École Nationale Supérieure Louis-Lumière
PROMOTION SON 2019

Mélia ROGER

THE VOICE IS VOICES

installation sonore autour du clonage vocal

Directeur interne : Alan Blum, ENS Louis-Lumière (Paris, France)
Directrice externe : Hanna Järveläinen, ICST (Zurich, Suisse)
Rapporteur : Eric Urbain, ENS Louis-Lumière (Paris, France)
Consultant : Nicolas Obin, IRCAM (Paris, France)

Mémoire soutenu le 12 juin 2019 à l'ENS Louis-Lumière.

Remerciements

Je tiens à remercier tous mes professeurs et camarades de classe qui m'ont soutenue dans ce projet, en particulier pendant mon échange universitaire au sein du Master Transdisciplinaire de la *Zurich University of the Arts* (ZHdK). Merci à Nicolas Obin et à l'IRCAM pour m'avoir permis de créer mon clone vocal. Merci à tous ceux qui m'ont hébergée pendant ces allers-retours entre Paris et Zurich.

Enfin, je remercie tout particulièrement :
Grégoire Bélien pour son travail d'éclairage et de vidéo et son soutien dans ma partie pratique.
Antoine Bertin pour m'avoir dit « tiens, on dirait ta voix » lors de mon premier essai de clone vocal en avril 2018.
Louise Buchart pour sa motivation et son aide lors de ma partie pratique.
Delphine Chapuis Schmitz pour ses retours littéraires et la découverte du travail sonore de Dominique Petitgand.
Julien Charpier pour son soutien dans mes expérimentations narcissiques.
Thomas Edelin, Alice Cagnat et Martin Peignier pour leurs retours et relectures.
Giulio Fiore pour sa patience lors de mes enregistrements sonores de pages Wikipédia.
Salomé Oyallon pour ses photographies masquées et son soutien pour ma partie pratique.
Klara Rundel et Gabriel Vieira pour leur présence dans notre appartement zurichois et nos conversations inspirantes.
Laurent Stellin pour sa disponibilité et ses solutions techniques.
Irene Vögeli pour m'avoir fait découvrir le travail d'Omer Fast.
Adrien Zanni pour son incroyable talent de programmation.
Mes parents pour m'avoir transmis leur passion de la communication et ma mère pour s'être déplacée jusqu'à Paris afin d'écouter ma voix de synthèse.

Résumé

La synthèse vocale a atteint un niveau où les voix artificielles sont au plus proche des caractéristiques humaines. Il est désormais possible d'implémenter n'importe quelle identité vocale dans une voix de synthèse, technique aussi appelée clonage vocal. De nouvelles entreprises proposent de créer son propre clone vocal, rendant cette technologie accessible aux particuliers. Deux exemples de ces services ont été étudiés dans ce mémoire. Il a été remarqué que les clones résultants ne sont malheureusement pas capables de restituer le timbre original de la voix. Cependant, dans un contexte de recherche, la synthèse vocale est capable de reproduire de manière crédible une signature vocale. Pour ce mémoire, ma voix a été clonée à l'IRCAM. Est-ce qu'un clone vocal peut rendre compte de notre identité ? En présentant *THE VOICE IS VOICES*, une installation sonore autour du clonage vocal, nous tentons de faire douter le spectateur sur ses capacités à discerner une « vraie voix » de sa jumelle artificielle.

Mots-clefs

Voix, clonage vocal, identité, synthèse vocale, *text-to-speech*, portrait, vérité, émotion, détournement, la vallée de l'étrange, clone.

Abstract

Voice technology has reached a level where it is possible to create credible speech synthesis, close to human characteristics. It is now possible to implement any vocal identity into an artificial voice, a technique also known as vocal cloning. New companies are offering to create someone's own vocal clone, bringing this technology accessible to individuals. Two examples of vocal clone services were studied in this work. It was noticed that the resulting clones, are unfortunately not capable of reproducing the original timbre of the voice. However, the research community produces speech synthesis that can credibly reproduce a specific vocal signature. For this work, I had my own voice cloned at IRCAM. Can a vocal clone give a glimpse of our identity ? By presenting *THE VOICE IS VOICES*, a sound installation about vocal cloning, we try to make the audience doubt its ability to discern a « real voice » from its artificial twin.

Key words

Voice, identity, vocal cloning, voice synthesis, text-to-speech, portrait, truth, emotion, hijacking, the uncanny valley, clone.

Sommaire

Introduction

> « *We are social beings by the voice and through the voice ; it seems that the voice stands as the axis of our social bonds, and that voices are the very texture of the social, as well as the intimate kernel of subjectivity.* »[1]
>
> DOLAR Mladen, *A Voice and nothing more*,
> Cambridge: MIT Press, 2006

La voix se distingue dans l'ensemble des sons qui nous entourent comme essence même d'être son porteur de sens. Notre écoute est voco-centrée, à la recherche d'indices de communication. Notre oreille cherche l'intelligibilité dans le contenu de la parole mais elle est aussi capable d'entendre l'intentionnalité d'un locuteur, indépendamment du signifiant. La voix est donc au cœur de nos relations sociales, elle est le liant de notre pensée vers l'extérieur, vers de possibles auditeurs et à travers elle, nous nous définissons dans un contexte social. Notre identité vocale, cette signature singulière, peut-être aujourd'hui clonée par une voix de synthèse. Dans ce mémoire, nous aborderons la manière dont la synthèse de sa propre voix peut nous donner un aperçu de notre identité.

En commençant par une description du fonctionnement mécanique de la voix parlée, nous nous pencherons ensuite sur les différents éléments qui rendent notre voix singulière. Ensuite, nous essayerons de donner un aperçu des différentes techniques de synthèse vocale, en aboutissant au possible clonage vocal et à ses potentiels détournements. Enfin, par la description des différentes étapes de création de l'installation sonore *THE VOICE IS VOICES*, nous tenterons de réfléchir sur les éléments qui permettent de différencier une voix organique de son clone numérique.

---

[1] traduction : « *Nous sommes des êtres sociaux par la voix et à travers la voix ; il semble que la voix soit l'axe de nos liens sociaux, et que les voix soient la texture même du social, ainsi que le noyau intime de la subjectivité.* »

I / 1. Voix et Identité

Il semble important, avant d'émettre une réflexion sur l'identité vocale, de réfléchir au concept même d'identité. Tout d'abord, nous pouvons distinguer une première notion qui serait celle de l'identité administrative, soit l'ensemble des données de fait et de droit qui permettent d'individualiser quelqu'un (date et lieu de naissance, nom, prénom, filiation, etc.). Ensuite, nous pourrions parler d'identité pour définir la singularité, c'est-à-dire ce caractère permanent et fondamental de quelqu'un, au sein d'un groupe, qui fait son individualité et ce qui permet donc de le distinguer d'un autre individu.

Comment l'individu se définit-il à travers sa voix ? L'identité vocale peut être décrite selon différents niveaux[2] :

- un facteur phonématique : correspond au timbre d'une voix, c'est-à-dire à l'ensemble des fréquences qui permettent de différencier une voix d'une autre. Ce facteur dépend des propriétés physiologiques et physiques de l'organe vocal du locuteur, ainsi que de son état émotionnel. Nous pourrons décrire le timbre d'une voix selon son caractère chaud, rauque, aigü, nasillard, granuleux, fort, voilé, doux…

- un facteur prosodique : correspond aux composantes de l'expression et du style, c'est-à-dire l'intonation et l'accent.

- un facteur linguistique : inclue l'intention et l'identité sociale du locuteur dans sa manière de s'exprimer, par un choix lexical, sémantique et syntaxique et selon une prononciation personnelle (emploi de certaines liaisons).

---

[2] DENIS, G.,Transformation de l'identité d'une voix. Rapport de stage DEA ATIAM, 2003.

L'identité vocale serait ainsi les caractéristiques sonores qui permettent d'identifier un locuteur. Par la composition fréquentielle de son timbre, de sa manière d'articuler (accent) et de se rendre expressive, la voix donne des informations liées au corps du locuteur : taille, sexe, âge, origine. Au delà de ces éléments sonores constitutifs de notre identité légale, notre personnalité s'entend par la singularité de notre voix, c'est-à-dire la façon dont elle est socialisée. Les tics de langage, nos disfluences[3], nos expressions renseignent sur notre milieu social, notre humeur et notre personnalité. Ces éléments sont modulables et permettent au locuteur de jouer de la voix, de se faire entendre comme il le souhaite en fonction du contexte. Notre voix permet donc d'endosser différents masques, jouer différents rôles, sans pour autant être totalement détaché de notre propre identité. À titre d'anecdote, nous pouvons noter que d'un point de vue étymologique, le mot « voix » est invariable[4] en français. Cette voix unique est alors plurielle, dépendante du discours, de l'intention, de notre émotion ou de notre talent d'acteur, preuve de la largeur de notre palette vocale.

Après avoir analysé le fonctionnement mécanique de la voix parlée, nous nous intéresserons à la façon dont notre voix exprime nos intentions en étant intimement liée à notre corps et comment elle nous permet de moduler notre discours en fonction des situations sociales. Par la suite, nous utiliserons ces connaissances pour comprendre comment nous pouvons synthétiser un modèle vocal, de son intelligibilité à son expressivité.

---

[3] Les disfluences sont les artéfacts liés à la parole improvisée, les marques de l'oralité tels les bégaiements, les hésitations et les bruits de bouche.

[4] Voix, nom féminin, invariable ; Voix humaine (définition Larousse) : faculté d'émettre des sons, en parlant de l'homme ; ensemble des sons produits par les vibrations périodiques des cordes vocales.

a. Fonctionnement mécanique de la voix parlée

La voix humaine est le produit de la mise en vibration de nos plis vocaux dans le larynx par l'air expiré par nos poumons. Cette vibration, appelée « phonation », est ensuite modulée par notre gorge, nez, bouche et langue pour former des mots et nous exprimer. La production d'un son vocal est donc un équilibre entre expiration, phonation et articulation, réalisé par un système complexe de différents organes qui collaborent à la mise en mouvement d'air, la vibration et la mise en résonance pour nous permettre de communiquer.

