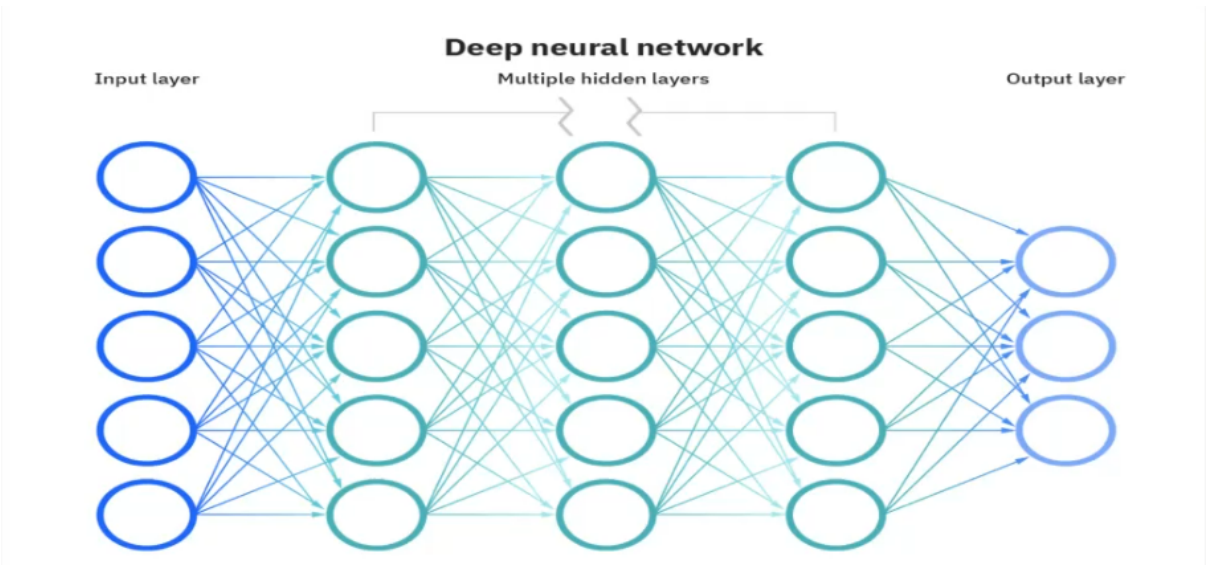
Compte-rendu de la réunion du département Son du 14 mars 2023 à 20 heures.

Au programme de la réunion de ce 14 mars 2023 : L'intelligence artificielle : quels bouleversements dans le monde de l'audio ? Présentation du projet Vox Protect et point sur les prochaines élections des représentants du département Son.

1. L'intelligence artificielle : quels bouleversements dans le monde de l'audio ?

Dans un premier temps, Christophe Henrotte, président de Studio Maia revient sur les grands principes qui régissent l'intelligence artificielle. L'IA est un ensemble de théories et de techniques développant des programmes informatiques complexes capables de simuler certains traits de l'intelligence humaine (raisonnement, apprentissage…). L'IA a ainsi besoin d'algorithmes et de données pour exister. L'apprentissage manuel fait toujours partie de l'intelligence artificielle avant de laisser place à l'apprentissage automatique aussi appelé machine learning. Il faudra attendre l'apparition du réseau de neurones profonds dans les années 2010 pour que l'intelligence artificielle opère un véritable tournant. Michel Monier explique ce qu'on appelle réseau de neurones profonds.



