Modified Rhyme Test (MRT) [20], speech quality via a MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) test [21], and speaker identity preservation with an A/B identification task. Additionally, we analyze results separately for male and female speakers to distinguish low- and high-pitched voices. Finally, we compare listening test outcomes with corresponding objective metrics to assess their suitability for evaluating body-conducted speech signals, both raw and enhanced. Before conducting t-tests, we identify and remove outliers using the interquartile range (IQR) method, discarding values deviating by more than $1.5 \times IQR$ from the first or third quartile. We then verify data normality with the Shapiro-Wilk test [22]. A significance level of 95 % is applied throughout our analysis.

Beyond assessing the impact of bandwidth extension, this study also contributes to a broader discussion on the relationship between objective metrics and human evaluations in speech enhancement. While automated measures are widely used for their efficiency and reproducibility, their ability to reflect perceptual quality remains an open question, particularly for body-conducted speech for which subjective studies are still scarce. By systematically comparing human judgments with metric-based assessments, our work provides insights into which metrics are best aligned with perception in this context. These findings could inform the development of more perceptually relevant evaluation frameworks for future speech enhancement systems.

2. Speech Intelligibility: MRT

2.1. Experimental Protocol

We evaluate the intelligibility of body-conducted speech with the MRT, the standard method for communication systems according to the American National Standards Institute [23]. The French adaptation of this test [20] quantifies consonantal confusion in a closed-response set of 50 lists, each containing six Consonant-Vowel-Consonant (CVC) words. In each list, the words differ by only one consonant. To create our test material, we recruited a male and a female participant to record the full set of 300 words, using the same sensors as in [14] un-

der quiet conditions. These participants, not included in the initial Vibravox dataset, pronounced each target word within a fixed carrier sentence: (*Le mot ... doit être indiqué.*¹). The resulting MRT test data² has been made publicly available on HuggingFace, along with an enhanced version³ processed using the appropriate EBEN models. In this study, we focus on three body-conduction sensors: the forehead accelerometer, the rigid in-ear microphone, and the throat microphone, evaluating both their raw and enhanced signals. Including the reference microphone, this results in a total of seven test conditions. 23 native French speakers participated in the MRT experiment, split into two sessions — one for the male speaker and one for the female speaker. In each session, participants heard 50 sentences (one per MRT list) for each condition. All signals were loudness-normalized to -36 LUFS [24]. After listening, each participant selected the target word from the six-word list. Each session lasted about 30 minutes.

2.2. Results

Figure 1(a) shows the distribution of MRT scores across all conditions. For body-conduction microphones, the average score stays above 80% for raw signals, but never matches the reference microphone’s performance. EBEN has little effect on the forehead accelerometer and rigid in-ear (RIE) microphone, likely due to their high-quality raw signals. However, with the throat microphone, EBEN improves the MRT score by over 5%, demonstrating its effectiveness in this case.

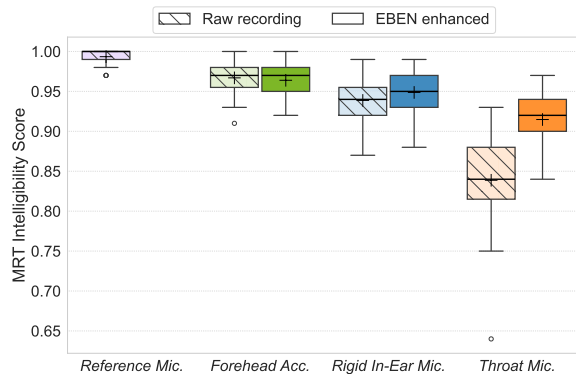
For all listeners, we compute the difference between average MRT scores with EBEN-enhanced and raw signals to measure intelligibility improvement. Figure 1(b) shows the distribution by speaker gender. As noted, EBEN has no effect on the forehead accelerometer but improves the throat microphone performance — by 5% for the male speaker and 10% for the female speaker. For the RIE microphone, there’s a 4% improvement for the male speaker but a slight 2% degradation for the female speaker. Statistical analysis supports these findings. After outlier removal and normality verification, we apply a one-sample t-test for each sensor and speaker to check if the mean intelligibility improvement differs from zero. p-values are shown in Figure 1(b). The female speaker’s raw RIE performance (96%) is higher than the male’s (91%), which may explain the slight degradation after enhancement.

