Vibravox: A Dataset of French Speech Captured with Body-conduction Audio Sensors

Julien Hauret, Malo Olivier, Thomas Joubaud, Christophe Langrenne, Sarah Poirée, Véronique Zimpfer, Éric Bavu

Vibravox: A Dataset of French Speech Captured with Body-conduction Audio Sensors

Julien Hauret, Malo Olivier, Thomas Joubaud, Christophe Langrenne, Sarah Poirée, Véronique Zimpfer, and Éric Bavu

***Abstract*—Vibravox is a dataset compliant with the General Data Protection Regulation (GDPR) containing audio recordings using five different body-conduction audio sensors : two in-ear microphones, two bone conduction vibration pickups and a laryngophone. The data set also includes audio data from an airborne microphone used as a reference. The Vibravox corpus contains 38 hours of speech samples and physiological sounds recorded by 188 participants under different acoustic conditions imposed by an high order ambisonics 3D spatializer. Annotations about the recording conditions and linguistic transcriptions are also included in the corpus. We conducted a series of experiments on various speech-related tasks, including speech recognition, speech enhancement and speaker verification. These experiments were carried out using state-of-the-art models to evaluate and compare their performances on signals captured by the different audio sensors offered by the Vibravox dataset, with the aim of gaining a better grasp of their individual characteristics.**

***Index Terms*—Body-Conduction audio sensors, Robust Communication, Speech enhancement, Speech recognition, Speaker verification**

I. Introduction

UNLIKE traditional microphones, which rely on airborne sound waves, body-conduction audio sensors – often referred to as body conduction microphones (BCMs) for simplicity – allow voice signals to be picked up directly from the body, offering potential advantages in noisy environments by greatly reducing the influence of ambient noise on recordings. Although BCMs have been available for decades [1]–[5], the limited bandwidth of speech captured by such sensors has so far restricted their widespread use. However, thanks to two tracks of improvements, this technology could be made available to a wider audience for speech capture and communication in noisy environments.

Research into physics and electronics is improving with some skin-attachable sensors such as [6]–[8]. Similarly to earlier bone and throat microphones, these new wearable sensors detect skin vibration, which is highly and linearly correlated with the acoustic pressure produced by the wearer's voice [9]. Not only do they improve on the state of the art by having superior sensitivity over the vocal frequency range – thereby improving the signal-to-noise ratio – but they also have superior skin compliance, which facilitates adhesion to curved skin surfaces. These new kinds of sensors, just like the previous ones, are however unable to capture the full bandwidth of the speech signal, due to the inherent low-pass filtering of tissues. The manufacturing process also needs to be stabilized, which currently prevents them from being commercially available. In addition, some earbuds, such as the Xiaomi Buds 4 Pro, already include some voice pickup unit thanks to a retro-action microphone integrated into the loudspeaker at the entrance to the ear canal. This exemplifies the significant role that BCMs will likely assume in the future.

Deep learning methods have shown excellent performance in a variety of speech and audio tasks. In this paper, we show that the Vibravox dataset can be used to advance research in three key tasks that would enable overcoming the current limitations of body conduction sensors: bandwidth extension, speech recognition, and speaker verification. Since body conduction sensors exhibit reduced bandwidth transduction efficiency, works such as [10]–[28] demonstrate the ability of deep learning approaches to regenerate mid and high frequencies from low frequency audio content. They have adapted recent deep learning trends to the specific problem of bandwidth extension, also known as audio super-resolution. For robust speech recognition, models such as Whisper [29] or Canary1B [30] have pushed the limits of usable signals. Finally, for speaker verification, approaches such as TitaNet [31], WavLM [32], Pyannote [33] and ECAPA2 [34] now allow a wide range of signals to be used thanks to their increased robustness.

The availability of large datasets is critical to advancing research and development of BCMs. These datasets allowing to train and evaluate deep learning models have been a key missing ingredient in achieving high-quality, intelligible speech with such audio sensors. The assumption that airborne and body-transmitted speech share identical excitation sources and possess a simple transfer function between them being inadequate, producing realistic synthetic data is therefore still a challenging task. In a previous studyn we performed Out of Distribution tests performed on real BCM signals with an EBEN model trained on simulated data [17], showing the importance of using a sufficiently large set of real data for training. Despite their importance, existing datasets still exhibit several notable limitations. Data collection being labor-intensive, the existing BCM datasets often lack the necessary scale and diversity to comprehensively cover the full range of acoustic scenarios encountered in real-world applications. Problems also persist with signal quality, including noise and artifacts. On top of these issues, the variety of sensors used in existing datasets remains limited, making it difficult to generalize results across different recording environments and equipment configurations. The Vibravox dataset is being made available to fill these gaps and stimulate research in the field of speech capture using non-conventional audio sensors.

Julien Hauret, Malo Olivier, Christophe Langrenne, Sarah Poirée, and Éric Bavu are with the Laboratoire de Mécanique des Structures et des Systèmes Couplés, Conservatoire national des arts et métiers, HESAM Université, 75003 Paris, France. e-mail for these authors: name.surname@lecnam.net.

Thomas Joubaud and Véronique Zimpfer are with the Department of Acoustics and Soldier Protection, French-German Research Institute of Saint-Louis (ISL). e-mail for these authors: name.surname@isl.eu

TABLE I
OPEN SOURCE BCM DATASET REVIEW

Dataset	Number of speakers	Number of BCM	Clean speech	Noisy speech	Data augmentation	Language	Text transcript	Phonetic transcript	Download page
ABCS [35]	100	1	2 × 42h	0h	0h	Chinese	✓	✗	GitHub
ESMB [36]	287	1	2 × 128h	0h	0h	Chinese	✗	✗	GitHub
EmoBone [37]	28	1	2 × 19h	0h	0h	English	✓	✗	Avail. upon request
Vibravox (ours) [38]	188	5	6 × 26h23	6 × 1h34	6 × 10h21	French	✓	✓	HuggingFace

II. RELATED WORK

The need for comprehensive and publicly available datasets of own-speech recordings using BCMs, such as those listed in Table I, is critical to the progress of research and development in the field of body-conducted speech capture. These datasets are essential for training and testing deep learning models.

A. Training datasets

Before delving into publicly available datasets, it is worth acknowledging that they have been made available for a relatively short period of time, are in some cases of insufficient size, or are not available in the target language. Consequently, several studies have employed low-pass filtering of high-quality audio to artificially generate body-conducted speech data. A hybrid approach involves the collection of individual transfer functions between a reference microphone and a body-conducted microphone, which is the focus of the Hearpiece database, as referenced in [39]. In their article, M. Ohlenbusch et al. [40] proposed several speech-dependent models of one's own voice transfer function to simulate the degradation induced by body conducted audio recordings. Their results indicate that the transfer function is of course speaker-dependent, but also phoneme-dependent due to the different position of the jaw. Their study brings a significant improvement in the ability to simulate data. Following on from this work, the authors also published [41] to propose speech-dependent data augmentation to compensate for the lack of a large dataset of own speech signals. They were able to reduce the performance gap when testing the model on real signals. However, by analogy with FineWeb-Edu [42], the performance improvement can also come from having a larger and highly curated dataset on which to perform classical supervised training.

A comprehensive literature review of existing, publicly available datasets of BCM-captured speech is essential to gain a full understanding of the subject. This encompasses all forms of capture, including bone, in-ear, and throat sensors. It should be noted that there exists a number of small private datasets [43]–[48], albeit not open-sourced. The few remaining public datasets are listed in Table I, including Vibravox.

In terms of size, the ESMB dataset [36] is the largest, with 128 hours of recorded speech, followed by ABCS [35] (42 hours), Vibravox (38 hours – with post-processing that exclude silent edges of speech segments[1].) and EmoBone [37] (19 hours). Although Vibravox is not the largest dataset, it overcomes several limitations found in existing datasets, such as the limited diversity of audio sensors and the lack of noise recordings. Vibravox addresses these issues by using five different BCMs and including recordings of speech in both noisy and quiet environments. Previous studies indicate that the transmission of external noise through the device is not only influenced by the device itself, but also by individual variance [39] and angle of arrival [49]. For this reason, the Vibravox dataset was recorded in a 3D sound spatializer to sample uniformly the sphere for noise emission.

[1]See section III-E for more details on the audio post-processing performed prior to the release of the public dataset.

B. Speech processing tasks

The body-conducted speech research community has made significant progress in recent years. The accessibility of public datasets such as Vibravox has become increasingly important for the development and improvement of deep learning models in this field. The continued availability of these resources is likely to facilitate further advances and innovations in body-conducted speech technology, for tasks such as speech enhancement, speech transcription, and speaker verification.

1) Speech Enhancement: BCM datasets, paired with airborne speech, are invaluable in the field of speech enhancement. This task refers to the process of improving the quality and intelligibility of speech, often in the presence of background noise. However, speech enhancement on BCM is mainly about bandwidth extension, as the sensors are robust to noise but cannot record high frequencies due to physical constraints. Several approaches have been published recently in this context. Among them, [20, 25, 48] model the long-term dependencies of speech by adapting the non-quadratic complexity variations of the attention mechanism [50]–[52]. Some studies show superior reconstruction metrics compared to EBEN, but as shown in Table 1 of [16], one of the approaches reporting the best metrics is Kuleshov's U-net of [53], which is also the one that performs worst in the MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) analysis. This evidence serves to illustrate the importance of conducting listening tests when comparing reconstructive and generative approaches. Furthermore, EBEN's computational and memory requirements are low, due not only to the limited number of parameters in the generator, but also to the highly strided embeddings facilitated by the Pseudo-Quadrature Mirror Filters (PQMF) representation [17]. In their paper [23], the authors proposed an innovative approach to deal with limited datasets. They developed a method to disentangle a speaker embedding

vector from the hidden representations of the network during the enhancement process, which allowed a fair generalization performance with a limited training set of 5 hours of 29 speakers and additional synthetic data. Conversely, the approach proposed in [25] involves pre-training with simulated data and fine-tuning on a device with real data, thereby creating a speaker-dependent speech enhancement model.

2) Speech-to-text: Large datasets of transcribed speech are essential for training models to accurately recognize and transcribe spoken words. The inclusion of speech in both noisy and quiet environments in the Vibravox dataset is particularly beneficial for this purpose, as it allows for the development of models that can perform well in a variety of acoustic conditions. To provide context, [35, 54] have previously addressed this task using this particular data modality. They have shown that the body-conducted speech can improve the performance of a basic speech recognition system in adverse environments solely based on airborne speech.

3) Speaker verification: Finally, speaker verification is a biometric authentication method that uses a person's voice to confirm their identity. As with all the other tasks, the system is less affected by reverberation and external noise, reducing the problem of domain mismatch and enabling higher recognition accuracy in challenging environments. Another benefit that is not immediately apparent is the amount of information available. Because everyone's skull and timbre are unique, the recorded signal is more descriptive of a person's identity. This wealth of biometric information opens up several potential applications, including those in the military domain as detailed in [55]. One such application is access control, where it can fortify security by adding an extra layer compared to conventional methods like keycards or codes. Another compelling application lies in intelligence gathering. By pinpointing known voices in intercepted enemy transmissions, speaker identification can reveal the involvement of high-value targets. Beyond this paper's focus, this rich biometric data (heart rate, breathing) could be a game-changer in battlefield forensics, aiding revealing emotional states and tracking life signal during critical events.

III. Building Vibravox

Creating the Vibravox BCM dataset involved a number of different tasks. These included designing pre-amplification and conditioning circuits for each transducer, coding the front-end and back-end software for the recording user interface, adapting the housings and fixings for some audio sensors to the human body through 3D modelling and printing, and selecting transducer technologies and recording parameters. For the sake of brevity, this section will only discuss the essential parts needed to gain a clear understanding of the collected data. All other details can be found at the following URL: https://vibravox.cnam.fr/.

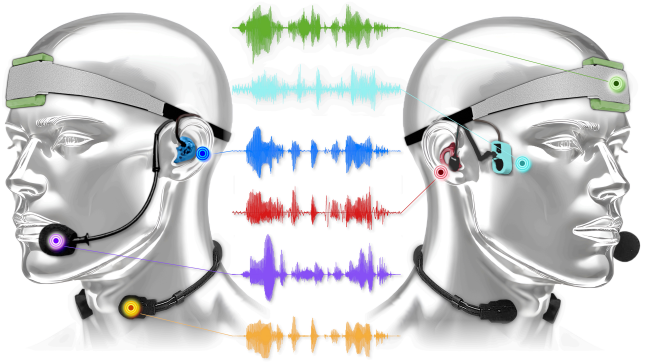


Fig. 1. Fully equipped participant with the six audio sensors

A. Hardware

A comprehensive list of all BCMs employed in the Vibravox dataset is presented in Tab. II. Their positioning on the participants' heads is shown in Figure 1. Although not depicted in Figure 1, a custom 3D-printed backpack and headset were also designed with the objective of facilitating the optimal placement of sensors, carrying the Zoom F8n field recorder, and routing audio cables to that device. The Zoom F8n records all six sensor signals simultaneously, operates at a sampling rate of 48 kHz and a resolution of 32 bits, thereby guaranteeing precise signal capture with high fidelity. To enhance audio data quality, a 20 Hz high-pass filter is uniformly applied to each acquisition channel, effectively attenuating low-frequency noise and artifacts.

