Algorithmes audio intelligents

# À propos de On-Hertz

On-Hertz est une startup technologique en pleine croissance qui crée des logiciels de production audio live pour l'industrie du broadcast (qui fait partie du secteur plus vaste des médias et divertissements). Nos clients sont actuellement principalement des stations de radio et de télévision d'Europe occidentale, mais la clientèle commence à s'étendre en dehors de l'Europe.

L'industrie du broadcast s'appuie encore largement sur des solutions matérielles pour produire du contenu en direct. Même si les outils vidéo évoluent plus rapidement vers des solutions logicielles, il y a un manque de solutions audio virtualisées sur le marché. L'orientation audio-centrée d'On-Hertz s'inscrit dans cette lacune du marché.

On-Hertz propose trois gammes de produits complémentaires:

* On-the-go (Lumo): applications bureautiques destinées aux créateurs de contenus (journalistes, animateurs radio, commentateurs sportifs, etc.) utilisées principalement pour la production à distance.
* On-demand (Nubo): services cloud également dédiés à la création à distance ou à la reprise après sinistre.
* On-prem (Artisto): destiné à être installé sur des serveurs et à s'intégrer au reste de l'infrastructure de diffusion et éventuellement à s'interconnecter avec les produits Lumo et Nubo.

# Contexte

L’industrie des médias connaît une transformation rapide, de nouveaux acteurs du monde numérique ont bouleversé les habitudes de consommation et la quantité de contenus audio et vidéo disponibles a explosé. Pour survivre à ces changements, les médias traditionnels doivent s’adapter: devenir plus efficaces, plus agiles et produire davantage de contenu pour davantage de plateformes différentes et ce avec des budgets réduits. Cela a évidemment un impact sur leur workflow et sur la qualité du contenu produit.

Même les productions professionnelles sont de moins en moins réalisées dans des conditions idéales. Par exemple, les studios sont de moins en moins isolés acoustiquement (privilégiant également l’aspect visuel par rapport à l’acoustique), et de plus en plus d’enregistrements sont réalisés en open space ou en extérieur. De plus, avec l’utilisation généralisée des téléphones portables et des vidéoconférences, la qualité audio des entretiens et interviews est souvent sous-optimale. Cela dépend également grandement de la qualité du réseau mobile.

Pour réduire les coûts, un ingénieur du son n'est affecté que si la production atteint une taille et un budget importants. Il est courant de voir des programmes (radio ou TV) réalisés sans assistance technique ou de voir un seul technicien devoir gérer toute la technologie, audio, vidéo, etc., potentiellement pour plusieurs productions simultanément.

On-Hertz s'inscrit parfaitement dans ces évolutions. Au-delà du simple remplacement des outils matériels existants par un équivalent logiciel, On-Hertz se positionne comme une solution permettant aux diffuseurs de repenser leurs workflows, de travailler de manière distribuée et d'introduire davantage d'automatisation.

En ajoutant des algorithmes intelligents à la gamme de traitements déjà disponibles dans son moteur audio piloté par API, On-Hertz vise à atteindre trois objectifs :

1. Gardez une longueur d'avance technologique sur la concurrence.
2. Améliorer la qualité sonore et maintenir la cohérence entre toutes les productions.
3. Assister les techniciens et les opérateurs, en suggérant des ajustements pour aider à accélérer le processus de production. Cet aspect est probablement très similaire au point précédent à la différence que les corrections seraient uniquement proposées à l'opérateur au lieu d'être appliquées automatiquement.

Par amélioration de la qualité et de la cohérence entre les différentes productions, nous entendons les choses suivantes :

1. Améliorer les voix capturées grâce au débruitage (y compris dé-écho), à la correction de la réponse en fréquence et à l'optimisation de la plage dynamique afin d'obtenir les voix les plus naturelles possibles. **L'objectif n'est pas nécessairement de réduire 100% du bruit, mais plutôt d'obtenir le meilleur compromis entre réduction de bruit et son « naturel »**.
2. L’automatisation du mixage de plusieurs sources entre elles afin d’obtenir le mix le plus cohérent et agréable possible.
3. L'application de corrections fréquentielles et dynamiques dans une perspective de branding (et donc avec une approche plus artistique) de certaines sources (généralement musicales) ou du mix final. Cela pourrait probablement être réalisé avec un algorithme similaire à 1.