Overall, EBEN enhances intelligibility when the body-conduction sensor introduces significant degradation, as with the throat microphone. Otherwise, its impact is minimal (forehead accelerometer) or slightly negative (RIE for the female speaker). Expanding the dataset with more speakers could confirm these trends.

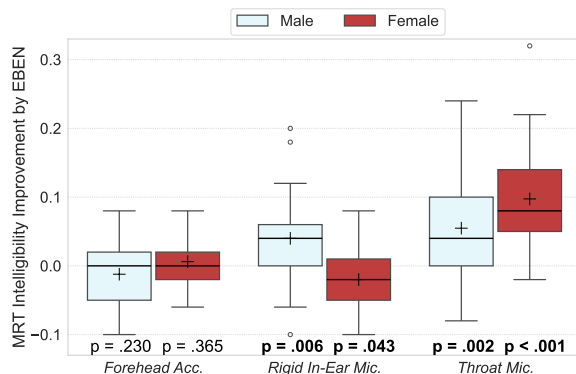
3. Speech Quality: MUSHRA

3.1. Experimental Protocol

We assess speech quality using a MUSHRA test [21] with the same body-conduction sensors and enhancement processing. From the Vibravox test set, we randomly selected 10 sentences from 5 women and 5 men. The headset microphone serves as the reference, and the temple contact microphone acts as a low-quality anchor due to its reduced bandwidth and high back-



(a) Distribution of MRT intelligibility score



(b) Distribution of EBEN-induced intelligibility improvement

Figure 1: (a) MRT intelligibility score for raw and EBEN-enhanced signals. (b) EBEN-induced intelligibility improvement per sensor and speaker gender. (Black cross: mean. p-values from one-sample t-tests for zero mean.)

ground noise. We evaluate raw and EBEN-enhanced signals from the forehead accelerometer, rigid in-ear microphone, and throat microphone, sourced from ⁴. Since EBEN operates at 16 kHz, the raw signals are downsampled accordingly, with only the reference remaining at 48 kHz, with a downsampled hidden version. All signals are loudness-normalized to -36 LUFS [24]. A total of 21 experienced listeners participated in the MUSHRA test. For each of the 10 sentences, they compared signals from the 8 conditions against the reference, rating speech quality on a 0–100 scale (0–20: bad, 20–40: poor, 40–60: fair, 60–80: good, 80–100: excellent). We discarded three participants who rated the hidden reference below 80 in more than 15% of trials. We also ensured that no listener rated the low-quality anchor as excellent in more than 15% of cases.

3.2. Results

Figure 2(a) shows the quality rating distribution for all sensors, comparing raw and EBEN-enhanced signals. Listeners correctly identified the hidden reference as the best. EBEN processing improves quality for body-conduction sensors by roughly 20% on average. However, high variability in MUSHRA scores suggests differences in individual rating strategies. To address this, we compute the quality differ-

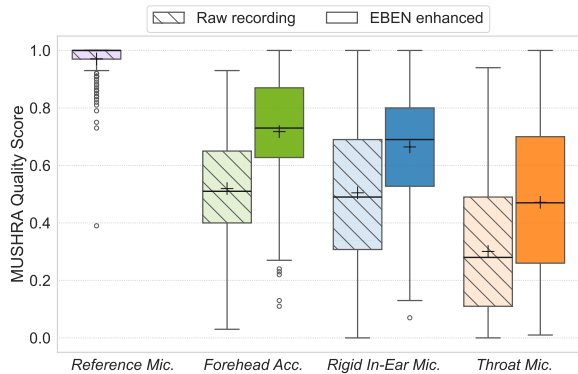
¹In English: *The word ... must be indicated.*