In order to provide a balanced comparison between body conduction and airborne audio sensors in terms of immunity to ambient noise, we selected an airborne microphone mounted on a headset that presents a cardioid directivity pattern towards the speaker's mouth, hence producing a dry reference. The in-ear soft foam-embedded microphone is a prototype designed by the ISL and used in [56]. The in-ear rigid earpiece-embedded microphone was initially presented in [57] as a versatile research earpiece with multiple drivers and microphones. In this study, we use the occluded version of this device, and only record the in-ear microphone signal. The laryngophone has been purchased from IXRAdio. This device consists of dual piezoelectric sensors on an adjustable neck rim, which captures skin vibrations from vocal cords and targets voice amplifiers, recorders, and Voice over Internet Protocol (VoIP) applications. The forehead accelerometer, affixed via a homemade headband, is a compact contact sensor that has been commercialized by Knowles. It is designed for use in high-noise environments and can be integrated into helmets for military and emergency service communications. Finally, the temple vibration sensor is an AKG product, a miniature vibration sensor originally designed for string instruments. In this study, this sensor has been diverted from its original purpose to capture bone vibrations by placing it in an adjustable 3D-printed enclosure adapted to the participants' zygomatic bones. In addition to the sensors specifications, Table II provides the median signal-to-noise ratio (SNR) of each sensor. This was computed as the difference between the

TABLE II
SENSORS SPECIFICATIONS

Name	Location	Technology	Reference	Median Signal-to-noise ratio
Headset microphone	Close to mouth	Cardioid electrodynamic microphone	Shure WH20XLR	36.0 dB
Forehead accelerometer	Forehead	One-axis piezeoceramic accelerometer	Knowles BU23173-000	23.7 dB
In-ear soft foam-embedded microphone	Left ear	Omni. electret condenser microphone	Knowles FG-23329-P07	22.8 dB
In-ear rigid earpiece-embedded microphone	Right ear	Omni. MEMS microphone	Knowles SPH1642HT5H	29.0 dB
Throat microphone	Larynx	Piezoelectric	Ixradio XVTM822D-D35	41.1dB
Temple vibration pickup	Temple	Figure of-eight condenser transducer	AKG - C411	9.5 dB

$L_{50}^{silence}$ and L_{50}^{speech} values, that will be defined in Section III-E, across the entire database. Note that all sensors had high SNR, except for the temple vibration sensor, which is typically used to measure much higher amplitude signals.

B. Textual data

The text utilized for participant readings originates from the French Wikipedia subset of Common Voice [58]. This textual data was further filtered to produce a simplified dataset with a minimum number of textual and phonemic symbols and to minimize pronunciation uncertainty. To create this simplified dataset, we applied textual filters that excluded most proper names, retaining only common first names, French town or region names, and country names. Moreover, utterances containing numbers, Greek letters, mathematical symbols, or syntactic errors were carefully removed, resulting in a more straightforward dataset that is suitable for accurate pronunciation analysis. Once the unwanted sentences were removed, we applied common text normalization operations, such as lowercasing text and removing ponctuation except for apostrophe. From this normalized text, we also generated the corresponding phonetic transcription using the International Phonetic Alphabet (IPA), which contains 33 phonemes in French. This operation was done with the phonemizer of the CoML lab [59]. For some words, the phonemizer produced some language switches that we wanted to avoid in order to preserve the minimal French phonetic alphabet. A phase of manual phonetic transcription was then necessary for these unrecognized words, such as ”algorithme”.

C. External ambiant noise data

The Vibravox contains 4 subsets : *speech-clean*, *speech-noisy*, *speechless-clean* and *speechless-noisy*. For the *speech-noisy* and *speechless-noisy* subsets, audio recordings were obtained from sensors worn by participants (either speaking or remaining silent) during the broadcast of 3D spatialized ambient noise. For all these ambient noise samples, the spatialization process was carried out using the `ambitools` library[2] [60] and rendered by a spherical array with a radius of 1.07 meters, made up of 56 loudspeakers placed around the participants [61]. A significant proportion of the noise excerpts were sourced from the Audio Set [62], with a particular focus on noise classes relevant to BCMs applications, such as those encountered in battlefield and industrial environments. The Audio Set noise samples were spatialized using a plane wave model and from random fixed positions uniformly distributed on the sphere. In addition, in-house recordings of concerts and street events in 3rd and 5th order ambisonics were also rendered around the participants.

[2]available at https://github.com/sekisushai/ambitools

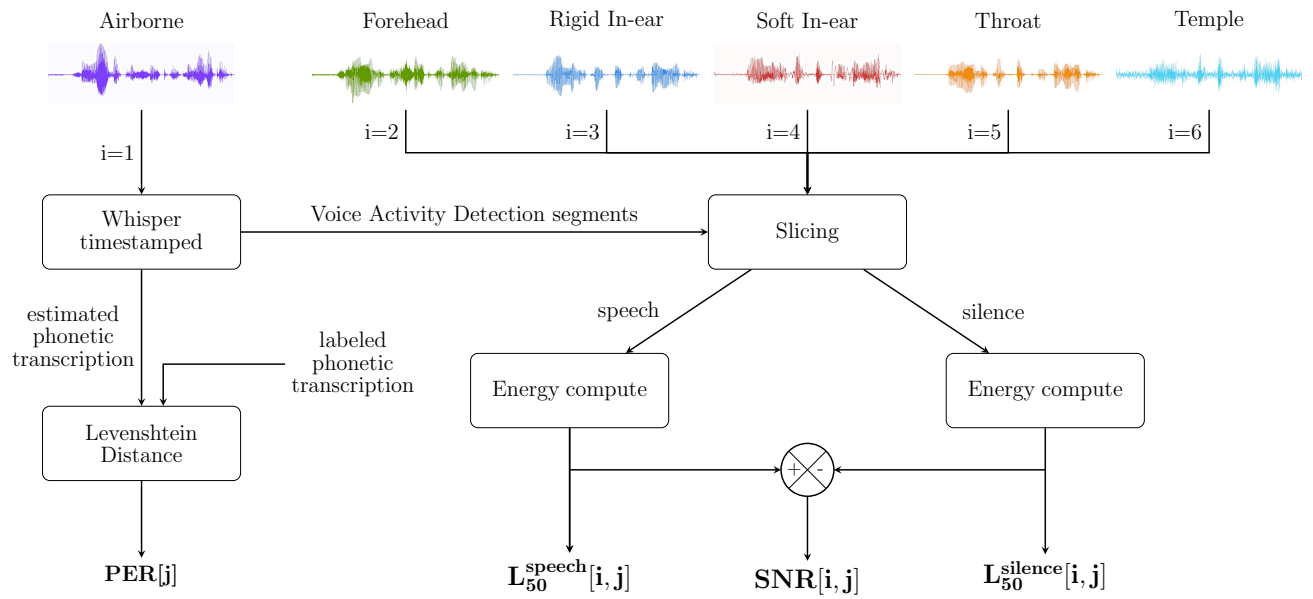
D. Recording protocol

The recording process for each participant involves four steps, each corresponding to a subset of the Vibravox dataset.

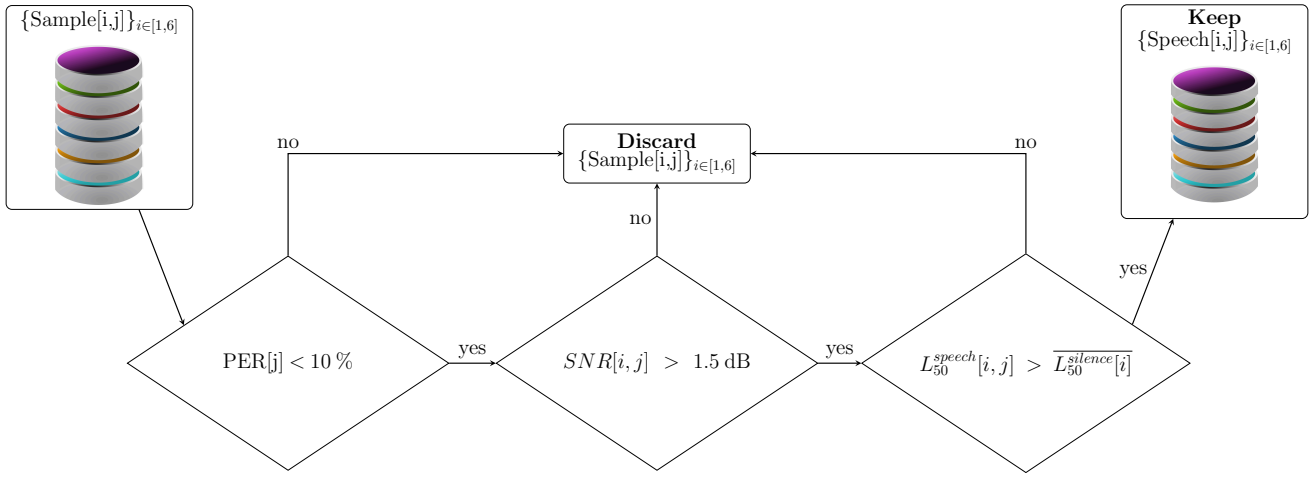
- *speech-clean*: The participant reads sentences sourced from the French Wikipedia for a duration of 15 minutes. Each utterance generates a new recording, and the transcriptions are retained. The recordings from this step populate the Vibravox subset that contains the most data for training.
- *speechless-noisy*: For a period of 2 minutes and 24 seconds, the subject is required to remain silent, yet free to move, swallow and breathe naturally in a noisy environment created from the AudioSet samples described in subsection III-C. Those samples are played in a spatialization sphere equipped with 56 loudspeakers surrounding the subject. The objective of this phase is to gather realistic background noises that will be combined with the *speech-clean* recordings to maintain a clean reference.
- *speechless-clean*: The procedure is repeated for 54 seconds in complete silence to record solely physiological and audio sensor noises. These samples can be valuable for tasks such as heart rate tracking or simply analyzing the noise properties of the various sensors. It could also be conveniently used to generate synthetic datasets with realistic physiological (and sensor-inherent) noise captured by body-conduction audio sensors.
- *speech-noisy*: The final phase (54 seconds) will serve to test the different systems (speech enhancement, automatic speech recognition, speaker verification, ...) that will be developed based on the recordings from the first three subsets. This real-world testing will provide valuable insights into the performance and effectiveness of these systems in practical scenarios. In this phase, the noise samples replayed in the 3D spatializer are sourced from in-house ambisonic recordings of concerts and street events.

E. Post processing

A total of J=33,948 sentences were initially collected from 200 participants. However, a small number of audio clips had various shortcomings: although we carefully monitored



(a) Extracting indicators



(b) Filtering process

Fig. 2. Post-processing filtering process

and validated each recording during the experimental process, the workflow was not entirely foolproof. Mispronounced sentences, sensors shifting from their initial positions, or more significant sensor malfunctions occasionally occurred. In instances where sensors were functional but not optimally positioned, such as when the participant's ear canal was too small for the rigid in-ear microphone to achieve proper acoustic sealing, we elected to retain samples in order to enhance the robustness of our models against the effects of misplaced sensors.

In order to address occasional shortcomings and to offer a high-quality dataset, we implemented a series of three automatic filters to retain only the best audio from the speech-clean subset. We preserved only those sentences where all sensors were in optimal recording condition, adhering to predefined criteria described in Figure 2.