L’objectif premier d’On-Hertz est de disposer d’algorithmes utilisables en temps réel. Il est cependant envisageable de proposer une solution hybride offrant la version temps réel la plus optimale possible et une version de traitement de fichiers offrant une qualité supérieure.

Nos cas d'utilisation sont:

1. Contribution: lorsque les invités sont mis en relation par téléphone ou par vidéoconférence.
2. Enregistrement « sur site » comme des événements sportifs, des débats, des podcasts, etc.
3. Capture dans des studios de radio ou de télévision (parfois plus bruyants qu'on ne l'imagine)
4. Capture dans des « cabines vocales »

Le cas de la contribution est probablement le plus complexe car rien n'est maîtrisé, ni l'environnement, ni le type et le positionnement du microphone ni le transport du signal qui fera également appel à des algorithmes de compression de données. Dans la plupart des cas, il s'agira d'une source unique.

La capture sur site est relativement similaire à la contribution. Dans ce cas, nous n'avons pas à nous soucier du transport des données. En revanche, il s’agira le plus souvent de gérer plusieurs sources. Même si les microphones sont souvent de bonne qualité, leur placement et leur utilisation ne sont souvent pas idéaux.

Pour l'enregistrement dans les studios de télévision (ou de radio), c'est probablement le plus idéal. Le type et le positionnement des microphones peuvent être contrôlés, et les protagonistes sont la plupart du temps des professionnels qui savent utiliser un microphone. En revanche, ces environnements présentent encore pas mal de bruits dus par exemple à la ventilation, au refroidissement des équipements (éclairage, informatique, mur vidéo, etc.), aux mouvements des équipements robotisés (caméras et éclairages), etc.

Les cabines vocales sont de petits studios, où une ou deux personnes peuvent s'asseoir, avec une acoustique relativement maîtrisée. Ils sont principalement utilisés pour des commentaires sportifs ou des enregistrements d'entretiens téléphoniques ou de voix off. Le principal défi dans certains cas est la mauvaise qualité des microphones utilisés (principalement des micro-casques) et le manque d'expérience dans l'utilisation des microphones.

# Qualité audio broadcast

Les professionnels de l’audiovisuel ont des attentes élevées :

* Besoin d'une large bande passante, entre 10 Hz et 24 000 Hz (celle-ci peut probablement être réduite pour le traitement de la voix).
* La fréquence d'échantillonnage est généralement de 48 kHz mais les algorithmes doivent pouvoir s'adapter à d'autres fréquences (principalement 44,1, 88,2, 96 et 192 kHz).
* Large plage dynamique (rapport signal/bruit élevé) >= 120 dB
* Faible latence (surveillance intra-auriculaire)
* Critères subjectifs : dans les usages professionnels, les sources audios doivent être les plus « propres » et naturelles possibles et les algorithmes doivent être les plus « transparents » possibles. Si pour des raisons artistiques, éditoriales ou de branding, les contenus doivent être « colorés », ces traitements sont généralement appliqués plus en aval de la chaîne. Cela permet au contenu d'être réutilisé avec différents traitements.

Les traitements appliqués à la voix doivent améliorer l'intelligibilité tout en préservant son aspect naturel. Dans le cadre du débruitage ou du désécho nous préférons ne pas viser un nettoyage total mais plutôt le meilleur compromis entre moins de bruit et un son naturel.

# Présentation du projet

Dans le cadre de ce projet, nous voulons nous focaliser sur les cas d’usage en contribution et en plateau télé présentés ci-dessus (point 1 et 3), et plus particulièrement sur le dé-bruitage et l’amélioration de la voix.