L'appareil vocal est d'abord un instrument à vent. Les muscles inspirateurs, dont le plus important est le diaphragme, remplissent les poumons d'air. Les muscles expirateurs vont ensuite souffler l'air dans le larynx. Les muscles inspirateurs et expirateurs régulent le débit d'air expiré par les poumons. Ces derniers servent de générateur pour créer le flux d'air, dont le volume et la pression sont plus importants que pour une respiration classique et qui sera à l'origine de la création du son. Pour produire la voix, le rythme respiratoire est considérablement modifié pour obtenir une inspiration courte, suivie d'une expiration allongée (qui correspond à la phonation). Les cordes vocales représentent un obstacle pour l'expiration de l'air appelée pression « sous-glottique » et ainsi, le débit d'air expiré par les poumons doit être plus important que pour la respiration au repos. Les muscles respiratoires doivent donc s'adapter pour produire cette élévation de pression, pour la maintenir pendant toute la durée de l'émission sonore et surtout pour la moduler en fonction des variations d'intensité, de tonalité et de timbre de la voix. La quantité d'air mobilisée est de 400 à 500 ml par respiration au repos alors qu'elle peut atteindre jusqu'à 1,5 litres lors de la voix parlée. En effet, dans la voix forte, un locuteur utilise de grands volumes pulmonaires (60 à 90 % de la capacité vitale)[5].

[5] Appareil phonatoire : URL (février 2019): http://phoniatriestrasbourg.free.fr/Site_6/Appareil_respiratoire.html

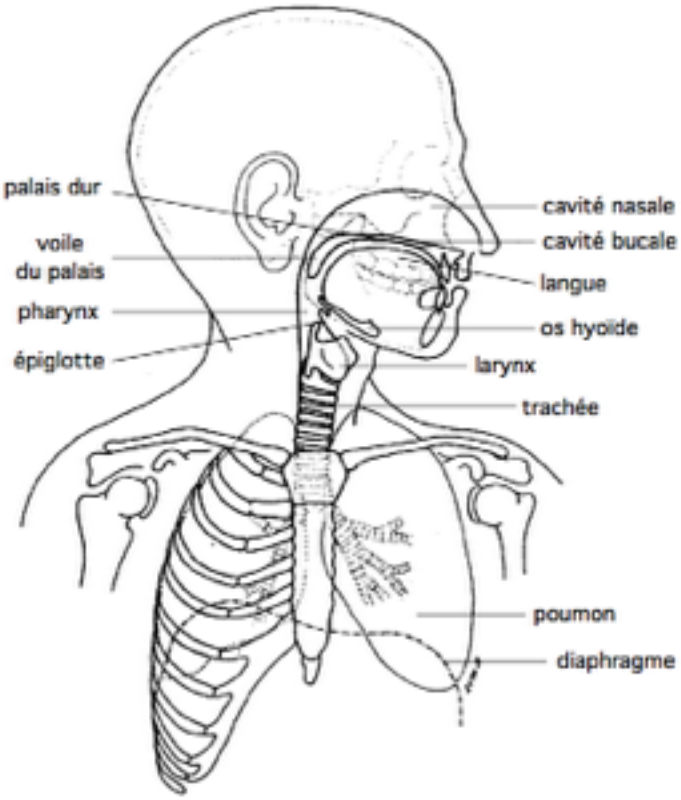


Figure 1 : Vue générale de l'appareil vocal[6] montrant les différents organes qui permettent la production de la voix

L'appareil vocal est également un instrument à cordes. Lorsque nous parlons, l'air expulsé par les poumons passe par la trachée, avant d'arriver au larynx où il fera vibrer les plis vocaux, aussi appelés cordes vocales. Ces dernières sont formées d'une paire de muscles et de ligaments de 20 à 25 millimètres de long, recouverts d'une muqueuse plus ou moins visqueuse. Elles sont tendues, attachées horizontalement dans le larynx. C'est en modifiant la position des cartilages (cartilage thyroïde et cartilages aryténoïdes, qui correspondent à la pomme d'Adam chez le garçon) accolés aux cordes vocales lorsque l'on parle, que l'on modifie alors leur longueur et leur position. Au cours d'une phrase, le locuteur produit différents sons ainsi modifiés par la fréquence de vibration des cordes vocales.

[6] HUSSON, R. *La voix chantée*, Paris, Gauthier-Villars, 1960.

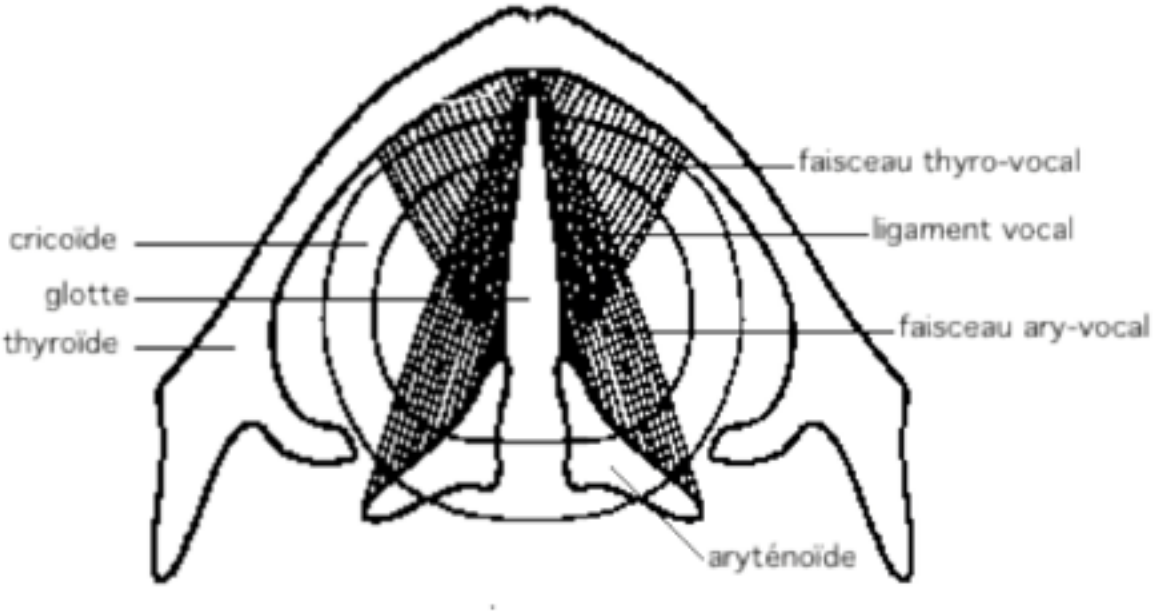


Figure 2 : Structure des plis vocaux[7]

Les sons émis par les vibrations de nos plis vocaux doivent encore être amplifiés pour être ensuite sculptés par nos modulateurs. Cette amplification est le rôle des résonateurs (fosses nasales, larynx et bouche), qui communiquent avec le pharynx (zone de rencontre entre les voies respiratoires et digestives). L'amplification va favoriser certaines fréquences et en atténuer d'autres. Elle filtre cette onde complexe émanant du canal vocal.

[7] HABERMANN, G. *Stimme und Sprache. Eine Einführung in ihre Psychologie und Hygiene*, Stuttgart, Georg Thieme Verlag, 1978.

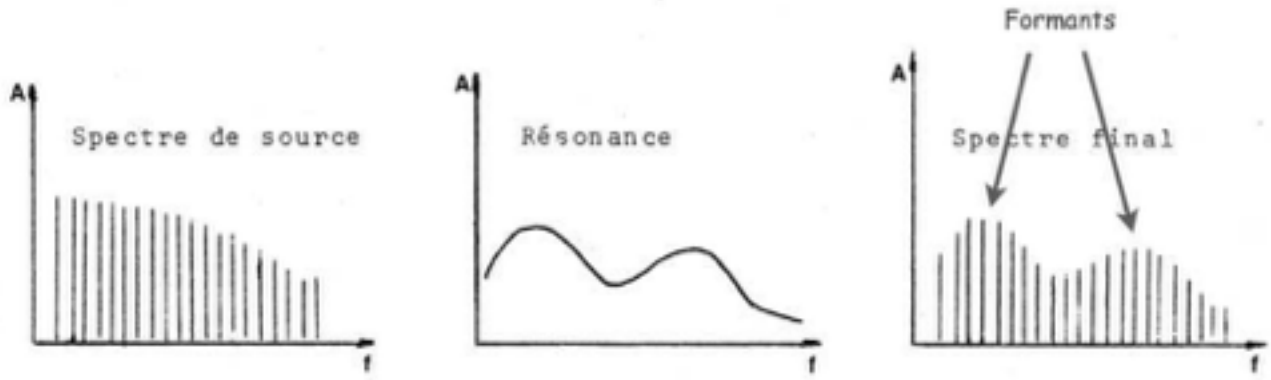


Figure 3 : schéma de la formation d'un formant, suite à la production d'un son complexe depuis les cordes vocales, vers la mise en résonance par les différentes cavités

Sur la Figure 3, nous pouvons observer que le spectre source fourni par les cordes vocales (premier graphique) comporte différentes harmoniques (spectre complexe). L'interaction entre la fréquence de vibration des cordes vocales et les résonances du tractus vocal d'un individu (second graphique), privilégie certaines harmoniques, créant les formants (troisième graphique). Les deux premiers formants jouent un rôle important dans la perception auditive des voyelles.

Cette amplification est ensuite complétée par la transformation fréquentielle des modulateurs (position du voile du palais, de la langue, des lèvres et des dents) qui agissent tels des filtres et articulent les sons pour former des phonèmes[8] et ainsi permettre la parole. L'appareil buccal est donc le principal organe qui permet la flexibilité et la souplesse de la voix.

[8] Phonème (définition Larousse) : élément minimal, non segmentable, de la représentation phonologique d'un énoncé, et dont la nature est déterminée par un ensemble de traits distinctifs. Autrement dit, c'est la plus petite unité discrète que l'on puisse isoler par segmentation dans la chaîne parlée.