C'est un processus de traitement des données qui s'inspire du système neuronal humain. Michel poursuit en expliquant en détails le fonctionnement du système neuronal humain et

dans quelle mesure l'intelligence artificielle s'en inspire à travers le machine et le deep learning.
Le deep learning se distingue du machine learning par les stades de procession et d'apprentissage. Ces avancées technologiques intéressent beaucoup les GAFAM qui sont devenus des leaders de l'I.A. C'est ainsi que se sont développés deux procédés cristallisant chacun des enjeux différents. A commencer par la synthèse vocale. A partir de 2010, la synthèse vocale a pris une autre dimension après des années d'expérimentations plus ou moins infructueuses. Christophe explique comment il est possible de reproduire une voix, son émotion, et ce que cela nécessite en termes d'heures d'apprentissage par la machine. Toutes les caractéristiques de la voix peuvent ainsi être reproduites par ce procédé. Exemple est donné d'une chanteuse qui a conçu la majeure partie de son album à partir d'un avatar vocal reproduisant sa voix à la perfection. Le deep learning permet aujourd'hui un apprentissage de moins en moins lourd pour arriver à ce résultat. Christophe évoque ensuite la reconnaissance vocale et ses enjeux. Il revient sur les dérives de l'I.A. via plusieurs exemples dont un faux discours de Barack Obama. Les enjeux autour de l'I.A. appliqués à l"image et au son sont d'ordre éthique et technique. Une discussion a ensuite lieu autour des perspectives et des limites de l'intelligence artificielle appliquée au son.

Michel revient ensuite sur les mises en application de l'I.A. dans les systèmes de compression audio. Aujourd'hui, l'intelligence artificielle permet une nouvelle approche des systèmes. Le système opératoire peut être affiné en fonction du résultat final. Parmi les exemples donnés, les systèmes de compression comme Lyra ou Soundstream, ou encore Encodec récemment développé par Meta.

Lyra utilise le système neuronal pour un meilleur traitement des dialogues à bas débit. Soundstream est un système mixte qui permet de traiter la parole et la musique grâce à l'I.A. Méthode de compression audio qui fait intervenir l'IA, EnCodec, permettrait d'obtenir un taux de réduction de bande passante importante - qui peut aller jusqu'à 10 fois par rapport au MP3 - pour un signal de dialogue et de musique mono sans perte de qualité. Selon Meta, cette technique pourrait améliorer considérablement la qualité sonore des discours sur les connexions à faible bande passante, comme les appels téléphoniques dans les zones où le service est irrégulier. Les chercheurs de Meta auraient obtenu des résultats de pointe en matière de compression audio vocale à faible débit (1,5 kbps à 12 kbps), évalués par des annotateurs humains qui ont comparé plusieurs méthodes de compression, dont le dernier codec Lyra-v2 de Google avec la méthode non compressée et les ont classées en conséquence. Meta décrit sa méthode comme un système en trois parties entraîné à compresser l'audio à une taille cible souhaitée.

Tout d'abord, l'encodeur transforme les données non compressées en une représentation « d'espace latent » de fréquence d'image inférieure. Le quantificateur comprime ensuite la représentation à la taille cible tout en conservant la trace des informations les plus importantes qui seront utilisées ultérieurement pour reconstruire le signal original. (C'est ce signal compressé qui est envoyé sur un réseau ou enregistré sur un disque). Enfin, le décodeur transforme les données compressées en audio en temps réel à l'aide d'un réseau neuronal sur une seule unité centrale.