The filtering process comprises two main stages. In the first stage, indicators for evaluating audio capture quality are extracted (Figure 2 a). In the second stage, these indicators are used to derive a set of filters (Figure 2 b). The whisper-timestamped model[3] constitutes a key component of our indicator extraction process. This model is a speech-to-text engine based on Whisper [29] that applies dynamic time warping [63] to cross-attention weights. By aligning the transcription with the raw waveform, accurate prediction of words timestamps in speech segments can be used to identify sequences containing vocalizations. The indicators we are using are listed below:

- $PER[j]$: Phoneme Error Rate are computed for each sample $j \in [1, J]$ using the Levenshtein distance between the phonetic labeled transcription and the phonetic transcription derived from Whisper's output. Note that we used the exact same phonemization procedure as described in subsection III-C.
- $L_{50}^{silence}[i,j]$: In acoustics, percentile levels are statistical measures used to describe the distribution of sound levels

[3]available at https://github.com/linto-ai/whisper-timestamped

over a specified period [64]. These levels indicate the sound level that is exceeded for a certain percentage of the time. In this case, $L_{50}^{silence}[i,j]$ represents the sound level (in dB) exceeded 50% of the time during non-voiced segments of the sample j recorded on sensor i. The "silence" segments are obtained by running the whisper-timestamped model on the headset microphone signals, which are considered as the reference in quiet conditions. The fluctuating sound levels for calculating L_{50} are determined using windows of 4092 samples at the original 48kHz sampling rate. Using $L_{50}^{silence}[i,j]$, $\overline{L_{50}^{silence}[i]}$ is also derived, which represents the mean L_{50} level for silence recorded by each sensor i.

- $L_{50}^{speech}[i,j]$: The sound level exceeded 50% of the time (in dB) during voiced segments of the sample j recorded on sensor i. The same processing as $L_{50}^{silence}[i,j]$ was used.

The aforementioned indicators are then employed in three filters. The first filter discards any audio samples where $PER[j] > 10\%$. This addresses potential discrepancies between the labeled transcription and actual pronunciation, ensuring high-quality labels for the speech-to-phoneme task.

The second filter is employed to verify that the sensor is functioning correctly, specifically by examining the ratio between speech and silence energy levels on a given sample. This ratio, denoted by $SNR[i,j] = L_{50}^{speech}[i,j] - L_{50}^{silence}[i,j] > 1.5dB$ is indicative of recordings with low vocal energy, or those that have been affected by sensor malfunction.

Finally, the last filter is designed to detect any potential sensitivity drift on the sensors. Such drift could be caused by a bug in the electronics or mechanical blockage of the transducer. To this end, the filter checks that $L_{50}^{speech}[i,j] > \overline{L_{50}^{silence}[i]}$.

Only *speech-clean* samples that pass the three filters are added to the Vibravox dataset. We systematically added all audio recordings for the three other subsets (*speech-noisy*, *speechless-clean*, *speechless-noisy*) if the corresponding *speech-clean* subset of the participant is not empty. Of the 200 participants, we kept 188 participants – due to two major sensor malfunctions – and 28,471 sentences.

Note that if a sample j passes all filters, it is not added directly to the dataset. Instead, we extend the timestamps generated by Voice Activity Detection (VAD) by 0.6 seconds on both sides. If the new start or end timestamps are outside the original audio (which is rare), we crop the audio by 15 ms on the corresponding side. This method helps eliminate mouse clicks at audio boundaries and increases the likelihood of capturing vocal segments without inadvertently excluding valid speech portions.

F. Signal processing analysis

The *speech-clean* subset of the Vibravox dataset provides an ideal resource for investigating the differences between various audio sensors. Figure 3 illustrates the coherence functions of all BCMs, based on the source-filter model, where the headset microphone serves as the input and the corresponding BCMs serve as the output. These coherence functions were averaged on the entire filtered dataset and computed using 2,048 frequency bins on the raw signals recorded at 48kHz.

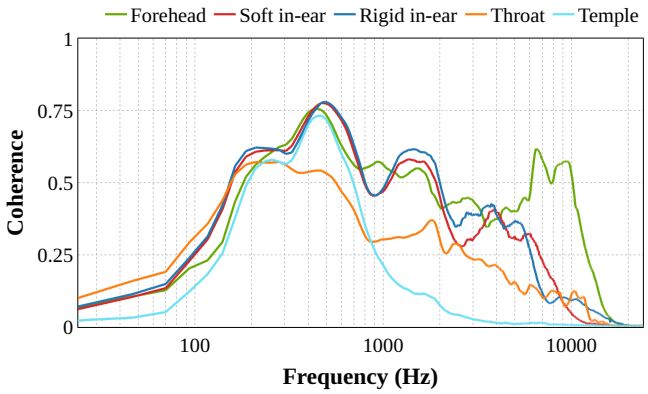


Fig. 3. Coherence function for each body-conducted sensor

Each sensor reveals distinct bandwidth size that the figure legend follows from left to right in decrasing order. The temple sensor exhibits the smallest bandwidth, barely recording any speech signal above 1500 Hz. The laryngophone follows, with coherence dropping below 0.25 at 2000 Hz. The soft and rigid in-ear sensors exhibit larger bandwidths, yet display clear antiresonance frequencies. The soft in-ear bandwidth is slightly larger than its rigid in-ear counterpart, which may be due to a less effective acoustic seal. Finally, the forehead accelerometer offers the largest bandwidth, yet does not effectively filter out external noise pollution. We will see below (Section IV) that this bandwidth is highly correlated with the performance of deep learning models on speech intelligibility, quality, speaker verification, and speech-to-phonemes.

Beyond bandwidth analysis, we can also observe that the shape of these coherences is not always flat, depending on the sensor placement. We cannot draw clear conclusions here because we also changed the technology at each placement, but some works like [65] have already explored the question of the best body-conducted microphone location. Their findings show that the forehead and temple are the best spots to capture speech with the highest intelligibility and sound quality, while the throat is among the worst positions.

IV. Using Vibravox

The Vibravox dataset offers considerable potential across various applications, particularly in improving communication in noisy environments. The *speech-noisy* subset is of particular interest for showcasing the noise-resilient capabilities of BCMs. However, this subset primarily serves testing purposes as it lacks a clean reference under such conditions. As a consequence, the majority of the dataset is contained within the *speech-clean* subset, where the airborne microphone serves as the reference. In order to simulate real-world noisy scenarios while retaining a clean reference, data augmentation techniques can be applied by adding audio segments of the *speechless-noisy* subset with BCM utterances. Furthermore, the Vibravox dataset displays potential for applications such as developing voice commands and monitoring vital signs,

including heart rate, particularly when utilising the *speechless-clean* subset. In the following, we propose three classic tasks in automatic speech processing to investigate the distinctive characteristics of BCMs, establishing baselines for each task.

A. Speech Enhancement

The high impedance of BCMs enables them to focus on the wearer's voice by rejecting ambient noise, but this also affects their ability to capture mid and high-frequency components due to the intrinsic low-pass characteristics of the biological pathway of vibrations between the vocal chords and the body conducted audio sensors. The absence of these frequencies in captured speech directly impacts the perceived quality, intelligibility, and distinctiveness. This highlights the necessity for a post-capture enhancement algorithm. For this purpose, we adopted the configurable EBEN [17] approach, enriched with recent trends in the audio field. This GAN [66] architecture offers several advantages. In particular, the lightweight design of the 1D U-net-like generator, which is compatible with integration in mobile and wearable systems where memory is limited and real-time is a priority. Additionally, the PQMF bank [67] offers the advantage of manipulating frequency content without the need to consider phase or manage complex values, as it operates directly on waveform signals. It is important to note that the limitations of GANs, including the difficulty of reaching a Nash equilibrium, the lengthy training period, and the high computational cost associated with the discriminators, do not affect the model once it has been deployed.

1) Architecture configuration: A separate EBEN model was trained for each sensor in the Vibravox dataset, with specific hyperparameters per sensor employed. The original number of subbands was maintained to a value of $M = 4$ for each sensor due to the significant variability in intra-sensor bandwidth among participants, rendering precise spectrogram slicing impractical. Furthermore, a smaller number of subbands allows for a more permissive design of the passband filters with a softer slope for the same resulting aliasing, thus shortening the kernel length, which is useful for reducing the algorithmic latency. The number of subbands provided to the model, denoted as P, was tailored to match the bandwidth of each sensor. For example, the temple vibration pickup captures signals within the 0-2 kHz range, while the forehead accelerometer exhibits good coherence with the reference microphone up to 8 kHz. Consequently, after resampling the signals to 16 kHz for training, only the first band was retained for the temple vibration pickup, while all bands were retained for the accelerometer. In addition, the last configurable parameter, Q, representing the number of bands fed to the PQMF discriminator, was set to match M at 4 for all models. It was observed that even when the audio sensors captures information within a specific frequency range, including the corresponding subbands in the discriminator input is beneficial for denoising minor artifacts present in these initial bands, which is often not observed with simulated degradations. This minor alteration necessitated an adjustment to the number of channels in the PQMF discriminators, in order to align with the divisibility constraints of the grouped convolution.

2) Training procedure: In comparison to the original implementation proposed in [17], we have incorporated a spectral loss $\mathcal{L}_{\mathcal{G}}^{spec}$ into the generator objective. This modification proved to be effective in stabilizing the training process. This loss incorporates the same FFT, hop size, and window length parameters as described in [68]. Implementation is taken from [69]. In contrast to some speech enhancement papers [70, 71], we did not include a reconstructive loss in the temporal domain due to the peculiarities of body-conduction speech. In fact, complex phase shifts occur, and attempting to align the signals to the reference at the sample level degraded performance. We retained the original EBEN implementation's use of the other two adversarial and a feature matching losses, respectively noted $\mathcal{L}_{\mathcal{G}}^{adv}$ and $\mathcal{L}_{\mathcal{G}}^{feat}$. Also note that a slight adjustment to $\mathcal{L}_{\mathcal{G}}^{feat}$ has been done by normalizing it using the enhanced feature norm. The training strategy remained consistent across experiments, with a constant learning rate of $3e^{-4}$ for 500 epochs on the Vibravox dataset. Additionally, we applied light data augmentation to compensate for the relatively small amount of data typical in deep learning contexts. This data augmentation is composed of speed perturbation, pitch shifting and time masking implemented with the Torchaudio library [72]. To ensure reproducibility of the experiment, the code has been made available on the GitHub repository https://github.com/jhauret/vibravox, which makes use of the following libraries [73]–[76].

3) Results: The correlation analysis presented in the configurable EBEN article [17] revealed that STOI [77] and Noresqa-MOS (N-MOS) [78] were the metrics with the highest correlation with MUSHRA studies in the context of BCMs speech enhancement. Consequently, these two metrics were retained for inclusion in Table III which presents the results of the bandwidth extension for each sensor. The results reveal a clear relationship between the metrics value and the sensors' available bandwidth. Yet, both STOI and N-MOS performance appear to hit a glass ceiling, plateauing at 0.88 and 4.3 respectively. This saturation could stem from the challenge of extracting fine details from body-conducted signals. While striving to produce authentic samples, the generator only picks one plausible clean signal out of all the possible ones, typical of one-to-many problems [79].

B. Speech-to-Phonemes

Automatic Speech Recognition (ASR) systems serve as crucial facilitators of seamless human-computer interaction, allowing users to dictate text, control devices, and access information through spoken language. Proficiency in this realm with BCMs can confer significant advantages, particularly in military operations or industrial settings.