Si le timbre de notre voix dépend de caractéristiques morphologiques et physiologiques comme nos résonateurs (taille des fosses nasales, positionnement des amygdales, taille du crâne et du massif facial), cela implique que notre timbre vocal nous est singulier et tributaire des caractéristiques de notre appareil phonatoire. Telle une empreinte digitale, notre timbre est unique et peut ainsi témoigner de notre identité. Le site appelé *pink trombone*[9] permet d'écouter les différentes influences de notre appareil vocal sur la voix. D'une certaine manière, notre timbre n'est pas un usage du corps mais est le corps lui-même. Toutefois, notre voix étant dépendante de notre corps, elle se modifie selon notre âge et nos habitudes de vie.

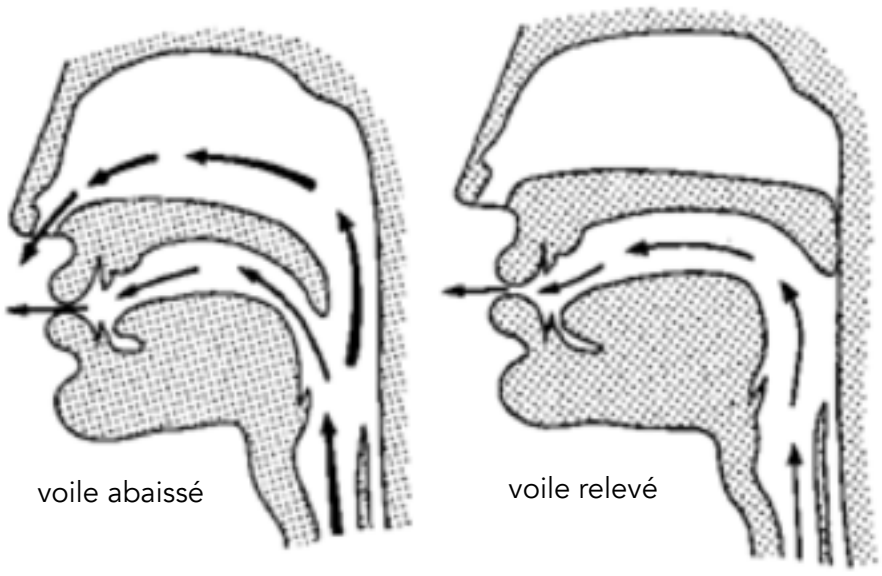


Figure 4 : Action du voile du palais sur le trajet de l'air expiré et donc sur le son prononcé[10]

[9] URL (mars 2019) : https://dood.al/pinktrombone/

[10] LEOTHAUD Gilles, *Théorie de la Phonation*, Cours de DEUG 2ème année, 2004.

b. La voix comme corps : transformations et socialisation

Langue maternelle

Notre manière d'articuler, de faire sonner certaines consonnes et voyelles, dépend de notre langue maternelle. L'apprentissage d'une langue éduque notre articulation, de la façon dont notre langue vient se coller au voile du palais à sa position derrière les dents. Ces habitudes phonétiques constituent notre manière propre de prononcer les mots, à la fois dans notre langue maternelle mais également dans l'établissement de notre accent lorsque nous parlons une langue étrangère. Le babil des nourrissons (entre 3 et 7 mois) permet au langage de se former par imitation, un bébé japonais n'ayant donc pas le même développement de babil qu'un bébé français. C'est ce qu'explique Bénédicte de BOYSSON-BARDIES, en mentionnant qu'« *une première organisation différentielle des capacités du nourrisson se fait à partir des données acoustiques de la langue maternelle. La structure phonétique, syllabique et rythmique de la langue parlée dans l'environnement va moduler l'espace perceptif de l'enfant.* »[11]. Le contexte d'éducation vocale du nourrisson influencera donc sa manière de s'exprimer.

La mue

De notre enfance à l'âge adulte, notre voix subit un changement radical au moment de la mue. Cette étape a lieu au moment de la puberté et est particulièrement marquée chez les garçons. La mue correspond à des modifications anatomiques, physiologiques et psychologiques dues à un changement hormonal pendant la puberté. Au niveau du larynx, les cartilages s'agrandissent et chez le garçon, le cartilage thyroïde change de forme (correspondant au relief de la pomme d'Adam), les muscles se renforcent, la partie membraneuse des cordes

---

[11] DE BOYSSON-BARDIES Bénédicte, *Comment la parole vient à aux enfants*, Article, Les Cahiers du MURS, 1998. Voir aussi : *Influence des langues-cibles sur le développement de la parole : études comparatives sur des enfants de 6 à 10 mois*, Publications de l'Université de Rouen, 1984.

vocales s'accroît et le registre vibratoire des plis vocaux se modifie, utilisant davantage le registre de poitrine. Aux modifications laryngées s'ajoutent la modification des résonateurs : croissance du crâne et du massif facial, rétraction des amygdales, développement des fosses nasales par croissance du nez. De plus, la cage thoracique s'élargit et les volumes d'air mobilisés pendant la parole sont plus importants. Durant quelques mois, la voix subit donc une mutation, passant du registre de la voix de tête de l'enfant au registre de poitrine de l'adulte. Chez le garçon, la mue correspond à un abaissement d'une octave tandis que chez la fille, la mue peut paraître inaperçue, par un abaissement d'une tierce seulement et d'un enrichissement d'harmoniques graves. Après la puberté, notre tessiture reste globalement stable, jusqu'à la fin de notre vie.

Vieillissement du corps, vieillissement de la voix

Le mécanisme vocal faisant appel à différents organes, ces derniers subissent aussi un vieillissement avec l'âge. Ainsi, la voix d'une personne âgée en bonne santé pourra se modifier, suite aux changements anatomiques et physiologiques de son appareil phonatoire. Plus précisément, les changements des structures du larynx incluent une détérioration histologique (c'est-à-dire des tissus) des plis vocaux, une calcification et une ossification des cartilages, une perte d'élasticité des ligaments et une atrophie des muscles et par conséquent, des plis vocaux. Ces modifications peuvent mener à une réduction d'amplitude et de vitesse de vibration des plis vocaux, une diminution de l'adduction des plis vocaux et une augmentation de la tension des muscles du larynx, surtout chez les hommes[12]. Nous pouvons noter que la voix s'atténue car les muscles inspirateurs et expirateurs ne permettent plus de maintenir le débit d'air suffisant et constant pour une voix pleinement portée. Ce phénomène d'abord musculaire a un impact sur le timbre du locuteur, qui

[12] LORTIE Catherine, *Le vieillissement de la voix : De la production à l'évaluation*, Thèse sous la direction de TREMBLAY Pascale, GUITTON Matthieu, Université LAVAL, Canada, 2017.

devient plus rugueux, amplifiant de manière plus faible les fréquences de la voix. Une telle transformation peut-être entendue sur le site *L'Encyclopédie de la parole*, en écoutant trois voix différentes de Marguerite Duras correspondant à trois âges : claire, puis aggravée par le vieillissement et le tabac, et enfin après avoir subi une trachéotomie[13].

Changements ponctuels de la voix

Notre voix est notre corps et subit avec lui ses changements, étant altérée de manière plus ou moins prononcée. Les habitudes de vie ayant un impact physique sur nos organes, elles modifient par conséquent notre timbre. Ainsi, nous pourrions évoquer la consommation de tabac à long terme qui favorise l'inflammation des cordes vocales dues à l'irritation des muqueuses qui les entourent. Cette détérioration pouvant s'accompagner par un dérèglement hormonale a un impact audible sur la voix, qui peut changer de tessiture et paraître davantage rocailleuse. Ce changement de timbre voire de tessiture est une transformation progressive de la voix. Cependant, une simple rhino-pharyngite peut influencer nos caractéristiques vocales sur un court terme. Dans cet exemple, nos cavités nasales étant obstruées, nos résonateurs sculptent notre voix de manière différente et change ainsi notre timbre de manière ponctuelle, ne nous permettant pas d'articuler l'ensemble des phonèmes de manière intelligible, d'où l'expression « parler du nez ».

Socialisation de la voix

Genre, âge, origine et santé sont les éléments liés à notre identité qui nous sont difficiles voire impossibles à masquer. Notre voix est pourtant manipulable, flexible, s'adaptant au contexte, à nos locuteurs, aux situations sociales, pouvant nous faire porter de nouveaux masques. La voix est un élément social, son expressivité fait partie d'un processus culturel inconscient qui nous définit et nous permet d'osciller entre nos différents rôles. La socialisation de la voix s'acquiert par imitation, par

[13] URL (février 2019) : https://encyclopediedelaparole.org/fr/taxonomy/term/123#notice

expérience de divers situations de communication, influencées par notre environnement. Ainsi, même si nous avons une voix caractéristique de sexe masculin, nous pouvons avoir une manière féminine de nous exprimer. Par exemple, les personnes trans « male to female » (suivant une transformation d'apparence masculine vers une apparence féminine) peuvent chercher à féminiser leur voix, sans pour autant essayer de changer de tessiture. Parler comme une femme relève donc de caractéristiques culturelles, souvent stéréotypées afin d'être reproduites, comme par exemple : le fait de parler à un niveau sonore en dessous de celui des hommes, marquer les hésitations ou faire entendre son sourire dans la voix. Ces "manières" de parler sont à aligner avec le choix du vocabulaire, pour que l'enveloppe de ce qui est dit soit en accord avec le contenu de la voix. La voix est donc un outil sociolinguistique et donc dans notre société aujourd'hui, est également genré[14], qui nous définit dans nos relations à autrui. Notre signature vocale est singulière et pourtant cette vibration est unique à chaque fois qu'elle est articulée.

Voix et physionomie

Si des caractéristiques vocales en fonction du corps peuvent être généralisées, il n'existe pas de rapport certain entre le genre, la taille, l'âge et la voix. Un homme peut avoir une tessiture plus haute que celle d'une femme comme un individu de grande taille peut avoir une voix plus ténue que celle d'une personne de petite taille. On peut d'ailleurs ici mentionner le cas particulier des castras avant le XIXème siècle, qui étaient des hommes dont l'apport de testostérone avait été coupé par castration et empêchait ainsi la mue au moment de la puberté. Les hommes gardaient donc leur voix d'enfant, c'est-à-dire de tête à l'âge adulte[15]. Nous pouvons entendre une voix de castra dans le film *Farinelli*

---

[14] ARNOLD, Aron, *La voix genrée, entre idéologies et pratiques – Une étude sociophonétique*. Linguistique. Université Sorbonne Paris Cité, 2015.