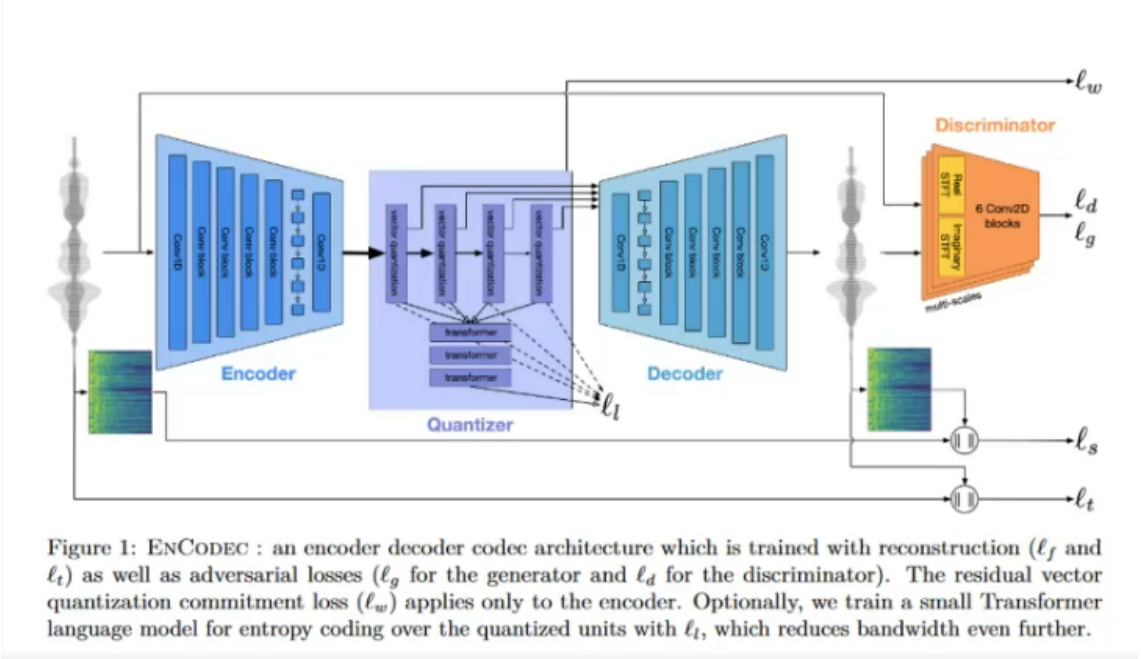


Figure 1: EnCodec : an encoder decoder codec architecture which is trained with reconstruction (ℓ_f and ℓ_t) as well as adversarial losses (ℓ_g for the generator and ℓ_d for the discriminator). The residual vector quantization commitment loss (ℓ_w) applies only to the encoder. Optionally, we train a small Transformer language model for entropy coding over the quantized units with ℓ_l, which reduces bandwidth even further.

Est ensuite évoquée la technologie RISC qui désigne un type d'architecture de processeur qui se caractérise par un jeu d'instructions visant la rapidité d'exécution grâce à la facilité de

décodage et d'exécution en pipeline des instructions machine. Encore exclue des marchés à fort volume (smartphones, ordinateurs etc…), Hans-Nikolas Locher remarque qu'il faut faire attention à la version de RISC que l'on utilise.

Sa disponibilité en open source en fait un terreau fertile pour l'I.A. L'IA doit être embarquée directement dans des composants proposant la capacité de mémoire et la puissance de calcul adéquats, hors d'atteinte des processeurs génériques, insuffisamment performants pour effectuer le calcul d'inférence IA en temps réel. L'enjeu se trouve dans l'architecture, ce qui place RISC-V au centre du jeu du fait de la capacité d'innovation qu'offre l'open source. L'innovation en IA et la nouvelle génération de processeurs offrent un nouveau paradigme dans l'expérience audio. Par un jeu d'apprentissage du ressenti de l'utilisateur, les algorithmes viennent adapter les performances au fil de l'eau, au point d'adapter les produits à la personne sans aucun préréglage. Le flux musical et les signaux de voix peuvent être propres aux habitudes et à l'audition de chacun, ce qui tend à changer l'usage des produits.

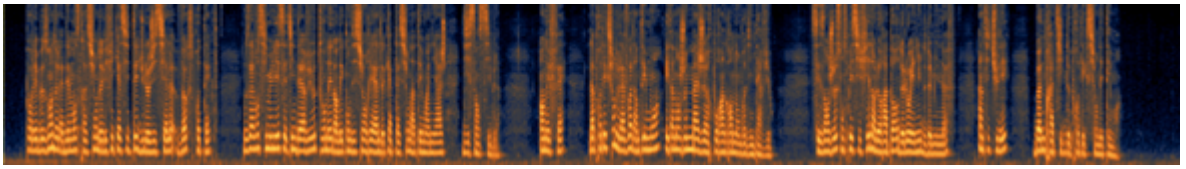
2. Présentation du projet Vox Protect

La transition est ainsi toute trouvée pour que Christophe Henrotte, directeur de studio MAIA, nous présente sa dernière innovation, VoxProtect, destinée à mieux protéger la voix des témoins. Témoignages télévisés, conversations privées, appels téléphoniques, notes vocales... toutes ces situations, pour la plupart quotidiennes, peuvent être exploitées à notre insu. Parce qu'il est possible de reconnaître l'identité, l'âge, le sexe, l'état de santé d'une personne grâce à sa signature vocale. La protéger est donc un enjeu majeur de notre société. Ainsi, le Règlement général sur la protection de données (RGPD) protège