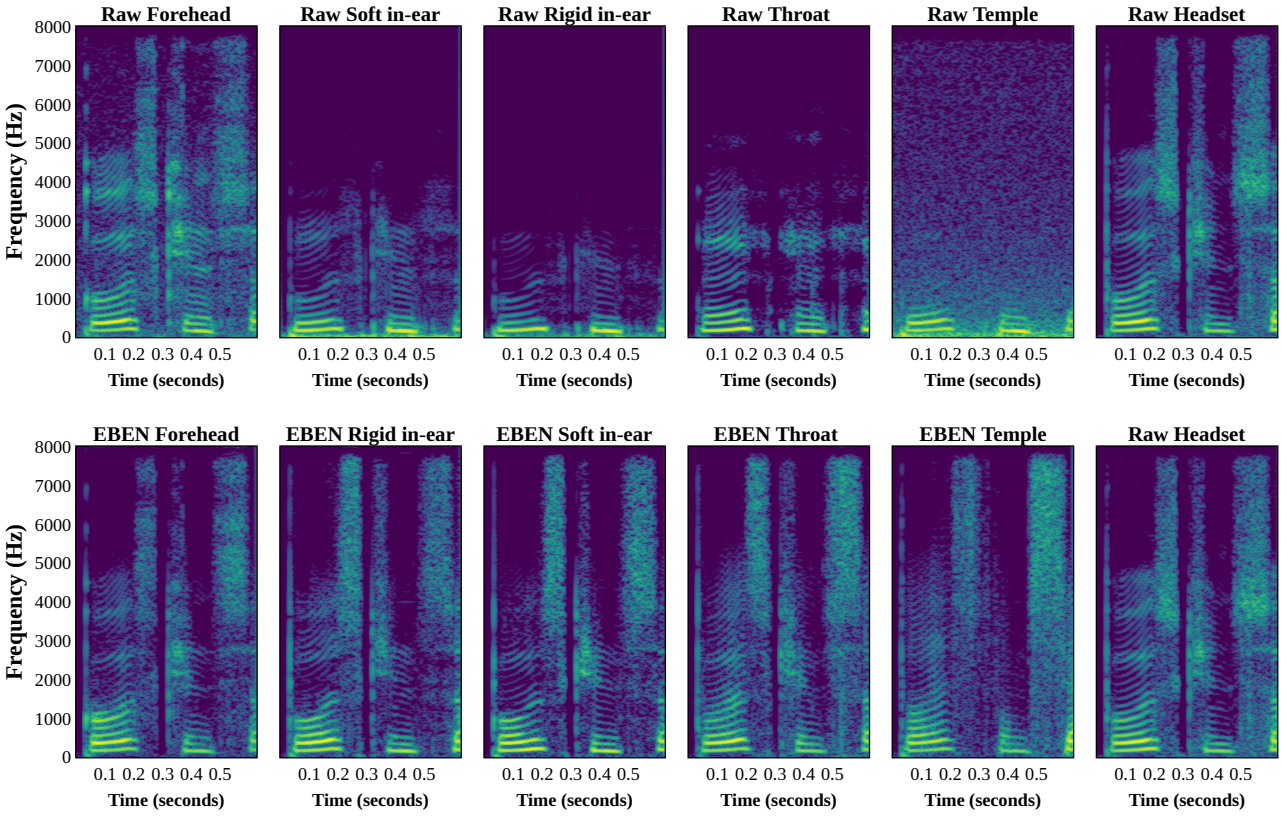


Fig. 4. Spectrograms of signals recorded by the different Vibravox audio sensors and their corresponding EBEN enhanced version

TABLE III
IN DISTRIBUTION TESTING OF EBEN MODELS AND RAW SIGNALS

Sensor	Configuration	STOI	Noresqua-MOS
Forehead	Raw	0.731	3.760
	EBEN (M=4, P=4, Q=4)	0.855	4.250
Rigid In-ear	Raw	0.782	3.392
	EBEN (M=4, P=2, Q=4)	0.877	4.285
Soft In-ear	Raw	0.752	3.315
	EBEN (M=4, P=2, Q=4)	0.868	4.331
Throat	Raw	0.677	3.097
	EBEN (M=4, P=2, Q=4)	0.834	3.862
Temple	Raw	0.602	2.905
	EBEN (M=4, P=1, Q=4)	0.763	3.632

1) Model selection: ASR is often divided into two models as in [80], an acoustic model that infers the tokens from the raw waveform and a language model that is used to refine the probabilities given by the acoustic model with a beam search algorithm. However, our focus primarily revolves around investigating the acoustic characteristics of various sensors. Hence, we have opted to exclude the language model and instead predict phonemes directly with greedy decoding of the acoustic model. The tokens predicted by the linear head of our model correspond to the 33 phonemes of the French phonetic alphabet.

2) Training procedure: Our acoustic model is a medium wav2vec2.0 model [80], pretrained on the multilingual VoxPopuli speech corpus [81]. The finetuning strategy consists of a constant learning rate of $1e^{-5}$ on the Vibravox dataset for 10 epochs, minimizing the CTC loss [82].

3) Results: We tested each fine-tuned model not only on its corresponding test set but also on every other audio sensors, with the PER (Phoneme Error Rate) results shown in Figure 5. As expected, the matrix shows dominance along the diagonal, which corresponds to in-distribution tests. From this diagonal, we infer that the temple vibration pickup and the throat microphone pose the greatest challenge for ASR, given their limited bandwidth. A comparison between models trained on airborne and temple vibration sensors reveals a significant trend: the temple vibration pickup demonstrates enhanced robustness when tested on other sensors data. This trend can be attributed to the model's reliance primarily on low-frequency information, which are present on all sensors. With regard to the in-ear microphones, there is a possibility that a single model can be developed in future studies. This is because they are highly similar and exhibit minimal performance loss when cross-tested. At last, the performance of the forehead accelerometer is approaching that of the in-ear microphone, despite the larger bandwidth. This may be attributed to the considerable variance in the coherence function and the lower signal-to-noise ratio.

A supplementary analysis of the PER has been performed, in order to quantify the effect of the bandwidth enhancement from EBEN models presented in Subsection IV-A for ASR.

TABLE IV
TOP FIVE EDITOPS COUNT FOR IN-DISTRIBUTION TESTING OF PHONEMIZERS ON RAW SPEECH SIGNALS

Sensor	1st editop	2nd editop	3rd editop	4th editop	5th editop
Headset Microphone	[o] → [ɔ] (98)	[e] → [ɛ] (75)	[ɑ] → [ɔ] (64)	[ɔ] → [o] (64)	[∅] → [l] (52)
Forehead Accelerometer	[∅] → [l] (124)	[ɔ] → [o] (106)	[ɛ] → [e] (92)	[o] → [ɔ] (80)	[∅] → [ɛ] (74)
Soft In-ear Microphone	[ɛ] → [e] (127)	[ɔ] → [o] (115)	[∅] → [l] (93)	[ɑ] → [ɔ] (68)	[∅] → [ɛ] (66)
Rigid In-ear Microphone	[ɛ] → [e] (122)	[∅] → [l] (103)	[ɔ] → [o] (97)	[o] → [ɔ] (93)	[∅] → [j] (84)
Throat Microphone	[∅] → [l] (160)	[∅] → [ɛ] (148)	[∅] → [i] (129)	[ɔ] → [o] (128)	[∅] → [ʁ] (107)
Temple Vibration Pickup	[∅] → [s] (294)	[∅] → [l] (290)	[∅] → [t] (281)	[∅] → [ʁ] (281)	[∅] → [i] (229)

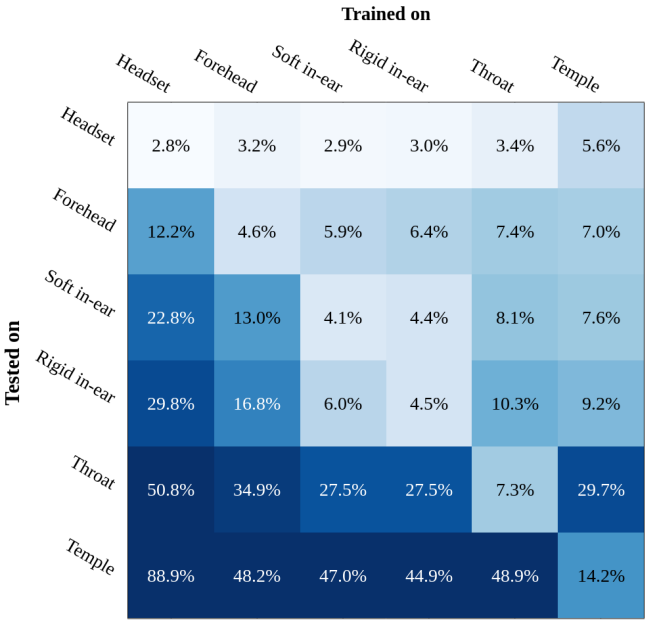


Fig. 5. PER of the finetuned phonemizer. Model trained and tested on the "speech-clean" subset

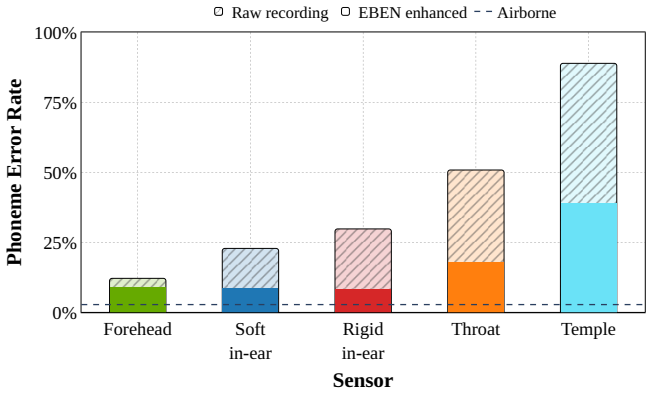


Fig. 6. PER of the raw and EBEN-enhanced test recordings phonemized by the headset phonemizer

This comparison was conducted employing solely the headset phonemizer. The results, presented as a histogram in Figure 6, demonstrate that all sensors exhibited a strong improvement in PER thanks to the enhancement obtained with EBEN models. This is a promising outcome, particularly in light of the fact that contemporary speech enhancement systems, when used as a pre-processing layer in ASR systems continue to face challenges in enhancing performance due to the production of artifacts [83]. The aforementioned findings thus corroborate the hypothesis that speech intelligibility is enhanced by EBEN models. However, it is essential to note that the direct parallel between the headset phonemizer and the human ear cannot be taken for granted. Subjective testing could be considered to provide an additional layer of validation.

Finally, in order to gain a detailed understanding of the acoustic properties of each sensor, we identified the top five phonetic confusions of all phonemizers when tested on their raw signals training distribution. This provided examination, presented in Table IV, offers valuable insights into the challenges inherent to each sensor placement and technology for the identification of phonemes due to the absence of high-frequency content. This table was obtained by counting the most frequently used editing operation to go from the model's decoded output to the labeled phonetic transcription when computing the Levenshtein distance [84]. It should be noted that only in-word phonetic confusions were considered in order to avoid including the suprasegmentals phenomenon [‿] (*i.e.* "liaison" in French), which poses difficulties for many native speakers during reading and which is not present in our labeled phonetic transcription. This phenomenon caused all models to either omit or add [z] (*i.e.* [∅] → [z] and [z] → [∅]) as their top errors.

Once this issue has been disregarded, Table IV reveals that the model trained on airborne microphone signals encounters difficulties in distinguishing the similar phoneme confusion we encounter in French, such as [e]/[ɛ] and [ɑ/][o]/[ɔ], which all are voiced sounds. Similar errors have been observed in other audio sensors with a sufficiently large bandwidth, albeit with varying extents. Of particular interest is the observation that the temple vibration pickup, which has a narrow bandwidth, most frequently fails to capture the voiceless alveolar and plosive fricatives [s]/[t] that rely on substantial high-frequency support. The wide band support [ʁ] is also left out to a significant extent for the temple vibration pickup and throat microphone. For the temple sensor, this could be due to the similarity of this phoneme with the sensor's self-noise. For the throat microphone, we do not have a clear explanation, especially since this sensor placement seems to be optimal for capturing this uvular fricative.

C. Speaker verification

This final section addresses speaker verification, which ascertains whether two audio files feature the same speaker,

providing a boolean answer. Speaker verification constitutes a subset of speaker recognition tasks [85], which also encompass speaker identification and diarization. This task is especially pertinent in radio communication scenarios with multiple speakers, where BCMs can exacerbate the probability of voice confusion. Such confusion can impede fluid conversations, particularly in noisy environments. Furthermore, robust speaker identification systems are essential in contexts where heavy machinery is operated via voice commands, ensuring rigorous safety protocols are upheld. As stated by [86], *"The presence and use of high-frequency information from the speech signal improves the recognition of individual speakers, particularly in noisy environments."*. Therefore, as highlighted by previous research [87]–[91], the lack of these frequencies presents a significant challenge to BCMs speaker verification, necessitating further investigation to produce a more robust technology for these sensors.

1) Model selection: In this paper, we assess the efficiency of speaker verification tasks using body-conducted sensors from the Vibravox dataset. The task is based on a distance metric comparison of speaker embedding results generated by a neural network from two distinct speech signals. Recently, Brydinskyi et al. [92] investigated the effectiveness of various models trained on English datasets for differentiating non-English speakers. Their findings indicated superior performance from TitaNet [31] and ECAPA-TDNN [93] compared to WavLM [32] or Pyannote [33]. In light of the recent advances in speaker verification, particularly the ECAPA2 model [34], which achieves state-of-the-art results, we elected to utilize its pre-trained version[4] on the Vibravox dataset for our analysis.

2) Testing procedure: It is crucial to highlight that the speaker verification task is only composed of a testing procedure, where we utilize a pretrained ECAPA2 model, without any further training. Consequently, we solely utilize the test split of the *clean-speech* subset. The speaker verification testing set is constructed by forming pairs of speech signals, designated as A and B, in alignment with the methodology outlined in [92]. Notably, A and B may correspond to different airborne or body-conduction microphones. For each speaker, we generate N_p pairs with their own recordings and an additional randomly selected N_p pairs with recordings from different speakers. To increase the difficulty of the task, we maintain the possibility of separating speakers by gender. Here, N_p represents the number of possible pairs for the speaker with the fewest recordings, which is set at 2346. We thus have a total of 98532 test pairs given the fact that the *clean-speech* test set is composed of 21 speakers.