[15] URL (avril 2019) : archive sonore d'une voix de castra : https://www.youtube.com/watch?v=_6FNKXtt95k

(CORDBIAU Gérard, 1994)[16] où, pour fabriquer cette voix si particulière, le département « Analyse et Synthèse de la Voix » de l'IRCAM a fusionné la voix d'une soprano avec la voix d'un ténor, pour établir un mélange entre les deux timbres et recréer cette voix masculine juvénile.

Figure 5 : Bruno Choël (à gauche),
voix française de Johnny Depp (à droite)[17]

Associer une voix et un corps différents est phénomène courant pour le cinéma, par le principe de doublage. Le mémoire de fin d'études de l'ENS Louis-Lumière de Thomas Edelin[18] mentionnait le talent de l'acteur français Bruno Choël, voix française de d'Ewan McGregor, Johnny Depp, Matthew McConaughey ou encore Joaquin Phoenix. Cette comparaison de faciès est intéressante car même si l'acteur peut éventuellement ressembler aux autres corps qu'il double, cette même

---

[16] CORDBIAU Gérard, *Farinelli*, 111 minutes, couleur, 1994.

[17] URL (mars 219) : https://www.youtube.com/watch?v=bi6srciWT1c compilation des voix de Bruno Choël, sa voix étant donc superposée à différents acteurs.

[18] EDELIN, Thomas, *Signature : une expérience sociale et interactive autour de la voix humaine*, mémoire (sous la direction de Thierry Coduys et Nicolas Obin), Son, ENS Louis-Lumière, 2017.

voix incarne plusieurs corps. Qui se souvient avoir sursauté en entendant la nouvelle voix française de Chandler dans la série *Friends*, qui, après 8 saisons avec le même comédien, change de voix à la Saison 9[19]? Dans les films d'animation, il est fréquent que les comédiens incarnent plusieurs personnages ou du moins, d'autres petits rôles en plus d'un personnage principal, modifiant ainsi leur voix pour se donner une nouvelle identité.

Aujourd'hui, les émissions radiophoniques sont filmées, enlevant à la radio cette magie des voix *sans corps*. La radio serait le masque le plus pur ? Voix célèbre de France Inter, dans son émission *Allô Macha*, Macha Béranger était une voix française les plus reconnues par le public, notamment par son timbre grave et profond. Ces voix reconnaissables, comme celle de la SNCF, prononcée par Simone Hérault, s'incarnent dans l'air, sans être ancrée sur un visage. D'après la SNCF, 2 français sur 3 associent cette voix à la compagnie de train. Associer un visage à une voix peut être déroutant, car notre imaginaire n'associait peut-être pas cette vison de la voix.

Figure 6 : Simone Hérault, comédienne incarnant
la voix de la SNCF depuis 1981

[19] URL (février 2019), ICHER Bruno, « Les *Friends* perdent des voix », *Libération*, 30 août 2003 :
https://www.liberation.fr/medias/2003/08/30/les-friends-perdent-des-voix_443347

À la manière de leur projet *The Next Rembrandt*[20], qui avait pour objectif de reproduire un tableau à la manière de Rembrandt (1607-1669), *ING and JWT Amsterdam* a permis à un groupe de recherche en Intelligence Artificielle (IA) de reproduire la voix du peintre. Son timbre a été re-créé à partir de l'analyse de ses autoportraits, afin de modéliser son appareil vocal. Ce qu'il allait dire a été défini par l'analyse de ses lettres manuscrites, afin d'extraire son vocabulaire courant ainsi que le lexique de l'époque (ancien néerlandais) mais également de pouvoir définir ses traits de personnalité afin de décrypter "sa manière de parler". Basée uniquement sur l'analyse de la physionomie du visage de l'artiste, la voix générée par ce groupe de recherche peut-être écoutée ici dans un premier tutoriel (en néerlandais)[21] : https://www.youtube.com/watch?v=8_A0sZi1ZEQ&feature=youtu.be. Je me permets ici d'émettre un doute sur la rigueur de cette démarche, sachant que le visage et l'appareil vocal n'ont pas de concordances définies mais propres à chacun, faisant de notre voix, un élément unique de notre personnalité, mais je trouve important de mentionner ce projet de création vocale, uniquement à partir de données visuelles.

Notre voix est notre masque social le plus flexible. Comment ces différentes identités s'expriment-elles ?

---

[20] URL (mars 2019) : *The Next Rembrandt,* utiliser une Intelligence Artificielle pour recréer une toile du maître hollandais https://www.youtube.com/watch?v=IuygOYZ1Ngo

[21] URL (mars 2019) : JWT Amsterdam, *« Learn how to paint from the Master himself in The Rembrandt Tutorials »,* 28 février 2019 : https://jwt-amsterdam.pr.co/171584-learn-how-to-paint-from-the-master-himself-in-the-rembrandt-tutorials

c. Expressivité de la voix, notre agent émotionnel

Notre voix dépend de notre corps ; sa flexibilité et sa richesse nous permettent de nous faire entendre comme nous le souhaitons, en fonction du contexte et de notre interlocuteur. La voix existe sans la parole mais la parole n'existe pas sans la voix. C'est avant tout dans son enveloppe (et ensuite dans son contenu) que la voix exprime nos émotions : un même mot peut être prononcé d'une multitude de façons différentes et pourra ainsi faire entendre différentes intentions pour nous permettre de communiquer justement.

En plus des vocalisations non linguistiques qui permettent d'exprimer différentes émotions (souffle expiré de fatigue, aspiration de surprise, rire de joie, gémissement de plaisir ou autres onomatopées…), notre parole est modulée par des paramètres acoustiques qui transmettent notre intention. Dans la littérature, les différentes recherches ne prennent pas en compte les mêmes paramètres dans l'établissement de règles pour exprimer les émotions dans la parole. Néanmoins, toutes les études s'accordent sur l'importance de paramètres principaux sur la prosodie, comme la hauteur et la variation de la fréquence fondamentale (F0) de la voix, sa vitesse d'élocution et son volume. Certaines études vont plus loin en incluant d'autres paramètres comme les pauses entre les mots (rythme d'élocution), la prosodie (c'est-à-dire la mélodie de la phrase) et la qualité d'élocution[22]. Ces paramètres permettent de moduler nos propos et transmettre un sous-texte, comme par exemple faire entendre du sarcasme ou de l'ironie.

Ces caractéristiques acoustiques ont un lien étroit avec nos expressions faciales. Par exemple, notre sourire modifie ainsi l'aspect fréquentiel de notre voix et par apprentissage culturel, nous associons un

[22] SCHRÖDER Marx, E*motional Speech Synthesis: A Review*, DFKI Saarbrücken, Institute of Phonetics, University of the Saarland, 2001.

certain spectre vocal à la joie car nous entendons inconsciemment un sourire[23]. De plus, entendre une voix souriante engendrerait un sourire sur le visage de l'auditeur. Voix et visage sont ainsi liés ; la voix fonctionnant tel un second masque émotionnel. Cette hypothèse permettrait aujourd'hui d'expliquer comment fonctionne la communication entre deux personnes ne partageant pas la même langue et venant de cultures très différentes. L'intention d'un locuteur peut être premièrement entendue dans sa voix, avant d'être confrontée à son expressivité corporelle. Ce qu'on appelle aujourd'hui « communication non verbale » fait référence aux émotions transmises par le comportement, et donc par le corps. Une voix peut dire « je suis triste » avec un ton perçu comme joyeux. C'est en ce creux que siège l'analyse de la communication non verbale. Le choix de la dénomination dite "non verbale" met bien en avant le fait qu'une émotion peut être oralisée sans pour autant être verbalisée. Le choix du contenu du discours devient alors un élément de référence tout aussi important que l'émotion oralisée, car il peut créer une distortion de perception.

C'est ce qu'a étudié le groupe de recherche CREAM à l'IRCAM depuis 2015 en analysant la prosodie autour du mot « bonjour », les chercheurs ont tenté de décrypter l'intention sociale d'un locuteur. Grâce à système de vocodeur de phase (développé sous le nom C.L.E.E.S.E.[24]), ils ont réussi à synthétiser plus de 70 000 « bonjour » différents, prononcés par la même voix[25]. En établissant la "fonction inverse" de la manière qu'a notre voix pour exprimer nos émotions, ils ont réussi à créer un algorithme qui permet de donner une couleur émotionnelle à n'importe quelle voix. Cet algorithme appelé D.A.V.I.D. (Da Amazing

---

[23] ARIAS Pablo, BELIN Pascal, AUCOUTURIER Jean-Julien, *Auditory smiles trigger unconscious facial imitation*, article, Current Biology, Vol. 28 (4), 2018.

[24] URL (février 2019) : C.L.E.E.S.E (Combinatorial Expressive Speech Engine) http://forumnet.ircam.fr/fr/produit/cleese/

[25] URL (février 2019) : http://cream.ircam.fr/

Voice Inflection Device) fonctionne comme un vocodeur émotionnel en modifiant les émotions prononcées par un locuteur. De plus, son fonctionnement est en « temps réel » car le *delay* entre la voix prononcée et celle modifiée peut être inférieur à 15 ms et peut être donc inaudible.

Ainsi, nous pouvons modifier l'intentionnalité d'une voix parmi trois émotions principales : joie, tristesse et peur. Nous pouvons alors entendre notre propre voix sonner plus joyeuse que la manière dont nous la prononçons acoustiquement. L'équipe de recherche a donc établi les liens entre les caractéristiques acoustiques que l'on associe aux émotions.

		Transformations		
Effects		Happy	Sad	Afraid
Time-varying	Vibrato			✓
	Inflection	✓		✓
Pitch shift	Up	✓		
	Down		✓	
Filter	High-shelf ("brighter")	✓		
	Low-shelf ("darker")		✓	

Figure 7 : Liste des effets audios utilisés dans l'algorithme et comment ils sont combinés pour former les transformations émotionnelles entre joyeux, triste et apeuré[26]

[26] RACHMAN, Laura, LIUNI Marco, ARIAS Pablo, LIND Andreas, JOHANSSON Petter, HALL Lars, RICHARDSON Daniel, WATANABE Katsumi, DUBAL Stéphanie, AUCOUTURIER Jean-Julien, *DAVID: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech*, publication 3 avril 2017.