²<https://huggingface.co/datasets/Cnam-LMSSC/french-mrt>

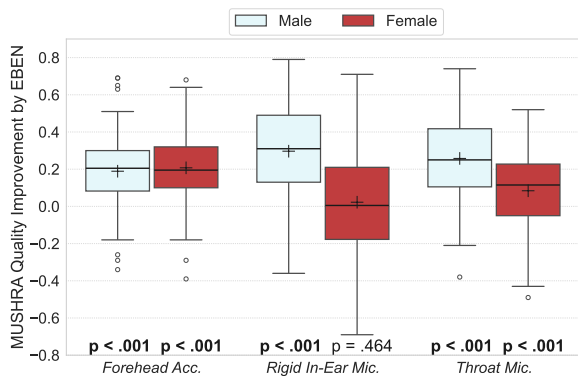
³https://huggingface.co/datasets/Cnam-LMSSC/french-mrt_enhanced_by_EBEN

⁴https://huggingface.co/datasets/Cnam-LMSSC/vibravox_enhanced_by_EBEN

ence between EBEN-enhanced and raw signals for each listener, shown in Figure 2(b). A positive value indicates improvement. With the forehead accelerometer, speech quality increases by 20% on average, for both male and female speakers. For the rigid in-ear and throat microphones, enhancement benefits male speakers more significantly.



(a) Distribution of MUSHRA quality rating



(b) Distribution of EBEN-induced quality improvement

Figure 2: (a) MUSHRA quality score for raw and EBEN-enhanced signals. (b) EBEN-induced quality improvement per sensor and speaker gender. (Black cross: mean. p -values from one-sample t -tests for zero mean.)

Using one-sample t -tests (p -values in Figure 2(b)), we find that the mean quality difference is significantly different from 0 in all conditions except for the RIE microphone with female speakers ($p = .464$). Notably, the raw signals for female speakers are rated as good on average, so the lack of improvement is not impactful. In contrast, raw (and enhanced) throat microphone signals for female speakers are rated as bad (and poor). Overall, EBEN significantly improves the quality of degraded body-conducted signals, though ratings may remain below fair (<40) for severely degraded raw signals.

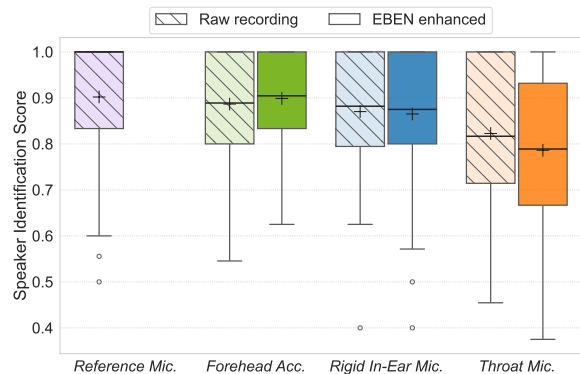
4. Speaker Identity: A/B Identification

4.1. Experimental Protocol

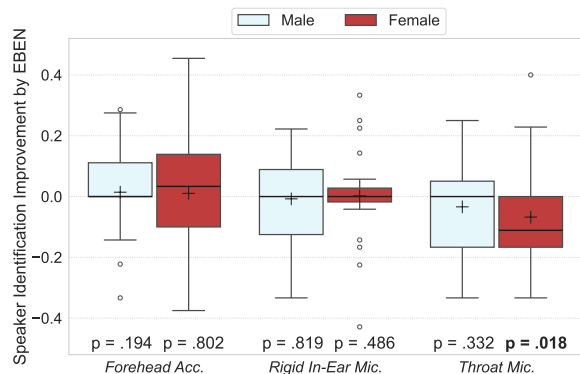
In [14], the authors used a speaker verification model [25] to show that EBEN-enhanced speech alters the Equal Error Rate (EER) compared to raw signals. In their study, they assessed speaker identity by comparing pairs of signals captured with the same sensor, either raw or EBEN-enhanced. To validate

these findings with a listening test, we use an A/B identification approach with the Vibravox *speech-clean* test set. We test the headset reference microphone, forehead accelerometer, rigid in-ear microphone, and throat microphone, in both raw and EBEN-enhanced conditions. For each test step, a sentence is randomly selected from the set, followed by another sentence either from the same speaker or a different one of the same gender. Loudness is normalized to -36 LUFS [24]. After listening to both signals, the listener indicates whether the sentences were recorded by the same speaker. The experiment consists of 100 test steps, with 22 volunteers participating.