When comparing pairs of signals, we compute the cosine similarity and the Euclidean distance between their normalized embeddings. After inference on the entire set of pairs, we compute the Equal Error Rate (EER) and the normalized minimum of the Detection Cost Function (DCF) [94] from the cosine similarity scores.

[4]available at https://huggingface.co/Jenthe/ECAPA2

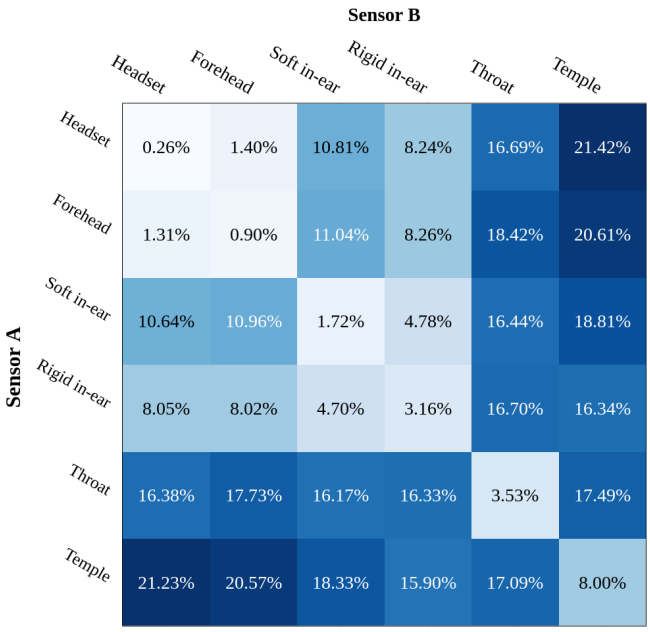


Fig. 7. EER of the speaker verification model for all sensor pairs

3) Results: The results of the ECAPA2 model test are presented in Figure 7 for all sensor combinations. As with Speech-to-Phonemes, the matrix emphasizes a diagonal with lower coefficients because it is easier to distinguish between two recordings taken from the same audio sensor. The asymmetry of the matrix is to be expected, given that the script for generating pairs is not commutative due to its randomness. Nevertheless, the results from the upper triangular and lower triangular parts are very close and show a low variance. It is immediately apparent that there is a considerable discrepancy in the diagonal EER between different sensors. To illustrate this point, the headset and temple vibration pickup exhibit an EER differing by a rate of 30. The cross-tested EER for the forehead/headset and soft/rigid in-ear microphones remains low compared to their initial performance, indicating the proximity of those sensors. Once more, we observe that the difficulty of the task is again mostly driven by the sensor's bandwidth.

To complete this study, we also compared the EER on raw signals and EBEN-enhanced ones. In contrast with the observations made in Section IV-B, a deterioration is observed in Figure 8. Note that the speaker identity preservation was not included among the three losses, namely, the adversarial loss $\mathcal{L}_{\mathcal{G}}^{adv}$, the feature loss $\mathcal{L}_{\mathcal{G}}^{feat}$ and the spectral loss $\mathcal{L}_{\mathcal{G}}^{spec}$, which were employed to adjust the EBEN parameters. There is a further possibility that the ECAPA2 model may be interpreting the EBEN outputs as being out of distribution, which could explain this poor performance. However, this default could then be compensated for in future work by emphasizing the importance of preserving speaker characteristics. This could be achieved by incorporating losses in that aim into the overall objective, as explored in [95].

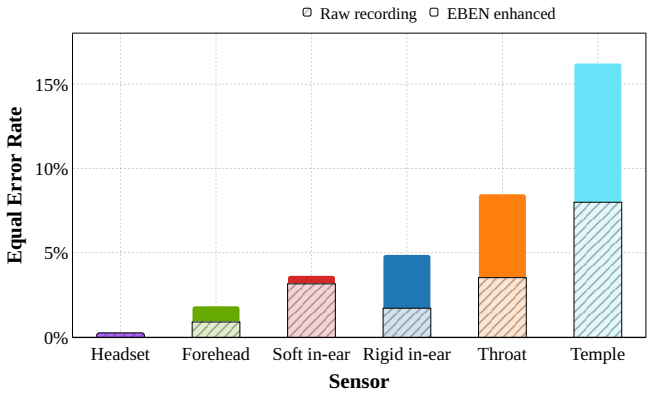


Fig. 8. EER of the raw and EBEN enhanced recordings with the speaker verification model on pairs of the same audio sensor

V. Discussion and Future Work

Building the Vibravox dataset has been a protracted undertaking that has necessitated the achievement of several milestones. These include the construction of equipment, the development of necessary software, the recording of participants, the filtering of data, and other related tasks. We are pleased to announce the release of this highly curated dataset, comprising four subsets: *speech-clean*, *speech-noisy*, *speechless-clean* and *speechless-noisy*. The total duration of the recordings is over 38 hours, with 188 participants. This corpus is being made available online on HuggingFace via Creative Commons Attribution 4.0 International license. Using Vibravox has proven to yield satisfactory results in speech enhancement, speech-to-phonemes and speaker verification, which are three fundamental speech processing tasks. The EBEN models used for speech enhancement demonstrated the capacity to enhance objective metrics such as Noresqua-MOS and STOI, while simultaneously improving phoneme transcription on every sensor. However, it was found to have a detrimental impact on speaker verification performance. Also, the aforementioned three tasks facilitated an in-depth comprehension of the distinctive acoustic subtleties to the various audio sensors and their respective placements. It is our hope that this release will drive progress in body-conducted speech analysis and facilitate the development of robust communication systems for real-world applications.

Acknowledgements

The authors would like to express their gratitude to all the 200 participants in the Vibravox measurement protocol, whose invaluable contributions made this research possible. We would also like to thank Jean-Baptiste Doc for his assistance in developing 3D-printed components to facilitate the positioning of sensors to the participants' diverse morphologies, and Philippe Chenevez for his expertise in electrical engineering for the pre-amplifier circuits. This work was granted access to the HPC/AI resources of [CINES / IDRIS / TGCC] under the grant 2022-AD011013469 awarded by GENCI and partially funded by the French National Research Agency under the ANR Grant No. ANR-20-THIA-0002.

VI. Ethical Considerations

In accordance with the guidelines of the General Data Protection Regulation (GDPR), which governs the processing and protection of personal data of citizens and residents of the European Union (EU), all necessary measures were taken to protect the privacy and rights of participants in this study. This included obtaining informed consent, clearly communicating the purpose of data collection, and ensuring anonymization and encryption of data during storage and analysis. In addition, the use of Wikipedia text ensures compliance with copyright and licensing issues. Finally, to increase the representativeness and inclusivity of the dataset, a deliberate effort was made to recruit a diverse and gender-balanced group of participants.

References

[1] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt, "Combining standard and throat microphones for robust speech recognition," *IEEE Signal Processing Letters*, vol. 10, no. 3, pp. 72–74, 2003.

[2] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, and X. Huang, "Air-and bone-conductive integrated microphones for robust speech detection and enhancement," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*. IEEE, 2003, pp. 249–254.

[3] J. C. Bos, D. W. Tack, and L. L. Bossi, "Speech input hardware investigation for future dismounted soldier computer systems," *DRCD Toronto CR*, vol. 64, p. 2005, 2005.

[4] B. Acker-Mills, A. Houtsma, and W. Ahroon, "Speech intelligibility with acoustic and contact microphones," Army Aeromedical Research Lab Fort Rucker Al, Tech. Rep., 2005.

[5] A. Shahina and B. Yegnanarayana, "Mapping speech spectra from throat microphone to close-speaking microphone: A neural network approach," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–10, 2007.

[6] S. Lee, J. Kim, I. Yun, G. Y. Bae, D. Kim, S. Park, I.-M. Yi, W. Moon, Y. Chung, and K. Cho, "An ultrathin conformable vibration-responsive electronic skin for quantitative vocal recognition," *Nature communications*, vol. 10, no. 1, p. 2468, 2019.

[7] S. Lee, H. Roh, J. Kim, S. Chung, D. Seo, W. Moon, and K. Cho, "An electret-powered skin-attachable auditory sensor that functions in harsh acoustic environments," *Advanced Materials*, vol. 34, no. 40, p. 2205537, 2022.

[8] Z. Che, X. Wan, J. Xu, C. Duan, T. Zheng, and J. Chen, "Speaking without vocal folds using a machine-learning-assisted wearable sensing-actuation system," *Nature Communications*, vol. 15, no. 1, pp. 1–11, 2024.

[9] S. Björklund and J. Sundberg, "Relationship between subglottal pressure and sound pressure level in untrained voices," *Journal of Voice*, vol. 30, no. 1, pp. 15–20, 2016.

[10] K. Zhang, Y. Ren, C. Xu, and Z. Zhao, "WSRGlow: A glow-based waveform generative model for audio super-resolution," *Proc. Interspeech*, vol. 1649-1653, 2021.

[11] S. Han and J. Lee, "NU-Wave 2: A general neural audio upsampling model for various sampling rates," *arXiv preprint arXiv:2206.08545*, 2022.

[12] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *arXiv preprint arXiv:2206.03065*, 2022.

[13] M. Mandel, O. Tal, and Y. Adi, "AERO: Audio super resolution in the spectral domain," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] S. Nisticò, L. Palopoli, and A. P. Romano, "Audio super-resolution via vision transformer," *Journal of Intelligent Information Systems*, pp. 1–15, 2023.

[15] C. Shuai, C. Shi, L. Gan, and H. Liu, "mdctGAN: Taming transformer-based gan for speech super-resolution with modified dct spectra," *arXiv preprint arXiv:2305.11104*, 2023.

[16] 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 *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[17] ——, "Configurable EBEN: Extreme bandwidth extension network to enhance body-conducted speech capture," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[18] C. Karthikeyan, T. R. Kumar, D. V. Babu, M. Baskar, R. Jayaraman, and M. Shahid, "Speech enhancement approach for body-conducted unvoiced speech based on taylor–boltzmann machines trained dnn," *Computer Speech & Language*, vol. 83, p. 101549, 2023.

[19] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: A unified framework for bandwidth extension and speech enhancement," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[20] C. Li, F. Yang, and J. Yang, "Restoration of bone-conducted speech with u-net-like model and energy distance loss," *IEEE Signal Processing Letters*, 2023.

[21] L. He, H. Hou, S. Shi, X. Shuai, and Z. Yan, "Towards bone-conducted vibration speech enhancement on head-mounted wearables," in *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, 2023, pp. 14–27.

[22] C. Li, F. Yang, and J. Yang, "A two-stage approach to quality restoration of bone-conducted speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[23] A. Edraki, W.-Y. Chan, J. Jensen, and D. Fogerty, "Speaker adaptation for enhancement of bone-conducted speech," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 456–10 460.

[24] M. Ohlenbusch, C. Rollwage, and S. Doclo, "Multi-microphone noise data augmentation for DNN-based own voice reconstruction for hearables in noisy environments," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 416–420.

[25] Y. Sui, M. Zhao, J. Xia, X. Jiang, and S. Xia, "TRAMBA: A hybrid transformer and mamba architecture for practical audio and bone conduction speech super resolution and enhancement on mobile and wearable platforms," *arXiv preprint arXiv:2405.01242*, 2024.

[26] G. Yu, X. Zheng, N. Li, R. Han, C. Zheng, C. Zhang, C. Zhou, Q. Huang, and B. Yu, "Bae-net: A low complexity and high fidelity bandwidth-adaptive neural network for speech super-resolution," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 571–575.

[27] "Universal score-based speech enhancement with high content preservation," in *Proc. Interspeech 2024*, 2024.

[28] H. Liu, K. Chen, Q. Tian, W. Wang, and M. D. Plumbley, "Audiosr: Versatile audio super-resolution at scale," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1076–1080.

[29] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[30] NVIDIA, "Canary-1B," https://huggingface.co/nvidia/canary-1b, 2024, accessed: 04/24/2024.

[31] N. R. Koluguri, T. Park, and B. Ginsburg, "Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8102–8106.

[32] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[33] H. Bredin, "Pyannote audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *24th INTERSPEECH Conference (INTERSPEECH 2023)*. ISCA, 2023, pp. 1983–1987.

[34] J. Thienpondt and K. Demuynck, "Ecapa2: A hybrid neural network architecture and training strategy for robust speaker embeddings," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[35] M. Wang, J. Chen, X.-L. Zhang, and S. Rahardja, "End-to-end multimodal speech recognition on an air and bone conducted speech corpus," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–12, 2022.