Comme nous l'avons vu avec le sourire, notre expressivité vocale serait donc étroitement liée à nos expressions faciales. Si l'on se place dans le contexte audiovisuel, c'est la synchrèse[27] qui nous permet d'entendre des émotions perçues par la voix lorsque l'on voit un visage synchronisé à une voix qui peut exprimer une émotion différente que celle qu'exprime le visage. L'importance de cette simultanéité du visuel et du sonore influence notre perception de la prosodie et dans le contexte audio-visuel, du jeu d'acteur par exemple. Au cinéma, on nomme « effet Koulechov »[28] l'effet qui, par le montage, nous fait percevoir une image différemment en fonction de celle qui la précède. Ainsi, une même image dégagera un sentiment de tristesse si elle précédée d'une image d'en enfant qui pleure mais pourra dégager un sentiment de joie si elle est précédée d'un enfant qui rigole. En transposant cet effet à un contenu audiovisuel jouant sur la simultanéité, on peut mentionner « l'effet McGurk »[29], illusion sonore qui se produit lorsque une voix est synchronisée à différents visages, pouvant faire entendre la consonne /Fa/ alors que la voix prononce /Ba/, car cette dernière était synchronisée à un visage prononçant le phonème /Fa/. D'un point de vu émotionnel, cette différence pourrait faire entendre une émotion opposée que celle jouée par la voix en synchronisant cette dernière à un visage exprimant une émotion totalement opposée. C'est ce qui se passe pour autant dans la communication non verbale : va-t-on ressentir l'expression entendue ou l'expression vue ? La voix est un instrument émotionnel indissociable de notre corps ; la voix communique ce que le corps ressent. Cet instrument nous est donc authentique et son unicité participe à faire de notre voix une caractéristique de notre identité.

---

[27] La synchrèse est un concept défini par Michel CHION dans *La voix au Cinéma*, ed. Cahiers du Cinéma, 1982. La synchrèse est l'émanation opérée par la synchronisation de l'image et du son qui associe un son perçu à une source visuelle car ces deux éléments interviennent de manière simultanée.

[28] URL effet Koulechov : https://www.youtube.com/watch?v=Mkgwo4GOOVk

[29] McGURK H., MACDONALD J. *Hearing lips and seeing voices*, Nature, vol. 264,1976.

d. Une oralité à soi : discours et disfluences

> Parler, c'est formuler *« des discours généralement non préparés à l'avance. Or lorsque nous produisons des discours non préparés à l'avance, nous les composons au fur et à mesure, en laissant des traces de cette production ».*
>
> *(BLANCHE-BENVENISTE Claire, 1990)*

Notre oralité est influencée par la socialisation de notre discours. Selon les contextes, les différentes propriétés de la voix sont mobilisées pour séduire, informer, vendre, convaincre, rassurer, terroriser, imiter ou prendre des masques. Par exemple, un présentateur de télévision qui livre un discours informatif soit-disant objectif essayera de gommer toutes les inflexions de sa voix qui seraient la preuve de sa subjectivité orale, autrement dit, il essaie de gommer son avis. En revanche, on attendrait de la voix d'un parent à son enfant un discours sincère, qui exprime de manière claire l'émotion envers son auditeur[30]. C'est la différence principale entre le langage oral et le langage écrit : il n'y a pas de texte préexistant pour l'oralité. Il y a donc dans l'oral nécessairement production, construction et performance.[31]

La manière dont notre discours se construit à travers notre voix et la façon dont les émotions qui s'en émanent nous sont personnelles. Notre expressivité est avant tout audible lorsque notre voix est improvisée. *A contrario*, la voix qui lit, ne pouvant improviser les mots qu'elle prononce, permettra de gommer l'expressivité d'un discours oral. Ces marques

---

[30] https://encyclopediedelaparole.org/fr/node/9772 : le site l'Encyclopédie de la parole répertorie dans son onglet TIMBRES, différentes voix en fonction du type de discours.

[31] DESPRATS, Pierre, *Recherche sur l'identité vocale dans la synthèse vocale et sa relation à la disfluence*, mémoire (sous le direction de Thierry Coduys et Greg Beller), Son, ENS Louis-Lumière, 2014.

d'une parole improvisée sont appelées disfluences. En d'autres termes, c'est l'action de chercher ses mots alors que la syntaxe de la phrase est déjà en place. Pouvant prendre la forme d'hésitations, de bégaiements, de tics de langage ou de bruits de bouches et de respirations, ces disfluences portent le discours improvisé, caractéristique d'une parole humaine. Les disfluences sont la preuve d'une réflexion en direct, d'un passage de la pensée à la voix. Elles varient également en fonction du contexte social et du type de discours employé. La disfluence est donc un élément constitutif de la prosodie d'un locuteur. Pour reprendre les mots de Roland Barthes définissant la notion de « bredouillement » : « la parole est irréversible, telle est sa fatalité. Ce qui a été dit ne peut se reprendre, sauf à s'augmenter : corriger, c'est, ici, bizarrement, ajouter. »[32]. Un parallèle trivial serait de comparer la rature de l'écrit avec l'hésitation à l'oral. Nous verrons par la suite que ces bredouillements, ces recherches de lexique ou de syntaxe durant un discours oral feront partie de l'effet du naturel recherché dans les voix de synthèse. En 2018, le « Hm-hm » d'approbation et le « Eeeeeuh » d'hésitation de la voix artificielle de OKGoogle viennent juste d'être implémentés[33].

La voix est donc un instrument émotionnel, notre moyen de communiquer le plus intime et le plus fidèle à notre identité. La voix est si fidèle qu'elle peut également trahir une intentionnalité qui se voulait masquée. « La voix ne trompe point même si les paroles trompent. » écrivait André Suarès[34].

---

[32] BARTHES, Roland, *Le Bruissement de la langue*, Ed. Seuil, Paris,1984.

[33] URL (mars 2019) : https://www.youtube.com/watch?time_continue=52&v=-qCanuYrR0g : Publicité pour le nouveau Google Assistant qui réserve une table au restaurant par téléphone. On peut entendre ses hésitations et un "Hm-hm" d'approbation.

[34] SUARÈS A., *Remarques*, Paris, Gallimard, Les Cahiers De La Nrf, 2000.

e. Un modèle vocal mécanique

En cherchant à identifier tous les paramètres qui forment notre identité vocale, nous sommes en mesure d'énumérer les caractéristiques à atteindre pour créer une voix artificielle, une voix de synthèse. Dans ce paragraphe, nous allons décrire le premier modèle vocal mécanique (sans électricité) cherchant à reproduire un timbre de voix humaine, ce qui nous amènera dans une deuxième partie à réfléchir sur une reproductibilité de notre identité vocale, par clonage numérique.

Si nous cherchions à reproduire de manière mécanique notre voix, il nous faudrait un instrument à vent, couplé à un instrument à cordes. Le chercheur hongrois Wolfgang Von Kempelen a tenté de modéliser ce double instrument en créant sa *Sprech-Maschine*[35], dont il décrit le fonctionnement dans son livre « Machanismus der menschlichen Sprache nebst Beschreibung diner sprechenden Maschine » (Le mécanisme de la parole humaine avec une description de la machine parlante), en 1791. Un soufflet joue le rôle de nos poumons, tandis qu'une anche double (en roseau) rentre en vibration à la manière de nos cordes vocales. Un réseau réglable de tuyaux permet à opérateur de modifier la hauteur du son. Ce modèle vocal est considéré comme la première machine de synthèse vocale mécanique concluante.

Cette machine à parler ne pouvait pas « parler » allemand mais permettait la production de quelques phonèmes. Quelques exemples de son vocabulaire étaient, en français et latin : « *Vous êtes mon ami - je vous aime de tout mon cœur - Leopoldus Secundus - Romanorum Imperator - Semper Augustus - papa, maman, ma femme, mon mari, le roi, allons à Paris* »[36].

---

[35] mot à mot : machine parlante

[36] Nous avons le rapport d'un M. Winfisch, *Lettres de M. Charles Gottlieb de Windisch sur Le joueur d'échecs de M. De Kempelen* (Bâle, 1783), cité par Parret 2002, p. 27.

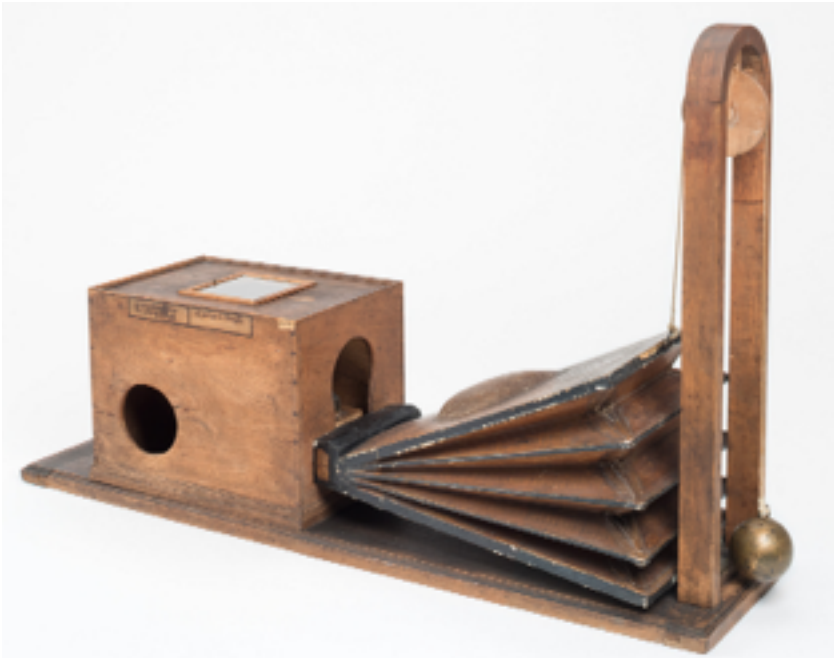


Figure 8 : modèle original de la *Sprech-Maschine* de Wolfgang Von Kempelen, au musée de Munich (un modèle en état de fonctionnement est également visible au musée)[37].