4.2. Results



(a) Distribution of A/B identification score



(b) Distribution of EBEN-induced identification improvement

Figure 3: (a) A/B speaker identification score for raw and EBEN-enhanced signals. (b) EBEN-induced identification improvement per sensor and speaker gender. (Black cross: mean. p -values from one-sample t -tests for zero mean.)

Figure 3(a) shows the distribution of identification scores for all sensors and processing conditions. With the reference microphone, half of the listeners achieve a perfect score, and the average score is 90%. Similar performance is observed with the forehead accelerometer and RIE microphone, with no noticeable effect of EBEN processing. For the throat microphone, the mean score is 82% for raw signals, dropping 4% with EBEN.

To further assess the impact of speech enhancement on speaker identification, we compute the difference between raw and EBEN-enhanced scores and perform one-sample t -tests for zero mean. Results, shown in Figure 3(b), indicate no significant effect for the forehead accelerometer and RIE microphone. For male speakers with the throat microphone, there is also no

significant change. However, for female speakers, there is a significant reduction in identification scores, with a mean decrease of -6% (median: -11%).

While the EBEN model was not trained to preserve speaker identity [3], our A/B identification test shows that speech enhancement generally does not affect speaker recognition. However, results also indicate that EBEN may hinder speaker identification in cases when the original signal quality and intelligibility are poor.

5. Comparison with objective metrics

In most studies related to speech enhancement, objective metrics are used to assess improvements without the need for listening tests, making it easier to compare results from different papers using the same datasets. However, the context in which these metrics have been developed may differ from the contexts in which they are applied. Our study leverages this opportunity to better understand the links and correlations between objective metrics and human evaluations for intelligibility, quality, and identity in body-conducted speech enhancement. While previous studies have attempted to align these two approaches, our work aims to provide deeper insights that can benefit the speech processing community by enhancing the understanding of how objective metrics relate to human assessments.

5.1. Intelligibility

The STOI (Short-Time Objective Intelligibility) [15] is one of the most widely used metrics for predicting intelligibility. We therefore compare it to the Articulation Band Correlation MRT (ABC-MRT) [26], adapted for the MRT paradigm, and a speech-to-phone (STP) transcription model. In [14], a Wav2Vec2.0 model [27] was fine-tuned with the Vibravox dataset. For our analysis, we use the model trained with the reference headset microphone. Intelligibility is predicted as $1 - PER$ (Phoneme Error Rate). To simulate the listening test, we computed the three metrics on the same MRT sentence recordings. The French MRT word lists target 17 consonants. We averaged the listening test performance, STOI, ABC-MRT, and STP predictions for each consonant across all raw and EBEN-enhanced sensors. The Pearson correlation coefficient ρ was used to assess the metrics' suitability for predicting MRT results. As shown in Table 1, ABC-MRT is the most correlated metric. However, ρ never exceeds 0.57, highlighting the need for an intelligibility metric specifically designed for body-conducted speech. STOI fails to capture variations across consonants within a recording condition, as noted in [28]. Lastly, STP transcription could improve by focusing only on the MRT word in the recorded carrier sentence.

Table 1: *Pearson correlation coefficients between listening tests and objective metrics for intelligibility, quality, and speaker identity.*

Listening Test	Metric	ρ
Intelligibility	STOI	.52
	ABC-MRT	.57
	1 - PER	.45
Quality	STOI	.87
	PESQ	.81
	N-MOS	.76
Identity	ECAPA2	.90

5.2. Quality

In [3], the authors found that STOI [15] and N-MOS [16] better predicted their MUSHRA test when comparing speech enhancement models. Since STOI focuses on intelligibility, we also include wideband PESQ (Perceptual Evaluation of Speech Quality) [29, 30] in this study. We compute the metrics on the same 10 sentences used in the MUSHRA test for all raw and EBEN-enhanced sensors. The quality section of Table 1 shows the Pearson correlation coefficients between the listening test results and predictions. All values are acceptable ($> .75$), with STOI being the best predictor ($\rho = .87$). Thus, STOI is more suitable for assessing quality than intelligibility in body-conduction sensors. In this study, PESQ is also a good indicator for quality, contrary to the findings of [3]. Further training of N-MOS with body-conducted speech data could improve its predictive accuracy.