[36] ESMB-corpus. (2021, https://github.com/elevoctech/ESMB-corpus) Elevoc simultaneously-recorded microphone/bone-sensor. Accessed on 2023-10-28. [Online]. Available: https://github.com/elevoctech/ESMB-corpus

[37] 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*, p. e24110, 2024.

[38] J. Hauret, M. Olivier, C. Langrenne, S. Poirée, and É. Bavu, "vibravox (revision 7990b7d)," 2024. [Online]. Available: https://huggingface.co/datasets/Cnam-LMSSC/vibravox

[39] F. Denk and B. Kollmeier, "The hearpiece database of individual transfer functions of an in-the-ear earpiece for hearing device research," 2021.

[40] M. Ohlenbusch, C. Rollwage, and S. Doclo, "Modeling of speech-dependent own voice transfer characteristics for hearables with in-ear microphones," *arXiv preprint arXiv:2310.06554*, 2023.

[41] ——, "Speech-dependent data augmentation for own voice reconstruction with hearable microphones in noisy environments," *arXiv preprint arXiv:2405.11592*, 2024.

[42] A. Lozhkov, L. Ben Allal, L. von Werra, and T. Wolf, "Fineweb-edu," May 2024. [Online]. Available: https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu

[43] E. Erzin, "Improving throat microphone speech recognition by joint analysis of throat and acoustic microphone recordings," *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 7, pp. 1316–1324, 2009.

[44] M. T. Turan and E. Erzin, "Source and filter estimation for throat-microphone speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 265–275, 2015.

[45] D. Shan, X. Zhang, C. Zhang, and L. Li, "A novel encoder-decoder model via NS-LSTM used for bone-conducted speech enhancement," *IEEE Access*, vol. 6, pp. 62 638–62 644, 2018.

[46] S. K. Davis, P. T. Calamia, W. J. Murphy, and C. J. Smalt, "In-ear and on-body measurements of impulse-noise exposure," *International journal of audiology*, vol. 58, no. sup1, pp. S49–S57, 2019.

[47] M. Tagliasacchi, Y. Li, K. Misiunas, and D. Roblek, "SEANet: A multi-modal speech enhancement network," *Proc. Interspeech 2020*, pp. 1126–1130, 2020.

[48] C. Zheng, L. Xu, X. Fan, J. Yang, J. Fan, and X. Huang, "Dual-path transformer-based network with equalization-generation components prediction for flexible vibrational sensor speech enhancement in the time domain," *The Journal of the Acoustical Society of America*, vol. 151, no. 5, pp. 2814–2825, 2022.

[49] S. Liebich, J.-G. Richter, J. Fabry, C. Durand, J. Fels, and P. Jax, "Direction-of-arrival dependency of active noise cancellation headphones," in *Noise Control and Acoustics Division Conference*, vol. 51425. American Society of Mechanical Engineers, 2018, p. V001T08A003.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[51] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[52] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[53] V. Kuleshov, S. Z. Enam, and S. Ermon, "Audio super-resolution using neural nets," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2017.

[54] S. Ishimitsu, "Body-conducted speech recognition and its application to speech support system," *SCIYO. COM*, p. 51, 2010.

[55] S. Everett and N. R. L. W. DC, "Automatic speaker recognition for military applications: Applications survey and operational requirements." 1985.

[56] C. Blondé-Weinmann, T. Joubaud, V. Zimpfer, P. Hamery, and S. Roth, "Numerical and experimental investigation of the sound transmission delay from a skin vibration to the occluded ear canal," *Journal of Sound and Vibration*, vol. 542, p. 117345, 2023.

[57] F. Denk, M. Lettau, H. Schepker, S. Doclo, R. Roden, M. Blau, J. Wellmann, B. Kollmeier *et al.*, "A one-size-fits-all earpiece with multiple microphones and drivers for hearing device research," in *Audio Engineering Society Conference: 2019 AES International Conference on Headphone Technology*. Audio Engineering Society, 2019.

[58] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020.

[59] M. Bernard and H. Titeux, "Phonemizer: Text to phones transcription for multiple languages in python," *Journal of Open Source Software*, vol. 6, no. 68, p. 3958, 2021. [Online]. Available: https://doi.org/10.21105/joss.03958

[60] P. Lecomte and P.-A. Gauthier, "Real-time 3d ambisonics using faust, processing, pure data, and osc," in *18th International Conference on Digital Audio Effects, DAFx-15*, 2015.

[61] P. Lecomte, P.-A. Gauthier, C. Langrenne, A. Berry, and A. Garcia, "A fifty-node Lebedev grid and its applications to ambisonics," *Journal of the Audio Engineering Society*, vol. 64, no. 11, pp. 868–881, 2016.

[62] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[63] T. Giorgino, "Computing and visualizing dynamic time warping alignments in r: The dtw package," *Journal of Statistical Software*, vol. 31, no. 7, 2009.

[64] M. J. Crocker, "Rating measures, descriptors, criteria, and procedures for determining human response to noise," *Handbook of Noise and Vibration Control*, pp. 394–413, 2007.

[65] 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, no. 3, pp. 495–502, 2011.

[66] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[67] T. Q. Nguyen, "Near-perfect-reconstruction pseudo-QMF banks," *IEEE Transactions on signal processing*, vol. 42, no. 1, pp. 65–76, 1994.

[68] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[69] C. J. Steinmetz and J. D. Reiss, "auraloss: Audio focused loss functions in PyTorch," in *Digital Music Research Network One-day Workshop (DMRN+15)*, 2020.

[70] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *Proc. Interspeech*, 2020.

[71] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25.

[72] J. Hwang, M. Hira, C. Chen, X. Zhang, Z. Ni, G. Sun, P. Ma, R. Huang, V. Pratap, Y. Zhang, A. Kumar, C.-Y. Yu, C. Zhu, C. Liu, J. Kahn, M. Ravanelli, P. Sun, S. Watanabe, Y. Shi, Y. Tao, R. Scheibler, S. Cornell, S. Kim, and S. Petridis, "Torchaudio 2.1: Advancing speech recognition, self-supervised learning, and audio processing components for pytorch," 2023.

[73] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[74] W. Falcon and P. L. team, "Pytorch lightning." [Online]. Available: https://www.pytorchlightning.ai

[75] O. Yadan, "Hydra - a framework for elegantly configuring complex applications," Github, 2019. [Online]. Available: https://github.com/facebookresearch/hydra

[76] Q. Lhoest, A. Villanova del Moral, Y. Jernite, A. Thakur, P. von Platen, S. Patil, J. Chaumond, M. Drame, J. Plu, L. Tunstall, J. Davison, M. Šaško, G. Chhablani, B. Malik, S. Brandeis, T. Le Scao, V. Sanh, C. Xu, N. Patry, A. McMillan-Major, P. Schmid, S. Gugger, C. Delangue, T. Matussière, L. Debut, S. Bekman, P. Cistac, T. Goehringer, V. Mustar, F. Lagunas, A. Rush, and T. Wolf, "Datasets: A community library for natural language processing," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 175–184. [Online]. Available: https://aclanthology.org/2021.emnlp-demo.21

[77] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[78] P. Manocha and A. Kumar, "Speech quality assessment through mos using non-matching references," in *Interspeech*, 2022. [Online]. Available: https://arxiv.org/abs/2206.12285

[79] X. Tan, T. Qin, J. Bian, T.-Y. Liu, and Y. Bengio, "Regeneration learning: A learning paradigm for data generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 20, 2024, pp. 22 614–22 622.

[80] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[81] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 993–1003. [Online]. Available: https://aclanthology.org/2021.acl-long.80

[82] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[83] K. Iwamoto, T. Ochiai, M. Delcroix, R. Ikeshita, H. Sato, S. Araki, and S. Katagiri, "How bad are artifacts?: Analyzing the impact of speech enhancement errors on asr," *arXiv preprint arXiv:2201.06685*, 2022.

[84] V. I. Levenshtein *et al.*, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8. Soviet Union, 1966, pp. 707–710.

[85] D. A. Reynolds and W. M. Campbell, "Text-independent speaker recognition," in *Springer Handbook of Speech Processing*, 2008, pp. 763–782.

[86] J. C. Schwartz, A. T. Whyte, M. Al-Nuaimi, and J. J. Donai, "Effects of signal bandwidth and noise on individual speaker identification," *The Journal of the Acoustical Society of America*, vol. 144, no. 5, pp. EL447–EL452, 2018.

[87] M. Sahidullah, R. Gonzalez Hautamäki, T. D. A. Lehmann, T. Kinnunen, Z.-H. Tan, V. Hautamäki, R. Parts, and M. Pitkänen, "Robust speaker recognition with combined use of acoustic and throat microphone speech," 2016.

[88] M. Sahidullah, D. A. L. Thomsen, R. G. Hautamäki, T. Kinnunen, Z.-H. Tan, R. Parts, and M. Pitkänen, "Robust voice liveness detection and speaker verification using throat microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 44–56, 2017.

[89] R. Liu, C. Cornelius, R. Rawassizadeh, R. Peterson, and D. Kotz, "Vocal resonance: Using internal body voice for wearable authentication," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–23, 2018.

[90] J. Shang and J. Wu, "Enabling secure voice input on augmented reality headsets using internal body voice," in *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 2019, pp. 1–9.

[91] Y. Gao, Y. Jin, J. Chauhan, S. Choi, J. Li, and Z. Jin, "Voice in ear: Spoofing-resistant and passphrase-independent body sound authentication," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–25, 2021.

[92] V. Brydinskyi, Y. Khoma, D. Sabodashko, M. Podpora, V. Khoma, A. Konovalov, and M. Kostiak, "Comparison of modern deep learning models for speaker verification," vol. 14, no. 1329, 2024.

[93] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[94] US National Institute of Standards and Technology, "NIST 2018 Speaker Recognition Evaluation Plan," 2018. [Online]. Available: https://www.nist.gov/system/files/documents/2018/08/17/sre18_eval_plan_2018-05-31_v6.pdf

[95] D. Wang, S. Liu, X. Wu, H. Lu, L. Sun, X. Liu, and H. Meng, "Speaker identity preservation in dysarthric speech reconstruction by adversarial speaker adaptation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6677–6681.

Julien Hauret is a PhD candidate at Cnam Paris, pursuing research in machine learning applied to speech processing. He holds two MSc degrees from ENS Paris Saclay, one in Electrical Engineering (2020) and the other in Applied Mathematics (2021). His research training is evidenced by his experiences at Columbia University, the French Ministry of the Armed Forces and the Pulse Audition start-up. Additionally, he has lectured for two consecutive years on algorithms and data structures at the École des Ponts ParisTech. His research focuses on the use of deep learning for speech enhancement applied to body-conducted speech. With a passion for interdisciplinary collaboration, Julien aims to improve human communication through technology.

Malo Olivier is a engineer-student at INSA of Lyon, that pursued an internship at the Laboratoire de Mécanique des Structures et des Systèmes Couplés (LMSSC) in the Conservatoire National des Arts et Métiers (CNAM), Paris, France. He is following his graduate studies at the Computer Sciences department of INSA Lyon for which he will receive his degree in 2024. Malo has valuable skills implementing different solutions from Information Systems problematics to deep neural networks architectures, including web applications. He foresees to pursue a Ph.D. in Artificial Intelligence, specializing in deep neural networks applied to science domains and hopes his engineer profile highlights his implementing abilities within projects of high interest.

Thomas Joubaud is a Research Associate at the Acoustics and Soldier Protection department within the French-German Research Institute of Saint-Louis (ISL), France, since 2019. In 2013, he received the graduate degree from Ecole Centrale Marseille, France, as well as the master's degree in Mechanics, Physics and Engineering, specialized in Acoustical Research, of the Aix-Marseille University, France. He earned the Ph.D. degree in Mechanics, specialized in Acoustics, of the Conservatoire National des Arts et Métiers (Cnam), Paris, France, in 2017. The thesis was carried out in collaboration with and within the ISL. From 2017 to 2019, he worked as a post-doctorate research engineer with Orange SA company in Cesson-Sévigné, France. His research interests include audio signal processing, hearing protection, psychoacoustics, especially speech intelligibility and sound localization, and high-level continuous and impulse noise measurement.

Christophe Langrenne is a Research Engineer at the Laboratoire de Mécanique des Structures et des Systèmes Couplés (LMSSC) at the Conservatoire National des Arts et Métiers (Cnam), Paris, France. After completing his PhD on the regularization of inverse problems, he developed a fast multipole method (FMM) algorithm for solving large-scale scattering and propagation problems. Also interested in 3D audio, he co-supervised 3 PhD students on this theme, in particular on Ambisonic (recording and decoding) and binaural restitution (front/back confusions).