La voix de cette *Sprech-Maschine* doit encore être activée par l'homme. Elle s'incarne à la fois dans l'outil mécanique mais aussi dans les gestes de celui qui la manipule. L'outil devient un objet doté d'une dimension subjective, cette parole modulable dans son contenu et son ton, par les mouvements de l'opérateur. Cette dualité de l'objet potentiellement sensible est d'ailleurs ressentie dans l'exemple de son vocabulaire : entre la déclaration d'amour et l'éloge de la règle. Reproduire le sensible sans pour autant cacher le fonctionnement mécanique de l'outil, la démarche de Wolgang Von Kempelen était à double sens. Il semble important de noter que, parallèlement à la machine parlante, Wolgang Von Kempelen inventa le Turc mécanique (1769), une première tentative de simulation d'intelligence artificielle. Le Turc mécanique est un automate d'apparence humaine qui actionnait les

[37] URL (février 2019) pour entendre une reproduction de la machine : https://www.youtube.com/watch?time_continue=31&v=oIjkzZGe2I8

pièces d'un jeu d'échecs ; faisant mine qu'aucun homme ne pouvait être caché dans le meuble de la machine, le public croyait à l'intelligence artificielle de l'automate, même s'il y avait un « truc ». Dans les années 1780, Wolgang Von Kempelen présentait les deux inventions lors d'événements communs, la machine parlante comme introduction au Turc mécanique et dévoilait ainsi deux aspects de cette double-créature, la parole d'un part, suivie de la pensée. La simultanéité des deux inventions semble pouvoir résonner avec le développement actuel des Intelligences Artificielles (IA) dont on tente de faire entendre leur raisonnement via une synthèse vocale incluant des hésitations, simulant alors un système intelligent. La réunion de la pensée et de la voix tendent vers la reproduction de la parole humaine, improvisée.

Suivant les instructions de Wolfgang Von Kempelen, un certain Charles Wheastone répliqua la machine parlante en 1838 et cette dernière inspira un certain Alexander Graham Bell qui, poursuivant ces recherches, abouti à la création du téléphone. Malgré le filtre du combiné, il suffit de prononcer « allô, c'est moi » pour que l'auditeur reconnaisse la voix du locuteur.

I / 2. Synthétiser la voix : dépasser *the uncanny valley of speech*[38]

« L'objectif de la synthèse de la parole est de produire des sons de parole à partir d'une représentation phonétique du message »

(Calliope, 1989)[39]

Depuis cette première tentative d'imiter la parole humaine de manière mécanique, la volonté de re-créer une voix crédible et réaliste n'a cesser d'augmenter. Aujourd'hui, les voix de synthèses font partie de notre quotidien, que ce soit par la voix de Siri sur notre iPhone ou bien la nouvelle voix E-Mone de la SNCF. Le développement des appareils à commande vocale, comme la toute récente sortie (juin 2018) de l'utilitaire Alexa d'Amazon, rendu populaire par leur enceinte Echo[40]. Ces voix sont de plus en plus réalistes, que ce soit en terme de timbre et de prosodie. Dans ces appareils à commande vocale, c'est un véritable dialogue qui peut s'ouvrir, entre l'homme et la machine. Cependant, il existe une étape à franchir, ou plutôt une « vallée », celle de la vallée de l'étrange (phénomène décrit par Masahiro Mori dans son article publié en anglais au sein de la revue *Energy 7* en 1970). Ce concept explique qu'à la frontière entre un être presque humain mais toujours robotique, l'ambiguïté de sa similarité le rend étrangement inquiétant. Cette inquiétante étrangeté, traduite en anglais par « *the uncanny* », provient du terme freudien *das Unheimliche*[41] en allemand. Dans notre cas, ce concept décrit le sentiment d'ambiguité entre une voix aux caractéristiques humaines et qui pourtant garde de son artificialité synthétique.

---

[38] *the uncanny valley of speech*, mot-à-mot "la vallée de l'étrange de la parole"

[39] CALLIOPE, TUBACH J., *La parole et son traitement automatique* - Collection technique et scientifique des télécommunications (ENST), 1989, Paris: Masson.

[40] URL (février 2019) : https://www.journaldelavoix.com/06/06/2018/alexa-est-sortie-en-france/

[41] FREUD Sigmund, L'inquiétant familier (suivi de : "Le marchand de sable" de E.T.A. Hoffmann), Paris, Payot, coll. "Petite Bibliothèque Payot", 2012.).

La synthèse de la parole consiste en la lecture par une voix synthétique d'un texte numérique[42], aussi appelé en anglais « *text-to-speech* » (mot-à-mot : du texte à la parole). En commençant par évoquer le paradoxe entre l'écrit et l'oralité dans la fabrication de la voix de synthèse, nous décrierons le fonctionnement général de la synthèse par formants, puis concaténative et nous aboutirons à l'utilisation des réseaux de neurones pour construire ces voix expressives. Ensemble, nous traverserons la vallée de l'étrange de la parole, *the uncanny valley of speech,* c'est-à-dire le moment où la voix de synthèse se veut indifférenciable d'une voix humaine organique. Même si, d'un point de vue d'enveloppe, une voix de synthèse peut être indifférenciable d'une voix organique, cette dernière ne pourra être entièrement perçue comme humaine que si son contenu est, quant à lui, aussi proche d'un discours humain. Ce qui nous amène à se poser la question du paradoxe entre une synthèse passant par l'écrit pour être oralisée : comment donner l'impression d'une pensée qui précèderait le langage ? Pour dépasser *the uncanny valley of speech,* il faudrait ajouter la pensée qui précède le langage.

a. Paradoxe du *text-to-speech* (TTS) : entre écrit et oralité

Toutes ces voix de synthèses partent d'un texte écrit pour l'oraliser. Les voix construites pour la domotique ont chacune leur personnalité et se distinguent dans leur manière de répondre, dans leur hésitations et ainsi, leur réalisme. Dans un discours oral, la parole est improvisée. Cette improvisation s'entend par nos bégaiements, nos respirations, nos pauses entre les mots, qui sont la preuve d'un discours construit en temps réel, brisant la régularité d'un énoncé. La difficulté de ces voix artificielles qui mettent à l'oral un texte lu et non improvisé est d'avoir ces défauts de parole, ces détails qui nous rendent humains, implémentés dans leur

---

[42] D'ALESSANDRO, Christophe, TZOUKERMANN, Evelyne (sous la direction de), *Synthèse de la parole à partir du texte, numéro de Traitement Automatique des Langues (TAL)*, Hermès, Vol. 42, No 1, 2001.

texte écrit. Comment écrire le doute ? Comme noter les disfluences et tous ces sons qui font sens dans le bruissement des mots ?

Nous avons vu précédemment que la disfluence faisait partie intégrante de notre identité vocale. La synthèse vocale générant une voix à partir d'un fichier texte, cette oralité doit passer par une transcription écrite à implémenter dans le texte qui sera oralisé par la voix artificielle. L'histoire de la littérature a su nous montrer qu'il existait une fine barrière entre une erreur et un effet de style. Par exemple, *Le Journal d'un vieux dégueulasse* de Charles BUKOWSKI (1969), vise à transcrire un témoignage sensible, où chaque phrase commence par une minuscule et où la ponctuation est désordonnée. Dans *Le grain de la voix*, Roland BARTHES explique dans un de ses entretiens[43] que « *ce qui se perd dans la transcription, c'est tout simplement le corps - du moins ce corps extérieur (contingent) qui, en situation de dialogue, lance vers un autre corps, tout aussi fragile (ou affolé) que lui, des messages intellectuellement vides, dont la seule fonction est en quelque sorte d'*accrocher *l'autre (voire au sens prostitutif du terme) et de la maintenir dans son état de partenaire. »* Pour établir un dialogue, il faudrait alors maintenir son interlocuteur en communication, et pour cela la voix et ses aspérités qui font entendre le corps du locuteur sont des outils indispensables. Comment faire entendre le corps d'une voix artificielle qui n'est par définition qu'une voix sans corps ? Son corps serait la machine dont s'émane le son ? Les artefacts de fabrication de cette voix pourraient-ils être comparés aux bruits de bouche d'un locuteur acoustique ? La synthèse vocale *text-to-speech* fait le chemin inverse et part donc de l'écrit pour arriver à une voix spontanée orale.

[43] BARTHES, Roland, *Le grain de la voix*, entretien *De la parole à l'écriture*, La Quainzaine littéraire 1er-15 mars 1974, Essais, 1981.

Il existe cinq catégories de disfluences : les pauses vides, les pauses remplies, les amorces, les répétitions et les autocorrections[44]. Comprendre les différentes techniques pour noter ces disfluences à l'écrit ferait l'objet d'un autre mémoire mais l'on peut ici déjà comprendre que l'analyse du discours oral fait partie intégrante de la recherche de spontanéité dans le discours des voix artificielles. De plus en plus utilisées pour incarner des IA (comme Siri, d'Apple), leur discours doit relever d'un mécanisme psycholinguistique simulé, pour rendre compte du processus de construction et préciser le cheminement menant à l'énonciation. En reprenant l'exemple du mémoire de fin d'études de Pierre DESPRATS (2014), dans la phrase :

« Je suis allé: dans: euh à la bibliothèque »

L'accès au mot « bibliothèque » a été fait après la construction de la syntaxe contenant « à ». Cet exemple est écrit selon la notation de Rémi BOVE, où l'allongement syllabique est noté à l'aide d'un « : » accolé à la syllabe allongée. À la différence d'une pause silencieuse, il y a ici restructuration de la phrase après la disfluence, telle une correction de l'erreur syntaxique. La pause remplie présente l'avantage, dans le dialogue, de témoigner d'un discours en cours de construction et donc que le tour de parole n'est pas encore terminé ; elle permet ainsi d'occuper un territoire de parole, le temps de finir son cheminement syntaxique. On pourra noter qu'en fonction de la langue parlée, ces pauses ne placent pas aux mêmes endroits dans la syntaxe. La manière de « couper » une phrase est aussi différente ; par exemple, en allemand, le verbe étant le plus souvent à la fin de la phrase, la suspension au discours de quelqu'un est bien plus importante qu'en français où un interlocuteur peut facilement finir la construction syntaxique de la

[44] BOVE, Rémi, *Analyse syntaxique automatique de l'oral : étude des disfluences.* Aix: Université d'Aix-Marseille, (2008).

personne avec qui il dialogue. En allemand, on doit attendre la fin de la phrase afin de savoir exactement ce que le locuteur voulait réellement exprimer.