5.3. Identity

Similarly to [14], we employ a pre-trained⁵ ECAPA2-TDNN model [25] to extract speaker embeddings of two tested sentences. We then compute the cosine similarity between the embeddings as a prediction that the same speaker pronounced the sentences. For all raw and EBEN-enhanced sensors, we average this metric and the identification results for each of the 21 listeners of the Vibravox test set and separately if the second sentence is from the same speaker or not. The obtained Pearson correlation coefficient of .90 in Table 1 is high, confirming the metric's suitability. However, in [14], the authors found EBEN-induced degradation of speaker identification, which we don't observe in the listening test (except for female speakers with the throat microphone). This difference may stem from human variability in listening tests, which could blur possible EBEN-induced effects.

6. Conclusion

In this study, we conducted listening tests to evaluate the effectiveness of the EBEN model for body-conducted speech enhancement. When the sensor significantly degrades the signal, EBEN improves both quality and intelligibility. If the initial quality is adequate, EBEN generally maintains it, with the only exception being a slight intelligibility reduction with the RIE microphone for the tested female speaker. Moreover, despite not being explicitly trained to preserve speaker identity, EBEN does not significantly affect speaker identification, except in the case of female speakers with the throat microphone. Lastly, we compared the listening test results with popular objective metrics. STOI and the ECAPA2 model proved to be strong predictors for speech quality and speaker identity, respectively. However, the findings suggest that other neural network-based methods, like NORESQA-MOS, would significantly benefit from training on data featuring body-conduction degradation. Additionally, prediction methods for the intelligibility of short MRT words are still underdeveloped and require further refinement to enhance their predictive accuracy for speech enhancement models. Lastly, while this study focuses on EBEN, we believe the findings generalize to a wider class of speech enhancement models. Indeed, EBEN adopts an architecture and training procedure similar to those used in widely recognized models such as Demucs [31], SEANet [4], and MelGAN [32], and shares key design principles with recent neural audio codecs like SoundStream [33], Encodec [34] and Mimi [35].