Sarah Poiree is a technician at the Laboratoire de Mécanique et des Systèmes Couplés (LMSSC) within the Conservatoire National des Arts et Métiers (Cnam), Paris, France. Her activities focus on the design and development of experimental setups. Notably, she contributed to the creation of the 3D sound spatialization system used during the recording of the Vibravox dataset.

Véronique Zimpfer is a Scientific Researcher at the Acoustics and Soldier Protection department within the French-German Research Institute of Saint-Louis (ISL), Saint-Louis, France, since 1997. She holds a M.Sc in Signal Processing from the Grenoble INP, France and obtained a PhD in Acoustics from INSA Lyon, France, in 2000. Her expertise lies at the intersection of communication in noisy environments and auditory protection. Her research focuses on improving adaptive auditory protectors, refining radio communication strategies through unconventional microphone methods, and enhancing auditory perception while utilizing protective gear.

Éric Bavu is a Full Professor of Acoustics and Signal Processing at the Laboratoire de Mécanique des Structures et des Systèmes Couplés (LMSSC) within the Conservatoire National des Arts et Métiers (Cnam), Paris, France. He completed his undergraduate studies at École Normale Supérieure de Cachan, France, from 2001 to 2005. In 2005, he earned an M.Sc in Acoustics, Signal Processing, and Computer Science Applied to Music from Université Pierre et Marie Curie Sorbonne University (UPMC), followed by a Ph.D. in Acoustics jointly awarded by Université de Sherbrooke, Canada, and UPMC, France, in 2008. He also conducted post-doctoral research on biological soft tissues imaging at the Langevin Institute at École Supérieure de Physique et Chimie ParisTech (ESPCI), France. Since 2009, he has supervised 8 Ph.D. students at LMSSC, focusing on time domain audio signal processing for inverse problems, 3D audio, and deep learning for audio. His current research interests encompass deep learning methods applied to inverse problems in acoustics, moving sound source localization and tracking, speech enhancement, and speech recognition.

APPENDIX A
DATASET STATISTICS

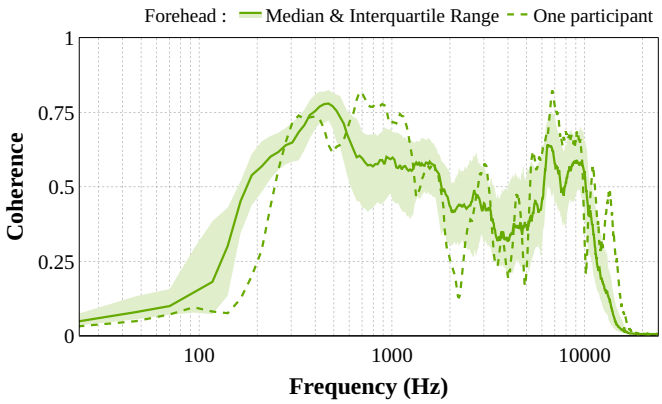
TABLE V
SPEAKERS AGE BALANCE

Gender	Mean age (years)	Median age (years)	Min age (years)	Max age (years)
Female	25.9	22	19	59
Male	31.4	27	18	82
All	**28.55**	**25**	**18**	**82**

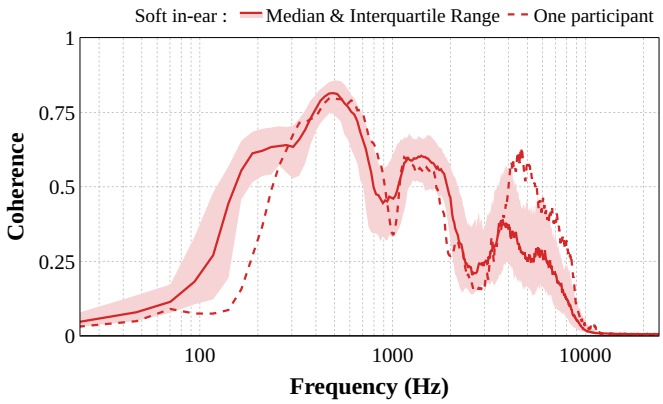
TABLE VI
AUDIO DATA SUMMARY

Subset	Split	Audio duration (hours)	Number of audio clips	Download size	Number of Speakers (Female/Male)	F/M Gender repartition (audio duration)	Mean audio duration (s)	Median audio duration (s)	Max audio duration (s)	Min audio duration (s)
speech_clean	train	6x20.94	6x20,981	108.32GB	77F/72M	52.46%/47.54%	3.59	3.50	12.20	0.52
	validation	6x2.42	6x2,523	12.79GB	9F/9M	52.13%/47.87%	3.46	3.38	9.44	0.66
	test	6x3.03	6x3,064	15.84GB	11F/10M	55.74%/44.26%	3.56	3.48	9.58	0.58
speech_noisy	train	6x1.26	6x1,220	6.52GB	77F/72M	54.31%/45.69%	3.71	3.64	8.66	0.46
	validation	6x0.13	6x132	0.71GB	9F/9M	56.61%/43.39%	3.67	3.47	7.36	1.10
	test	6x0.18	6x175	0.94GB	11F/10M	55.54%/44.46%	3.66	3.70	6.88	1.00
speechless_clean	train	6x2.24	6x149	8.44GB	77F/72M	51.68%/48.32%	54.10	54.10	54.10	53.99
	validation	6x0.27	6x18	1.02GB	9F/9M	50.00%/50.00%	54.10	54.10	54.10	54.05
	test	6x0.32	6x21	1.19GB	11F/10M	52.38%/47.62%	54.10	54.10	54.10	54.10
speechless_noisy	train	6x5.96	6x149	24.48GB	77F/72M	51.68%/48.32%	144.03	144.03	144.17	143.84
	validation	6x0.72	6x18	2.96GB	9F/9M	50.00%/50.00%	144.03	144.03	144.05	143.94
	test	6x0.84	6x21	3.45GB	11F/10M	52.38%/47.62%	144.04	144.03	144.05	144.03
Total		**6x38.31**	**6x28,471**	**186.64GB**	**97F/91M**	**52.55%/47.45%**				

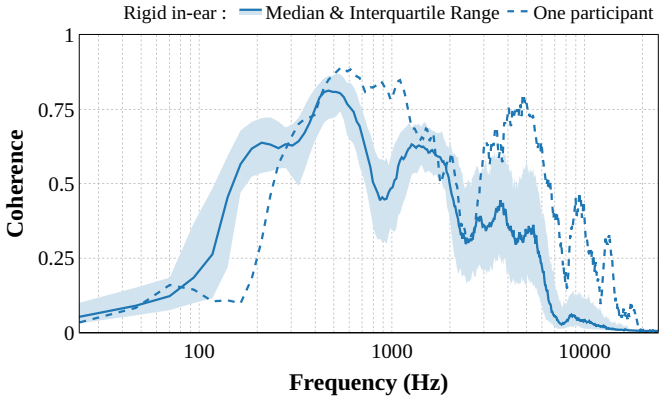
APPENDIX B
COHERENCE FUNCTION STATISTICS



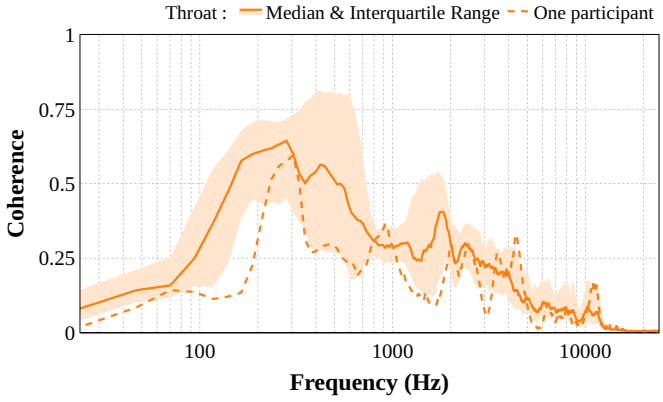
(a) Forehead



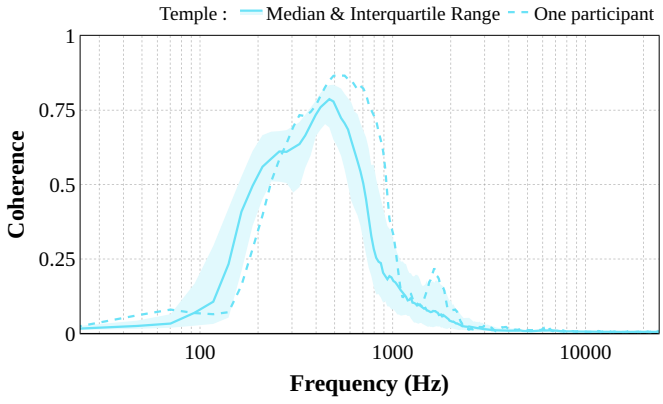
(b) Rigid in-ear



(c) Soft in-ear



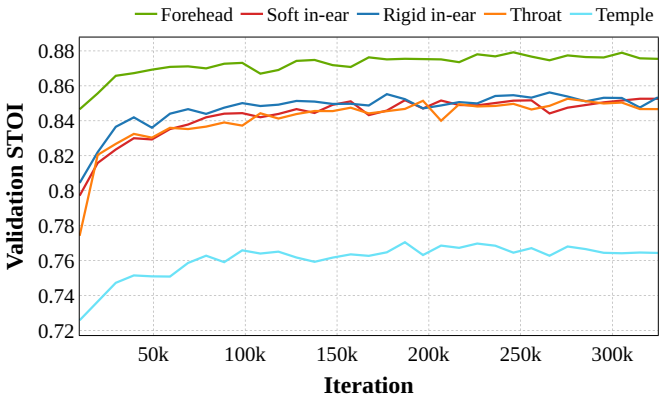
(d) Throat



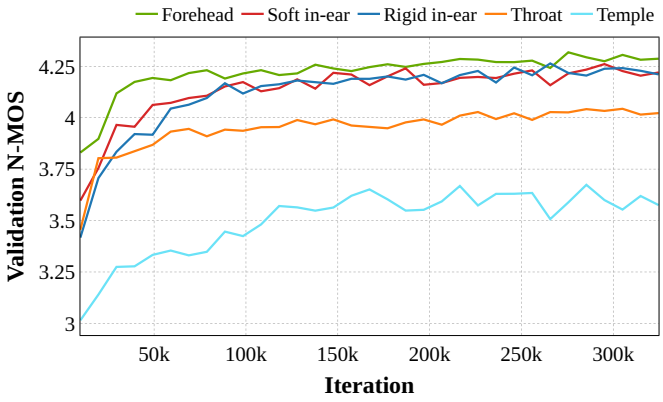
(e) Temple

Fig. 9. Median coherence function per sensor given with their 25-75% IQR and a single participant