Ces disfluences, qui participent à la spontanéité du discours oral, sont accompagnées de sons non syntaxiques, qui participent à la fluidité du dialogue et rendent la répartition de parole plus naturelle. Ainsi, les mots d'approbation tels que « oui », « je vois », ou simple « hm-hm » émis par l'un des locuteurs pendant le discours de l'autre, encourage celui qui énonce son discours dans la progression de son énonciation. Ces marqueurs de compréhension sont la preuve que la communication est établie sur le même canal et que le locuteur est bien compris par son auditeur.

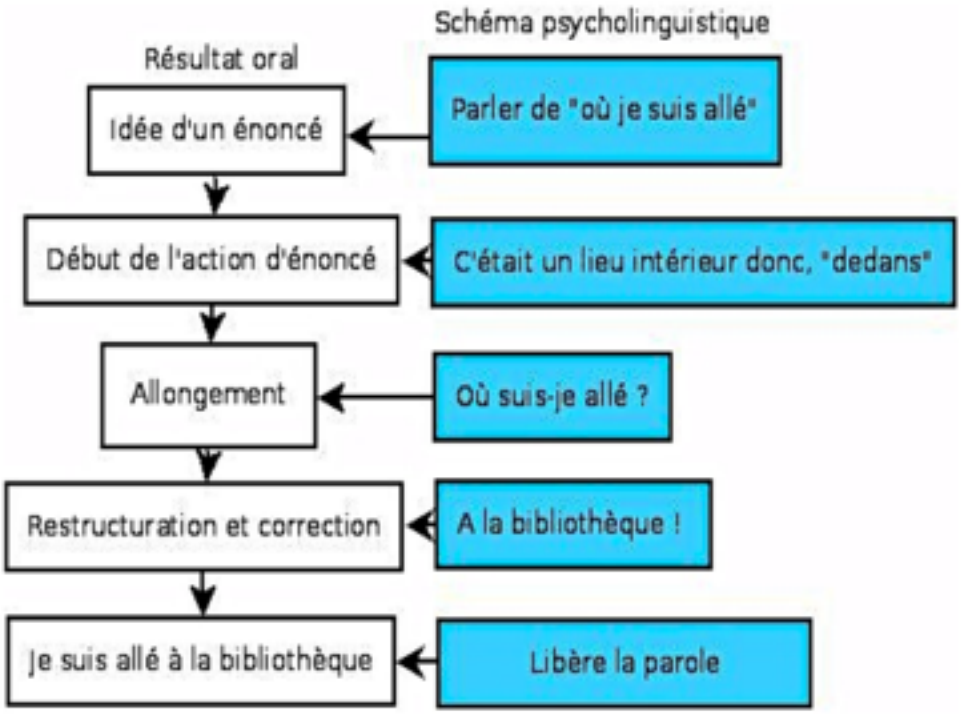


Figure 9 : Schéma fonctionnel d'une pause remplie

Afin de comprendre davantage ce processus de *text-to-speech* et de prendre conscience de mes propres disfluences, j'ai voulu me mettre à l'épreuve de la transcription de mon discours oral. Pour cela, je me suis enregistré pendant quelques heures (de 8h à 11h30, jour de classe) et ai mis sur papier tout ce que j'avais dit durant ces quelques heures. Ce texte, un extrait de ce processus de *speech-to-text* est en *annexe 1* de ce mémoire[45], on peut y lire la recherche de transcription, incluant un code graphique pour suivre la parole prononcée (par exemple, sauter une ligne = 1 min de parole). Ces preuves d'une pensée qui précèdent la parole font donc partie de l'identité possible d'une voix artificielle. Mais en quoi son identité vocale (qui passe par le contenu) passe-t-elle aussi par le choix de son timbre ? Il faut alors définir le but de l'utilisation de cette voix de synthèse : un agent conversationnel (comme l'intelligence artificielle d'*Apple*, *Siri*) n'aura pas la même fonctionnalité qu'une voix de *GPS* qui doit seulement donner des instructions, sans hésiter. Si le contenu d'une voix artificielle peut être crédible, à la fois dans une parole "lue" et dans une parole écrite pour paraître « improvisée », la synthèse vocale cherche alors à avoir un timbre vocal aussi crédible que son contenu.

b. Synthèse par formants

En 1939, les laboratoires Bells (New-York City) développent le « Voder » (*Voice Operating Demonstrator*), le premier appareil électronique synthétisant des phonèmes[46]. À l'aube de la seconde guerre mondiale, cet appareil inventé par Homer W. Dudley est d'abord conçu dans le cadre de la transmission de messages vocaux sécurisée. Un opérateur joue différents phonèmes, en choisissant à l'aide de son poignet, l'activation d'un oscillateur pour simuler des voyelles ou bien

---

[45] *annexe 1* : Extrait de *Quelques heures sur le bout de ma langue,* publication pour le Master Transdisciplinaire de l'Université d'Arts de Zurich (ZHdK), 2019.

[46] URL (février 2019) : https://www.youtube.com/watch?v=5hyI_dM5cGo documentaire de présentation du "Voder"

d'un générateur de bruit blanc qui sert à simuler les consonnes. Une fois l'un de ces deux sons choisi, l'opérateur peut commander la hauteur du son (le pitch) avec l'aide d'une pédale ou bien activer différentes bandes fréquentielles d'un filtre à l'aide d'un clavier. Cette coordination du poignet, des mains et du pied demandait à tout opérateur d'être formé, afin de créer des formants les plus réalistes possibles.

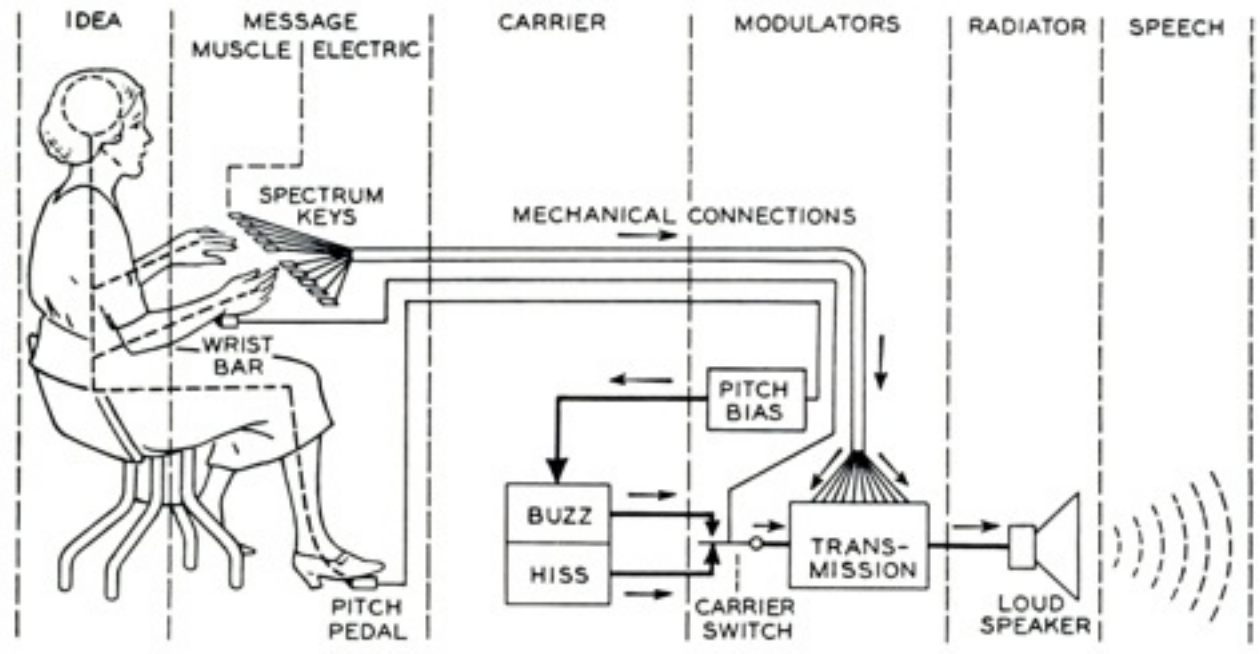


Figure 10 : Fonctionnement du Voder[47]

Suite à l'invention du Voder, une grande série de recherches voient le jour, mettant en évidence l'importance des transitions dans l'enchaînement des phonèmes. En 1961, l'ordinateur IBM 7094 fut le premier appareil électronique à chanter une chanson, appelée « Daisy »[48]. Programmés par John Kelly and Carol Lockbaum, les formants font entendre les paroles : « Daisy, Daisy, give me your answer do I'm half crazy all for the love of you It won't be a stylish marriage I can't afford a

[47] DUDLEY Homer W., *Bell System Technical Journal* 1940, p. 509, Fig.8 Schematic circuit of the *Voder*.

[48] URL (février 2019) : exemple de synthèse par formants à écouter ici : https://www.youtube.com/watch?v=41U78QP8nBk

carriage. But you'll look sweet upon the seat Of a bicycle built for two. » Cette nouvelle a inspiré l'auteur Arthur C. Clarke et fut donc ré-interprétée par HAL 9000 dans *2001, L'odyssée de l'espace* (KUBRICK Stanley, 1968).

Figure 11 : DECTalk DCT01[49] avec un chat pour donner un ordre de taille

En 1984, le chercheur Dennis KLATT propose le DECTalk, le premier synthétiser vocal basé sur un algorithme source-filtre, connu sous le nom de KlattTalk ou MITalk, permettant de générer de l'audio à partir de texte. Dennis KLATT fit lui même les enregistrements d'une des voix de synthèses proposées par l'appareil, appelée *Perfect Paul*. C'est cette voix que Stephen Hawking utilisa en premier et dont il chercha à garder l'identité vocale pendant sa maladie, malgré les nouvelles voix qui lui étaient offertes. Le DECTalk proposait également un choix de sept autres voix telles que *Beautiful Betty* (voix de femme) et de *Kit the Kid* (voix d'enfant)[50].