Voici une liste préliminaire des étapes envisagées:

1. Construire un dataset
2. Etat de l’art:
   1. Recherche des publications, algorithmes et publications existantes
   2. Evaluation de l’application [NVidia Broadcast](https://www.nvidia.com/fr-be/geforce/broadcasting/broadcast-app/) et de la librairie [Audio Effects SDK](http://ttps//docs.nvidia.com/deeplearning/maxine/audio-effects-sdk/index.html)
   3. Evaluation du noise-cancelling très efficace de Microsoft Teams
   4. Sélection des techniques et algorithmes permettant la réduction de bruit
   5. Sélection des techniques et algorithmes pour la synthèse ou reconstruction vocale
3. Entraînement des réseaux de neurones (le cas échéant)
4. Développements d’un premier POC si possible testable en situation réelle (idéalement en l’intégrant au moteur de traitement temps-réel de On-Hertz)

# Etat de l’art préliminaire

## Solutions existantes sur le marché

Les solutions suivantes ne semblent pas l'intelligence artificiel mais sont populaires dans notre secteur:

* <https://www.cedar-audio.com>
* <https://www.izotope.com>
* [Vocal Processing for Musicians with Izotope RX 10](https://www.linkedin.com/learning/vocal-processing-for-musicians-with-izotope-rx-10/cleaning-and-processing-vocals-with-rx-10)

## Solutions utilisant l’intelligence artificiel

[NVidia Broadcast](https://www.nvidia.com/fr-be/geforce/broadcasting/broadcast-app/) et son [Audio Effects SDK](http://ttps//docs.nvidia.com/deeplearning/maxine/audio-effects-sdk/index.html) -> Exemple mais je sais pas si c'est bien la version avec IA ou pas https://www.youtube.com/watch?v=7HB89DUgzRk

<https://www.waves.com/plugins/clarity-vx-pro>

<https://www.waves.com/plugins/clarity-vx-dereverb-pro>

<https://www.2wcom.com/ai-audio-enhancement-by-2wcom/>

<https://www.sonible.com/smarteq3/>

<https://www.izotope.com/en/products/vea.html>

[https://audo.ai](https://audo.ai/)

[https://cleanvoice.ai](https://cleanvoice.ai/)

<https://podcastle.ai/tools/remove-background-noise-from-audio>

<https://www.veed.io/tools/remove-background-noise-from-audio>

<https://www.lalal.ai/voice-cleaner/>

<https://freevstplugins.net/blue-lab-audio-denoiser/>

Open source

<https://jmvalin.ca/demo/rnnoise/> Lorenzo

<https://github.com/werman/noise-suppression-for-voice> Lorenzo

<https://github.com/Batlez/AMD-Noise-Suppression> Théo

<https://github.com/Dhriti03/Noise-Reduction>

<https://github.com/lucianodato/libspecbleach>

<https://github.com/sergree/matchering>

<https://github.com/ai-mastering>

<https://github.com/ai-mastering/mastering_comparison>

<https://magicmic.ai>

## Generative AI / speech synthesis

<https://ai.meta.com/blog/voicebox-generative-ai-model-speech/>

[https://play.ht](https://play.ht/)

[https://murf.ai](https://murf.ai/)

[https://www.resemble.ai](https://www.resemble.ai/)

[https://wellsaidlabs.com](https://wellsaidlabs.com/)

[https://elevenlabs.io](https://elevenlabs.io/)

<https://www.microsoft.com/en-us/research/project/vall-e-x/>

<https://audiocraft.metademolab.com/>

[https://www.superhifi.com](https://www.superhifi.com/)

[https://www.voicemod.net](https://www.voicemod.net/)

Open source

<https://github.com/ZDisket/TensorVox>

<https://github.com/facebookresearch/audiocraft>

<https://github.com/tuneflow/tuneflow-py>

<https://google-research.github.io/seanet/audiopalm/examples/>

<https://google-research.github.io/seanet/audiolm/examples/>

# Publications

<https://arxiv.org/pdf/2109.13731.pdf>

<https://www.christiansteinmetz.com/>

<https://ijcert.org/ems/ijcert_papers/V3I1103.pdf>

<https://arxiv.org/abs/2010.13154>

<https://op.europa.eu/en/publication-detail/-/publication/20b9f607-f7c4-11ea-991b-01aa75ed71a1>

<https://becominghuman.ai/deepfake-audio-with-wav2lip-263f0f0e84bc>

<https://c4dm.eecs.qmul.ac.uk/ProposedPhDTopics.html>

## Datasets

<https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research>

<https://github.com/Yuan-ManX/ai-audio-datasets-list>

Notes au vol réu avec le prof

1er délivrable -> Digérer l'état de l'art

-> Transparents, exemple pour avoir une bonne idée de ce qui se fait.