APPENDIX C
SPEECH ENHANCEMENT TRAINING CURVES



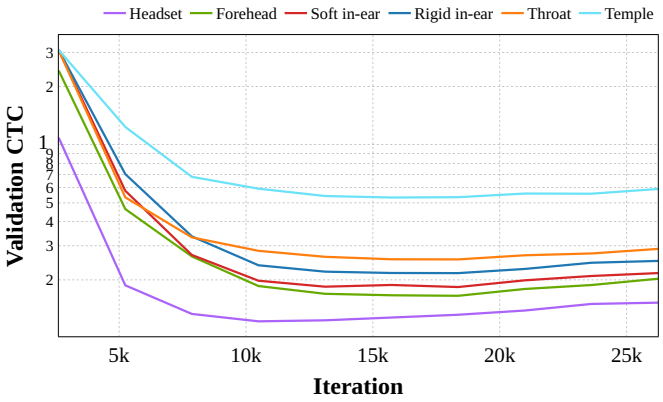
(a) Validation STOI



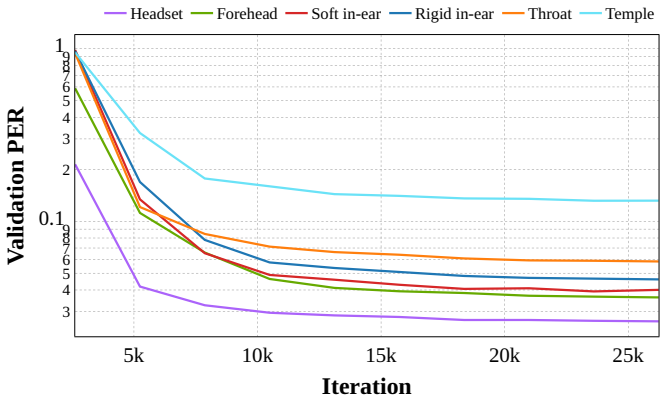
(b) Validation Noresqua-MOS

Fig. 10. EBEN validation curves for Speech Enhancement

APPENDIX D
SPEECH-TO-PHONEME TRAINING CURVES



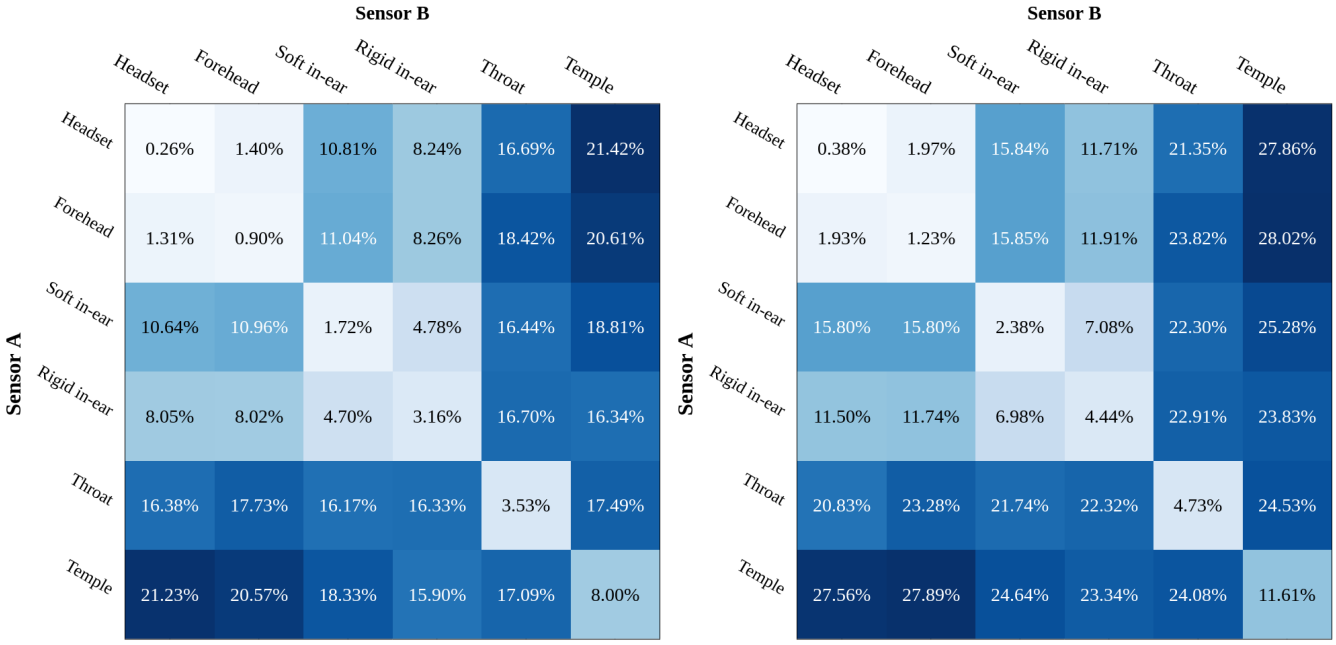
(a) Validation Connectionist Temporal Classification Loss



(b) Validation Phoneme-Error-Rate

Fig. 11. Wav2vec2 validation curves for Speech-to-Phonemes

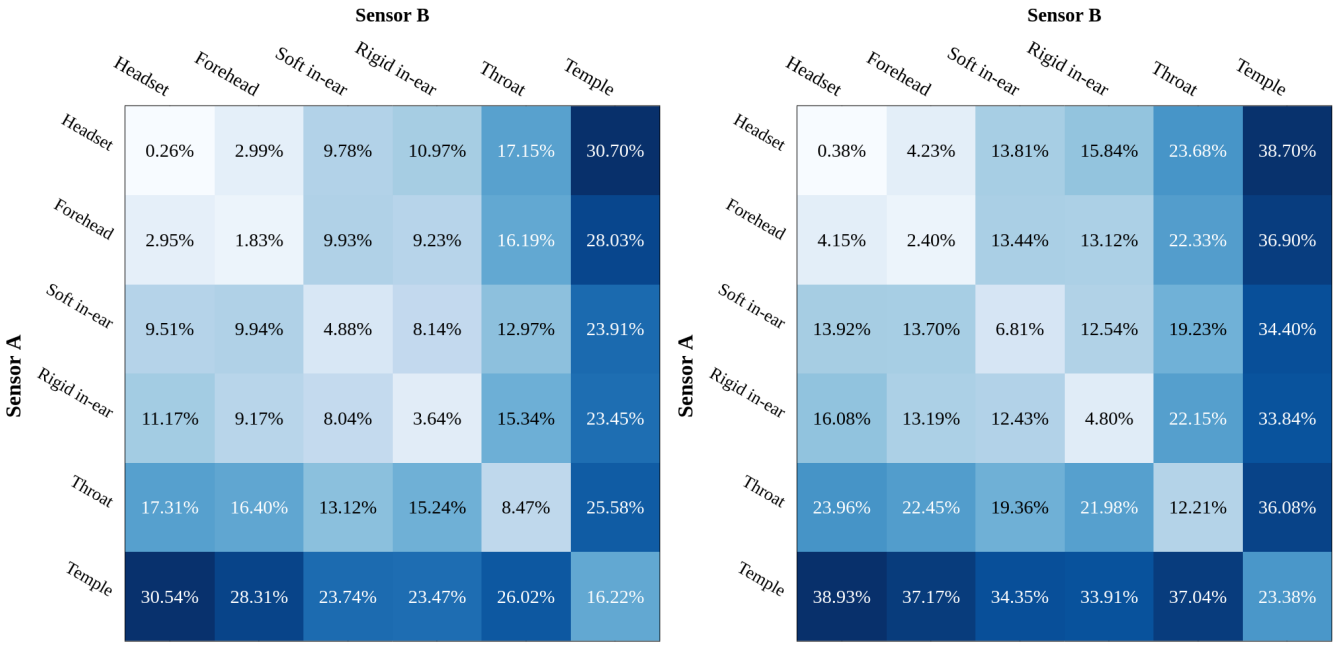
APPENDIX E
SPEAKER VERIFICATION EER



(a) Mixed gender pairs (already given on Figure 7)

(b) Same gender pairs

Fig. 12. EER obtained with the speaker verification model for sensor pairs

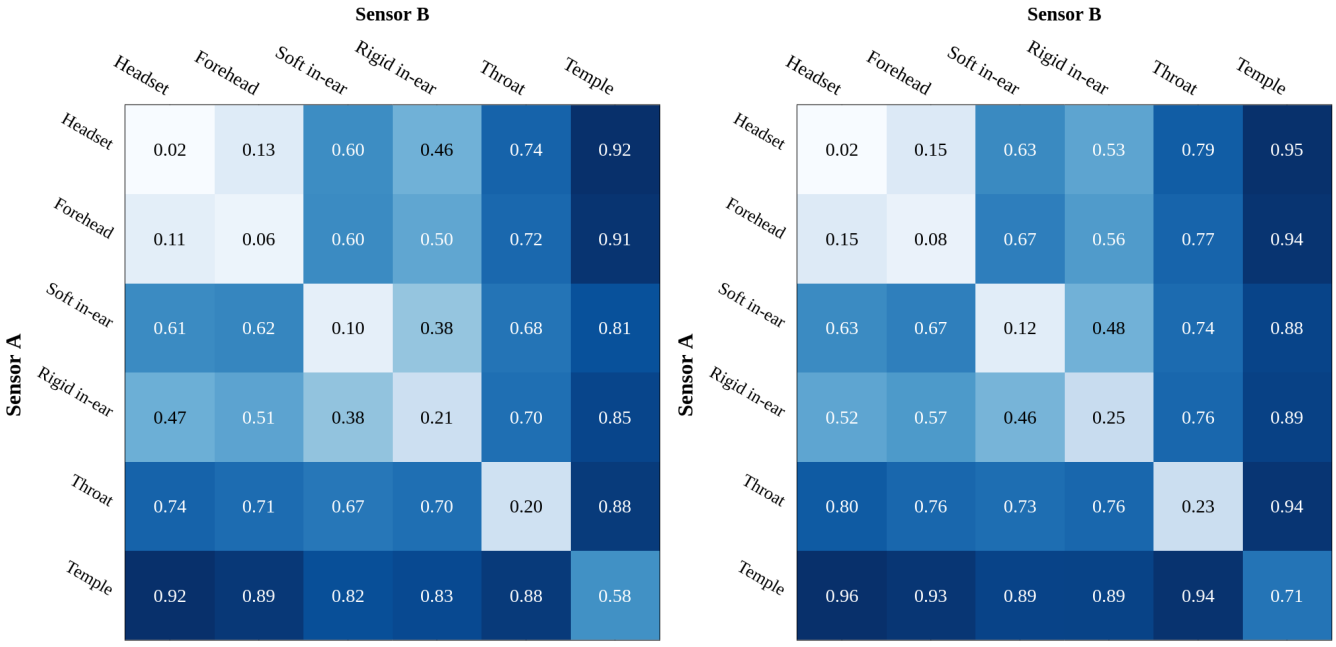


(a) Mixed gender pairs

(b) Same gender pairs

Fig. 13. EER obtained with the speaker verification model for sensor pairs enhanced by EBEN

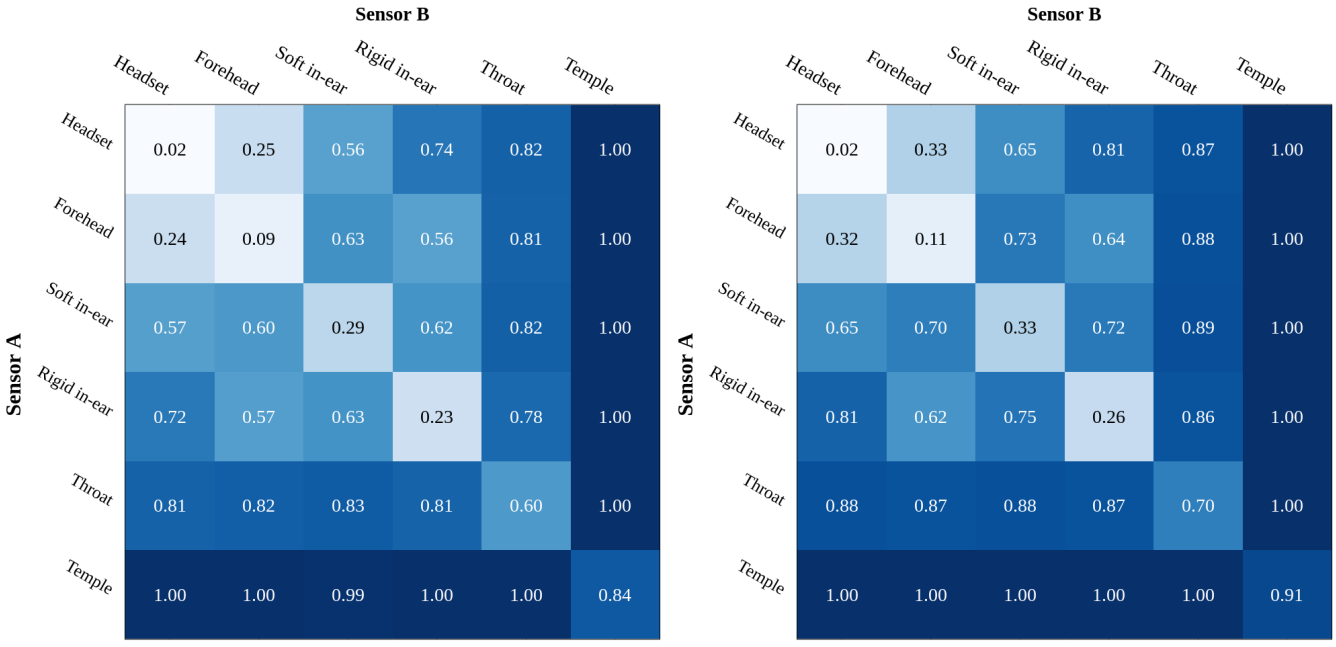
APPENDIX F
SPEAKER VERIFICATION MIN-DCF



(a) Mixed gender pairs

(b) Same gender pairs

Fig. 14. min-DCF obtained with the speaker verification model for sensor pairs



(a) Mixed gender pairs

(b) Same gender pairs

Fig. 15. min-DCF obtained with the speaker verification model for sensor pairs enhanced by EBEN