⁵<https://huggingface.co/Jenthe/ECAPA2>

7. References

- [1] M. McBride, P. Tran, T. Letowski, and R. Patrick, "The effect of bone conduction microphone locations on speech intelligibility and sound quality," *Applied Ergonomics*, vol. 42, pp. 495–502, 2011.
- [2] R. E. Bouserhal, T. H. Falk, and J. Voix, "In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension," *Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 1321–1331, 2017.
- [3] J. Hauret, T. Joubaud, V. Zimpfer, and É. Bavu, "Configurable EBEN: Extreme bandwidth extension network to enhance body-conducted speech capture," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, p. 3499–3512, 2023.
- [4] M. Tagliasacchi, Y. Li, K. Misiunas, and D. Roblek, "Seanet: a multi-modal speech enhancement network," in *Interspeech*, 2020, pp. 1126–1130.
- [5] Y. Li, Y. Wang, X. Liu, Y. Shi, S. Patel, and S.-F. Shih, "Enabling real-time on-chip audio super resolution for bone-conduction microphones," *Sensors*, vol. 23, no. 35, 2023.
- [6] R. E. Bouserhal, T. H. Falk, and J. Voix, "On the potential for artificial bandwidth extension of bone and tissue conducted speech: a mutual information study," in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2015, pp. 1–5.
- [7] H. Park, Y.-S. Shin, and S.-H. Shin, "Speech quality enhancement for in-ear microphone based on neural network," *IEICE Transactions on Information and Systems*, vol. E102.D, no. 8, pp. 1594–1597, aug 2019.
- [8] A. Shahina and B. Yegnanarayana, "Mapping speech spectra from throat microphone to close-speaking microphone: A neural network approach," *EURASIP Journal of Advanced Signal Processing*, vol. 2007, 2007.
- [9] M. T. Turan and E. Erzin, "Enhancement of throat microphone recordings by learning phone-dependent mappings of speech spectra," in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7049–7053.
- [10] J. Hauret, T. Joubaud, V. Zimpfer, and É. Bavu, "EBEN: Extreme bandwidth extension network applied to speech signals captured with noise-resilient body-conduction microphones," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2023, pp. 1–5.
- [11] M. Wang, J. Chen, X.-L. Zhang, and S. Rahardja, "End-to-end multi-modal speech recognition on an air and bone conducted speech corpus," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 513–524, 2023.
- [12] ESMB-corporus. (2021, <https://github.com/elevoctech/ESMB-corporus>) Elevoc simultaneously-recorded microphone/bone-sensor. Accessed on 2023-10-28.
- [13] M. S. Hosain, Y. Sugiura, M. S. Rahman, and T. Shimamura, "Emobone: A multinational audio dataset of emotional bone conducted speech," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 19, pp. 1492–1506, 2024.
- [14] J. Hauret, M. Olivier, T. Joubaud, C. Langrenne, S. Poirée, V. Zimpfer, and É. Bavu, "Vibravox: A dataset of french speech captured with body-conduction audio sensors," *Speech Communication*, vol. 172, p. 103238, 2025.
- [15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.
- [16] P. Manocha and A. Kumar, "Speech quality assessment through MOS using non-matching references," in *Interspeech*, 2022.
- [17] W. Zhang, R. Scheibler, K. Saijo, S. Cornell, C. Li, Z. Ni, J. Pirklbauer, M. Sach, S. Watanabe, T. Fingscheidt, and Y. Qian, "URGENT challenge: Universality, robustness, and generalizability for speech enhancement," in *Interspeech*, 2024, pp. 4868–4872.
- [18] A. Tjandra, Y.-C. Wu, B. Guo, J. Hoffman, B. Ellis, A. Vyas, B. Shi, S. Chen, M. Le, N. Zacharov, C. Wood, A. Lee, and W.-n. Hsu, "Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound," *Meta AI website*, 2025.
- [19] F. Denk, M. Hiipakka, B. Kollmeier, and S. M. A. Ernst, "An individualised acoustically transparent earpiece for hearing devices," *International Journal of Audiology*, vol. 57, pp. S62–S70, 2017.
- [20] V. Zimpfer, G. Andéol, G. Blanck, C. Suied, and T. Fux, "Development of a french version of the modified rhyme test," *Journal of the Acoustical Society of America*, vol. 147, no. EL55, pp. EL55–EL61, 2020.
- [21] "Recommendation ITU-R BS.1534-3 - Method for the subjective assessment of intermediate quality level of audio systems," International Telecommunication Union, 2015.
- [22] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, 1965.
- [23] "ANSI/ASA S3.2-2009 - Method for measuring the intelligibility of speech over communication systems," American National Standards Institute.
- [24] "Recommendation ITU-R BS.1770-5: Algorithms to measure audio programme loudness and true-peak audio level," International Telecommunication Union, 2023. [Online]. Available: <https://www.itu.int/rec/R-REC-BS.1770-5-202311-1>
- [25] J. Thienpondt and K. Demuynck, "Ecapa2: A hybrid neural network architecture and training strategy for robust speaker embeddings," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2023, pp. 1–8.
- [26] S. Voran, "Using articulation index band correlations to objectively estimate speech intelligibility consistent with the modified rhyme test," in *IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.
- [27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.
- [28] T. Joubaud and V. Zimpfer, "Convolutional neural network-based prediction of a french modified rhyme test recorded with a body-conduction microphone," in *18th International Workshop on Acoustic Signal Enhancement*, 2024, pp. 464–468.
- [29] "ITU-T Recommendation P.862 - Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech," International Telecommunication Union, 2001.
- [30] "ITU-T Recommendation P.862.2 - Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," International Telecommunication Union, 2007.
- [31] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Proc. Interspeech*. ISCA, 2020.
- [32] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.
- [33] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.
- [34] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, 2023, featured Certification, Reproducibility Certification. [Online]. Available: <https://openreview.net/forum?id=ivCd8z8zR2>
- [35] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, "Moshi: a speech-text foundation model for real-time dialogue," Kyutai, Tech. Rep., September 2024. [Online]. Available: <http://kyutai.org/Moshi.pdf>