Pour chaque Algo -> Trouver code source utilisé

-> Trouver les ressources utilisées

-> résultats

-> Articles scientifiques

\*Pour évaluer les algos, il faut tester

Dataset -> Voir s'il faut utiliser un des datasets qui sont donnés ou s'ils en donnent eux ?

Si c'est un dataset qui vient du 2e lien -> Lequel on choisit pcq il y en a pleins différents avec des caractéristiques différentes.

Transfert Learning -> Adapter le réseau de neurones à des données qui viennent de chez eux

Rapport : - Intro à l'IA pour l'audio

- Mise en œuvre du projet : Data, algo,..

- Tests

Réunion :

- Interview -> cas d'usage -> quelqu'un chez lui et rajouter des bruits possibles

-> quelqu'un dans un bureau

Voir s'il faut trouver des sources déjà bruités ou bien un bon son et rajouter des bruits ( 2e soluton peut etre bien pour l'IA, on a la source et la target.

Dutoit a peut être accès à un ensemble de bruits etc...

Voir aussi par rapport à la distorsion etc...

Ils vont demander à des clients des exemples d'enregistrements qu'on pourra utiliser

Les échos et réverbérations sont des bruits convolutifs, pas additifs

Voir aussi quand on parle trop fort dans un micro -> saturation

Voir le dé-noiser de teams

Tableau avec une comparaison entre les techniques, les IA etc..

Tableau avec les différents types de bruits et les caractéristiques.

Partie scientifique du projet : Lien Dutoit

* Lien initial [Speech Enhancement Review: Krisp Use Case - Krisp](https://krisp.ai/blog/speech-enhancement-review-krisp-use-case/)

Multiple channels : - The multi-channel case involves two or more microphones (channels). In this case, the extra channel(s) contain information on the noise signal and can help to reduce the noise signal in the primary channel. An example of such a method is [*adaptive noise filtering*](https://www.researchgate.net/publication/2994278_Adaptive_Noise_Cancelling_Principles_and_Applications).

This technique uses a reference signal from the auxiliary (secondary) microphone as an input to an adaptive digital filter, which estimates the noise in the primary signal and cancels it out (see Figure 2). Unlike a fixed filter, the adaptive filter automatically adjusts its [*impulse response*](https://en.wikipedia.org/wiki/Impulse_response#:~:text=In%20signal%20processing%20and%20control,response%20to%20some%20external%20change.). The adjustment is based on the error in the output. Therefore, with the proper adaptive algorithm, the filter can smoothly readjust itself under changing conditions to minimize the error.

- Autre méthode -> beamforming

* Lien sous-jacent : [The Hard Side of Noise Reduction - Hardware Based Approach via Beamforming (krisp.ai)](https://krisp.ai/blog/hardware-beamforming-noise-reduction/)

A common application of wave interference is active noise cancellation (ANC) in headsets.

By leveraging wave interference, one may then try to use several microphones to reinforce the signal coming from the direction of the main speaker (constructive interference) while suppressing signals from other directions (*destructive* interference). This is what [**beamforming**](https://en.wikipedia.org/wiki/Beamforming) does.

One of the applications of beamforming technology is sound source localization, which estimates the position of one or many sound sources relative to a given reference point. Remarkably, this happens all the time in our ears.

We’re able to tell where a sound is coming from when we hear it because of a sound localization process that is partially based on binaural cues. More precisely, it’s because of the differences in the arrival time or the intensity of the sounds at the left and right ears. Such differences are used mainly for left-right localization in space.