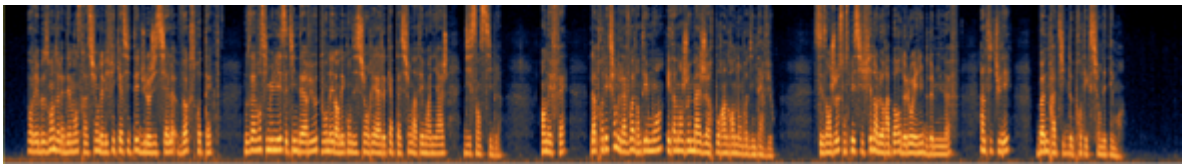
justement, à l'article 4[1], « la data personnelle », dont fait partie notre voix. Comme le droit à l'image d'une personne, le « droit à sa voix » fait partie, depuis un procès de 1982, de ce qu'on appelle en France « les droits de la personnalité », explique l'article[2] de la Commission nationale de l'informatique et des libertés (CNIL).
La CNIL a d'ailleurs rappelé dans un article publié le 4 janvier 2022, que la non ou mauvaise protection d'un témoin est « punie de 5 ans de prison et de 75 000 € d'amende ». Ainsi établie comme une donnée juridiquement et légalement protégée, le témoignage anonyme d'une personne doit rester anonyme, doit rester irréversible.

Protéger et garantir l'anonymat de la voix des témoins se doit donc d'être une priorité. Mais les nombreux effets linéaires, comme le pitch très majoritairement utilisé par les professionnels du son, ne protègent en rien les témoins, leur voix étant facilement retraçable. L'effet utilisé de pitch est *« une mesure de protection extrêmement faible »*, assure la CNIL. *« Il n'est pas compliqué techniquement de modifier {la voix} dans le sens inverse pour se rapprocher rapidement de la voix réelle et ainsi pouvoir ré-identifier la personne »*. En prenant l'exemple de représentation spectrale ci-dessous, l'analyse avant et après du pitch montre que le signal n'a pas été modifié, puisqu'il n'y a aucune différence sur les harmoniques du signal.

Représentation spectrale de la voix originale :



Représentation spectrale de la voix « pitchée » :



L'année 2013, en France, a prouvé à quel point il était facile d'identifier la signature vocale d'un individu, avec comme simples outils un enregistrement et un logiciel de reconnaissance vocale. C'est l'affaire Cahuzac, mettant en cause l'ex-ministre du Budget, qui met en avant ce risque. Le fameux enregistrement téléphonique, instrument de sa chute, dévoilé par *Mediapart*, a été nettoyé par les experts du Service central de l'informatique et des traces technologiques (SCITT), puis convertie en courbes afin d'établir une signature vocale.

---

[1] https://www.servicesmobiles.fr/la-voix-est-une-donnee-personnelle-sensible-que-dit-le-rgpd-45787

[2] https://linc.cnil.fr/fr/les-droits-de-la-voix-12-quelle-ecoute-pour-nos-systemes