* Lien sous-sous-jacent : [Krisp SDKs for Ensuring Voice Communication Quality & Noise Cancellation](https://krisp.ai/blog/voice-communication-quality-with-krisp-sdk/) & [World’s #1 Noise Cancellation SDK | Krisp](https://krisp.ai/developers/) -> Les 2 à voir

Single-channel : One of the first results in the monaural case is the [spectral subtraction method](https://www.semanticscholar.org/paper/Suppression-of-acoustic-noise-in-speech-using-Boll/04d4d26f0866a6e2c16d6666b66f7a67f9f0c526). There are various methods for this approach, but this is the idea behind the original method:

* Take the noisy input signal and apply a [short-time Fourier transform (STFT)](https://en.wikipedia.org/wiki/Short-time_Fourier_transform) algorithm
* Estimate background noise by averaging the spectral magnitudes of audio segments (frames) without speech
* Subtract noise estimation from spectral magnitudes of noisy frames
* Then, by using the original phases of the noisy frame spectrums, apply an Inverse Short-time Fourier (ISTFT) transform to get an approximated signal of clean speech (see Figure 3).

Another classical solution is the [minimum mean-square error](https://ieeexplore.ieee.org/document/1164453) (MMSE)

**Stationary noises** have a simpler structure. Their characteristics are mainly constant over time, such as fan noise, white noise, wind noise, and river sound.

**Non-stationary** noises have time-varying characteristics and are more widespread in real-life. They include traffic noises, construction noises, keyboard typing, cafeteria sounds, crowd noises, babies crying, clapping, animal sounds, and more. The traditional algorithms can effectively suppress stationary noises, but they have little to no effect when suppressing more-challenging non-stationary noises.

## **Deep learning approach**

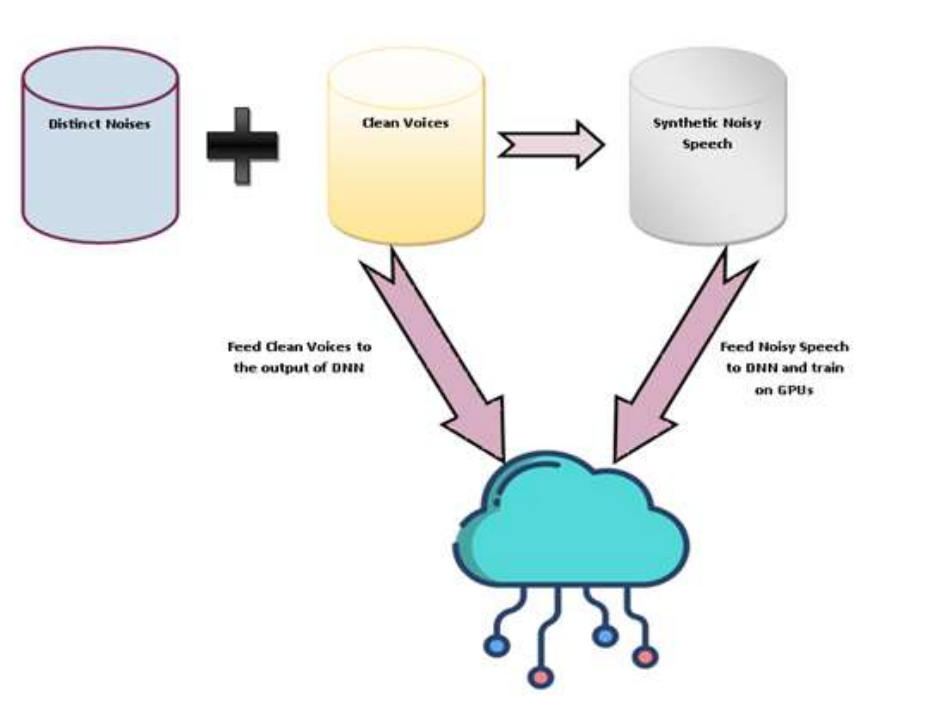
**Microsoft Teams ACTIVE NOISE CANCELLATION**

<https://aircconline.com/ijcsit/V15N1/15123ijcsit03.pdf>

<https://www.youtube.com/watch?v=DMtWR-SQd68&ab_channel=MikeTholfsen>

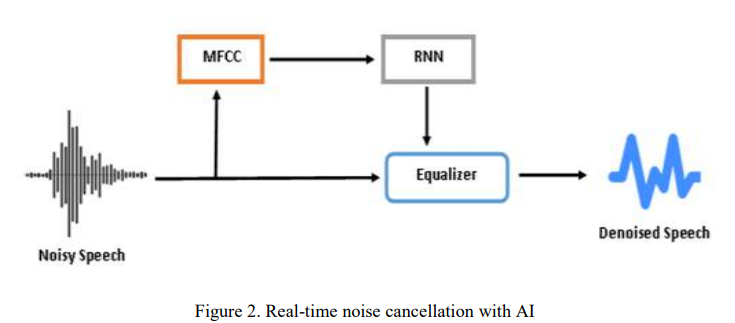
Ils utilisent AI & NLP POWERED ALGORITHMS

Voix claire et même voix claire avec un bruit “distinct” ajouter.



*”Cependant, la suppression du bruit ne traite que les bruits stationnaires, tels que le bruit généré par un ventilateur d'ordinateur ou un climatiseur en arrière-plan. Traditionnellement, la suppression du bruit pouvait être abordée en identifiant les pauses dans la parole, en estimant le bruit de fond, en supposant que le bruit de fond progressif reste constant, et en le filtrant [6]. Microsoft Teams travaille à la suppression des bruits non stationnaires, tels que l'ouverture d'une porte ou un chien qui aboie. Cette étape sera réalisée en utilisant l'apprentissage automatique, où un ensemble d'entraînement sera élaboré avec plusieurs bruits représentatifs.”*

“Les bruits qui ne seront pas filtrés comprennent les sons d'instruments de musique, l'annulation de bruit en temps réel, le chant et le rire. Les bruits seront laissés non filtrés car l'équipe de Microsoft ne peut pas séparer les sons des voix humaines. Ainsi, seuls les bruits distractifs seront éliminés. De plus, il est impossible de filtrer certains bruits car des bruits indésirables apparaissent dans les intervalles entre la parole et se superposent à la parole [8]. Lorsque le bruit d'une personne et sa parole se chevauchent, il n'est pas facile de les distinguer. Ainsi, Microsoft Teams formera un réseau neuronal pour différencier le bruit de la parole en utilisant l'intelligence artificielle.”



Pour combiner les données de bruit avec la parole, l'équipe utilise des scripts de synthétiseur qui intègrent différents rapports signal-bruit. Ils amplifient le bruit, simulant diverses situations réalistes rencontrées pendant les appels, en tenant compte des différences entre les appels de conférence et les livres audio. Le succès de la réduction du bruit par intelligence artificielle repose fortement sur l'apprentissage automatique, permettant au système de faire la distinction entre le bruit et la parole claire.

Le monde se compose de nombreux centres de modèles avec une grande expertise audio [25]. Les données doivent être en open source pour permettre des améliorations. La technologie permet aux employés de penser de manière créative dans le développement de diverses solutions pour relever les défis auxquels sont confrontés les modèles actuels [26]. L'expertise en apprentissage automatique recherche des connaissances auprès des personnes, et non uniquement à partir des données collectées.

RÉSULTATS

Cette recherche montre que l'intelligence artificielle peut filtrer les aboiements, la frappe sur un clavier et d'autres bruits lors des appels vidéo dans Microsoft Teams. L'utilisation de l'apprentissage automatique peut aider dans la suppression du bruit car il peut faire la distinction entre la parole et le bruit, ainsi qu'identifier et filtrer les bruits indésirables. De plus, un programme de formation peut aider à créer un ensemble d'entraînement avec plusieurs bruits représentatifs. Les modèles d'apprentissage profond nécessitent une infrastructure d'entraînement puissante et Microsoft Azure pour offrir à l'équipe une vision plus large du développement d'une version améliorée du modèle d'apprentissage automatique. Les technologies d'IA peuvent aider les utilisateurs à trouver les bonnes informations présentées au bon endroit pour résoudre une tâche spécifique. De plus, l'IA peut prendre des notes spécifiques et attribuer des tâches après la réunion, surveiller la tâche assignée, résumer le contenu et surveiller les délais. L'IA a le potentiel de rendre les appels vidéo plus efficaces et productifs et de réduire la frustration.