Ensuite, entre en jeu le logiciel Batvox, qui a permis de comparer l'enregistrement à des discours de l'ancien ministre Jérôme Cahuzac, et d'établir un degré de ressemblance. Aucun doute, donc, sur la possibilité de reconnaître l'identité d'une personne grâce à sa voix. Cependant, le risque est largement plus important pour les personnes témoignant dans des affaires dites « sensibles » ou souhaitant conserver leur anonymat pour leur sécurité.
En clair, une voix de témoin pitchée peut être facilement identifiable en inversant le processus. La voix des témoins n'est pas protégée. Christophe Henrotte explique que son équipe a travaillé plus de quatre ans avec des chercheurs pour développer un plug-in qui garantit la protection de la voix des témoins. Ces travaux étant concluants, ils ont créé la société Swealink et développé le produit Vox Protect afin de le commercialiser. Ils ont envisagé plusieurs pistes de réflexion, notamment la resynthèse vocale, qui est l'une des pistes couramment envisagées de nos jours. Le principe de la resynthèse vocale est assez simple. D'une part, on récupère le texte dit par le témoin par une brique de Speach2text et, d'autre part, on analyse la prosodie de la voix afin d'en conserver les émotions. Le texte et les informations de prosodies sont alors analysés et on reconstitue la nouvelle voix à partir d'un corpus distinct de voix existantes.

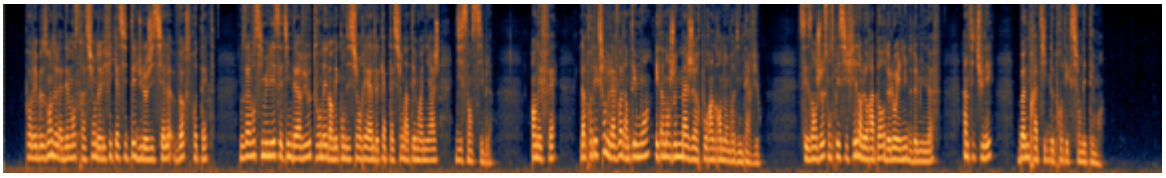
Cette approche présente plusieurs inconvénients, dont certains sont rédhibitoires dans le cas qui nous occupe, où l'on cherche que la voix du témoin soit protégée, mais également entendue par un spectateur (lors d'un journal télévisé ou sur un documentaire).

Pour reconstituer l'émotion de la voix initiale, il est nécessaire de réunir un corpus très important, avec plusieurs gigaoctets de voix disponibles. Ce corpus est également sensible à la langue, en fonction de l'accent de l'ancienne ou de la nouvelle voix. Il est donc probable qu'il faille stocker ce corpus en dehors de la machine de l'ordinateur sur lequel est fait le traitement, d'une part, et/ou qu'il faille envoyer les informations : voix originale, texte original, dans le cloud pour un traitement distant, ce qui crée une faille de sécurité. La puissance nécessaire pour faire fonctionner un tel process en temps réel impose des ressources machines également très importantes, parfois incompatibles avec ce qu'autorisent les outils de montages sons sur lesquels serait branché le plug-in. Une quête de performance viendrait à simplifier la prosodie, voire à créer des erreurs d'interprétation, ce qui pose des questions sur le plan déontologique. Par ailleurs, dans un contexte sensible de protection d'un témoin, nous pourrions rencontrer des limites en termes de droits d'auteurs et de droit moral du comédien qui a enregistré le corpus de resynthèse. A-t-il donné son autorisation pour que sa voix soit utilisée pour évoquer une attaque terroriste ? Pour l'ensemble de ces raisons, Christophe et son équipe ont cherché à concilier une approche plus classique, mais avec une originalité dans le traitement aléatoire qui rend la transformation indétectable.

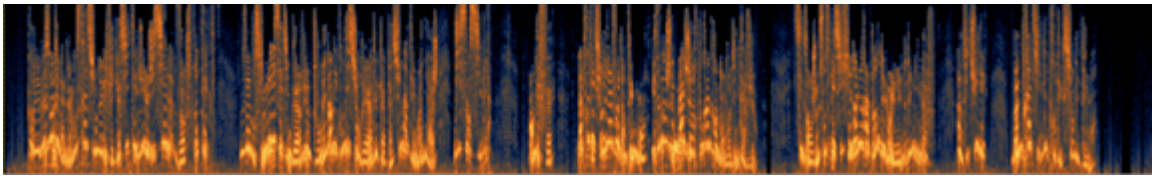
Ils ont donc retenu l'option d'un traitement par modulation aléatoire qui permet de travailler en temps réel. Il s'agit de faire subir au signal trois variations indépendantes les unes des autres. Ces variations fonctionnent sur un temps très court, de moins d'une seconde, et rendent impossible la réversibilité. Par ailleurs, le logiciel ne gardant aucune trace, la protection du témoin est ainsi assurée : si on passe la même voix deux fois à travers ce

processus de transformation, la voix transformée sera donc différente. Comme il s'agit d'une adjonction de transformations linéaires, on conserve la prosodie et donc l'émotion du témoignage. Si on reprend la comparaison entre la voix originale et la voix traitée, nous constatons que le signal a bel et bien subi des transformations importantes, protégeant donc le témoin d'une reconnaissance vocale par analogie des signaux.

Voix originale :



Voix traitée par le logiciel Vox Protect :



Deux niveaux de protection ont été mis en place pour s'adapter au mieux au besoin de chaque témoin. Le premier est le « Low Level of Protection » qui correspond à un niveau bas de protection, soit un témoignage sur des affaires courantes sans risque majeur autre que la réversibilité du processus.

Il protège d'une réversion effectuée par un pitch. On peut le renforcer par une légère variation

des modulations. En revanche, ce niveau ne protège pas d'une attaque par des services « spécialisés » qui ont accès à des outils beaucoup plus perfectionnés comme la reconnaissance vocale assistée par ordinateur.

Le second, le « High Level of Protection », permet, comme son nom l'indique, d'atteindre un niveau de protection plus important. Il est donc conseillé pour des cas comme le témoignage dans une affaire terroriste. En revanche, ce niveau nécessite un ajustement pour trouver le meilleur compromis entre protection et lisibilité du message. L'équipe de VoxProtect a la volonté de proposer des améliorations au fil de l'eau par des mises à jour régulières et des changements de versions, a priori tous les deux ans.

3. Point sur les prochaines élections des représentants du département Son.

L'année 2023 est une année élective pour la CST avec le renouvellement ou la reconduction de vos représentants de départements. Vianney Aube ne reconduit pas son mandat de représentant de département, il est remplacé par Erwan Kerzanet, nouvellement élu. Il sera accompagné de Michel Monier qui a décidé de se représenter pour un nouveau mandat. A leurs côtés, on trouve Christophe Henrotte, élu au poste de représentant adjoint.

Diplômé de l'École Nationale Louis-Lumière et du Conservatoire National des Arts et Métiers (CNAM), Michel Monier a démarré sa carrière comme responsable technique pour Paris Studios Billancourt et Télé Europe puis consultant pour Dolby. Depuis 2016, il est consultant postproduction Cinéma & TV. Michel se représente au poste de responsable du département son aux côtés d'Erwan Kerzanet Erwan Kerzanet étudie le son de manière transversale à l'École Nationale Supérieure Louis-Lumière (ENSLL) dont il sort diplômé en 1997. Il continue ensuite ses études en effectuant une recherche sur la photographie documentaire en soutenant un DEA d'histoire de l'art à l'université Paris X Nanterre (Lee Friedlander, The American Monument). Erwan Kerzanet travaille comme ingénieur du son auprès d'artistes français et internationaux parmi lesquels Leos Carax, Kiyoshi Kurosawa, Jacques Doillon, Amos Gitai, Karl Lagerfeld, Philippe Parreno ou Pietro Marcello. Il a été aux génériques de nombreux films de *L'Afrance* d'Alain Gomis à *L'Envol* de Pietro Marcello en passant par *Annette* de Leos Carax ou encore *Rodin* de Jacques Doillon. Compositeur, ingénieur du son, mixeur, Christophe Henrotte est un homme aux multiples casquettes et talents. Depuis 1993, il est également président de Studio Maia notamment spécialisé dans le doublage, le mixage, la composition sonore ou le RCS. Il est également à l'origine du logiciel Vox Protect dont l'objectif est de renforcer la protection des témoins par un meilleur encryptage de leurs voix.

4. Conclusion

Voilà qui conclut cette dernière réunion du département Son placée sous le signe de l'innovation et qui aura vu l'expression de nombreuses intelligences… toutes sauf artificielles !