---

[49] Photo de Emgaol - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=10796356

[50] URL (mars 2019) : écouter les différentes voix de DECTalk : https://www.youtube.com/watch?v=8pewe2gPDk4

La synthèse par formant se fait à l'origine à l'aide de plusieurs F.O.F. (Fonction d'Onde Formantique) qui sont des sinus synchrones à la partie voisée (contenant la voyelle) et asynchrone pour la partie fricative (simulant l'effet de bruit des cavités supra-glottiques), modulés par une enveloppe qui permet la formation du formant. Cette technique de synthèse se caractérise donc par une imitation de la voix par des procédés de filtrages. Ne dépendant pas d'un corpus vocal enregistré au préalable, la voix synthétisée est facilement modulable dans son expressivité (par ajustement de filtre). De plus, son intelligibilité est remarquable, même à une grande vitesse d'élocution.

Cependant, son timbre reste assez robotique, ne comportant aucun « défaut » dû à une articulation humaine (bruit de bouche, grain de voix singulier ou autre aspect prenant part à une identité vocale particulière). Par la suite, afin d'obtenir ce timbre plus proche de la voix humaine, la recherche s'est tournée vers l'utilisation d'un corpus vocal dans lequel irait piocher un algorithme pour former différentes syllabes, mots, phrases. C'est ce que nous appelons « synthèse à concaténation d'unités ».

c. Synthèse concaténative

La synthèse par concaténation d'unités est une synthèse dite "directe", car elle consiste en la lecture d'une série d'échantillons sonores (appelés unités), enregistrés dans un bibliothèque d'échantillons (corpus vocal). Plus les échantillons sont nombreux et longs, plus la synthèse sera de bonne qualité. À la différence de la synthèse granulaire qui met bout à bout différents grains sonores, la synthèse concaténative prend en compte les données acoustiques (hauteur, énergie, spectre, aussi appelés descripteurs sonores) des unités, afin d'identifier celles qui correspondent le mieux au critère spécifié pour former la phrase. En effet, la dernière étape de ce modèle de synthèse sonore est de choisir et combiner les

segments du corpus qui conviennent le mieux pour la phrase que nous souhaitons synthétiser.

Comme nous l'avons vu précédemment avec la SNCF, la voix de Simone Hérault a été aussi synthétisée par concaténation d'unités : d'abord en choisissant des mots entiers comme échantillon, puis par la suite, des phonèmes. Par exemple, pour annoncer la venue d'un train, le schéma de la phrase peut être :

> Le train */ numéro du train /* en provenance de */ nom de la provenance /* et à destination de */ nom de la destination /* partira voie */ numéro de la voie / . /*

Les éléments variables, écrits en italique (numéro de train, noms de provenance et destination, numéro de la voie), sont issus d'une bibliothèque d'échantillons (corpus vocal).

Pour créer un synthétiseur à concaténation d'unité, il faut un texte défini et l'enregistrement de la voix lisant ce texte, formant alors un corpus vocal qui servira de bibliothèque d'échantillon pour la synthèse. Il faut ensuite concaténer à différentes échelles (phrase, mot, demi-syllabe, phonème), puis aligner chaque élément avec le texte correspondant ainsi que les informations acoustiques. Une fois ces différents segments d'audio réalisés, plusieurs étapes se suivent pour passer du texte écrit à de l'audio généré.

Premièrement, une étape de « grapheme-to-phoneme » convertit une forme écrite (lettres françaises par exemple) en une forme phonétique (utilisant un alphabet phonétique type XSAMPA[51]). Ensuite,

[51] *XSAMPA* : eXtended Speech Assessment Methods Phonetic Alphabet (alphabet phonétique étendu)

un modèle de segmentation localise les différents phonèmes prononcés par la voix. Lors de la saisie d'un texte, il sera ensuite, de la même manière, décomposé à ces différentes échelles afin de créer un signal continu respectant le sens sémantique du message d'entrée. La concaténation d'unités permet de garder l'authenticité du timbre de la voix enregistrée, elle prend les échantillons contenant alors les bruits de bouche et les respirations enregistrées. Cependant, sa modulation expressive est assez rigide, car dépendante du corpus vocal qui lui est fourni.

La synthèse concaténative privilégie donc l'identité spectrale unique du locuteur. Elle pose donc la question du choix de cette identité vocale pour les voix qui incarnent les intelligences artificielles, qui servent d'agent conversationnels ou de service domotique. Nous pouvons remarquer que la voix par défaut de Siri en français est systématiquement une voix de femme. Les tâches de ces « assistants » renforcent notre perception des actions genrées, ayant une voix féminine qui ne peut jamais dire « non ». Les recherches autour d'une voix non genrée permettent enfin de choisir une option qui n'influence pas notre perception de ces tâches d'assistanat, faites par une femme. De plus, si vous interrogez Siri sur son genre, celle-ci vous répondra qu'elle n'en a pas. Pourtant, avec sa voix de femme, elle accorde tous ces mots au genre féminin… Définir une voix sans genre pour une intelligence artificielle permet de lutter contre l'association d'un genre et d'un service, ce qui modifie également la manière dont nous percevons les voix organiques. Nous sommes en mesure de nous demander si nous arrêtions d'associer « voix féminine » et « service », nous accepterions davantage d'avoir une femme au pouvoir ou d'entendre une femme être indépendante. À l'heure où l'usage de ces voix artificielles ne cesse d'infiltrer notre paysage sonore, où notre perception de ces voix influence notre rapport aux machines, le plus souvent en leur donnant un genre binaire « masculin » ou « féminin », un groupe de recherche a commencé à développer sur voix de synthèse sans genre, une voix neutre. Cette voix

s'appelle « Q »[52], elle a été forgée à partir d'enregistrements anglais de voix d'hommes, de femmes et de personnes trans, afin de n'être perçue ni comme une voix d'homme (autour de 145 Hz), ni comme une voix de femme (autour de 175 Hz), sa hauteur moyenne étant à 153 Hz.

Même si le genre de ces voix de synthèse influence notre perception de l'identité des intelligences artificielles, ces voix restent un outil de communication. Elles ont vocation à rendre le dialogue homme-machine de plus en plus naturel, sans qu'une distinction d'adresse ait lieu lorsque qu'un utilisateur s'adresse à son outil de domotique ou bien à son ami. La ressemblance de l'humain passe donc par la considération (service ou compagnon) que nous avons de ces outils. L'implémentation d'une voix connue dans ces outils rend-elle leur considération plus humaine ?

d. Synthèse par réseau de neurones

À la fin des années 1990, la synthèse dite statistique ou paramétrique est apparue, avec les premières tentatives de modéliser les paramètres d'une voix comme l'intonation, le rythme, et le timbre, par des modèles statistiques génératifs comme les chaînes de Markov cachées. Depuis les années 2010, les synthétiseurs basés sur l'apprentissage profond utilisent les réseaux de neurones profonds (*DNN* pour *Deep Neural Network*), qui sont entraînés sur des données de la parole enregistrée. Ils fonctionnent sur le même principe que la synthèse concaténative mais s'adaptent par apprentissage au corpus de voix pour que la synthèse soit encore plus réaliste dans sa prosodie ; c'est-à-dire que chaque mot synthétisé dans une phrase a pris en compte les mots qui l'entourent, afin de rendre la mélodie et l'articulation de la phrase plus réaliste, et moins systématique qu'une mélodie de phrase ne variant pas selon différentes ponctuation (question, phrase en suspension etc). La

[52] *Q*, site officiel (URL avril 2019) : https://www.genderlessvoice.com

synthèse par réseau de neurones (*neural text-to-speech* ou *NTTS*) apprend ces variations de prosodie (et donc de lien entre les phonèmes) après l'analyse de différents corpus, qui mettent en avant la singularité de la voix (l'idiosyncrasie du locuteur, c'est-à-dire sa manière propre d'articuler), en utilisant un modèle statistique comprenant les chaînes de Markov cachés (HMM, *Hidden Markov Marks*). Ces modèles permettent l'analyse de l'évolution spectrale des phonèmes dans le temps et dans un contexte linguistique précis. Ces voix de synthèse réalisées par *neural text-to-speech* permettent donc de jouer sur la prosodie de manière précise et réaliste afin d'accentuer certains mots, de modifier l'intonation si la phrase est une question ou une phrase en suspension, de synthétiser des phrases où la prononciation phonétique doit comprendre le rôle de certains mots dans la phrase. Par exemple, dans la phrase « les présidents président », la synthèse *text-to-speech* doit définir le phonème /e/ pour la fin du mot « président » afin de différencier le nom commun « les présidents » du verbe conjugué « président ».

Dans la suite du mémoire, nous verrons que c'est cette technique de synthèse par réseau de neurones que nous utiliserons pour réaliser le clone vocal (utilisation de IrcamTTS[53]), afin de pouvoir le comparer à sa voix organique d'origine.

C'est aujourd'hui la technique de synthèse vocale qui s'approche le plus du timbre et de l'expressivité de la voix humaine. Les principaux algorithmes de synthèse vocale qui utilisent des réseaux de neurones sont : WaveNet de DeepMind, Tacotron de Google et Deep Voice (qui utilise la technologie WaveNet) de Baidu. Afin d'évaluer et de comparer les techniques de synthèse, les résultats sont communiqués principalement suivant le critère d'appréciation globale subjective nommé test d'opinion moyenne (Mean Opinion Score ou MOS en anglais). Ainsi, les résultats de *Tacotron 2* ont un MOS de 4.526 ± 0.066 (sur 5), tandis que la synthèse concaténative a un MOS de 4.166 ± 0.091 (sur 5) suivant

[53] *ircamTTS* est le modèle de synthèse *text-to-speech* de l'IRCAM

le même corpus[54]. Chaque modèle se différencie dans son choix d'apprentissage par réseau de neurones.

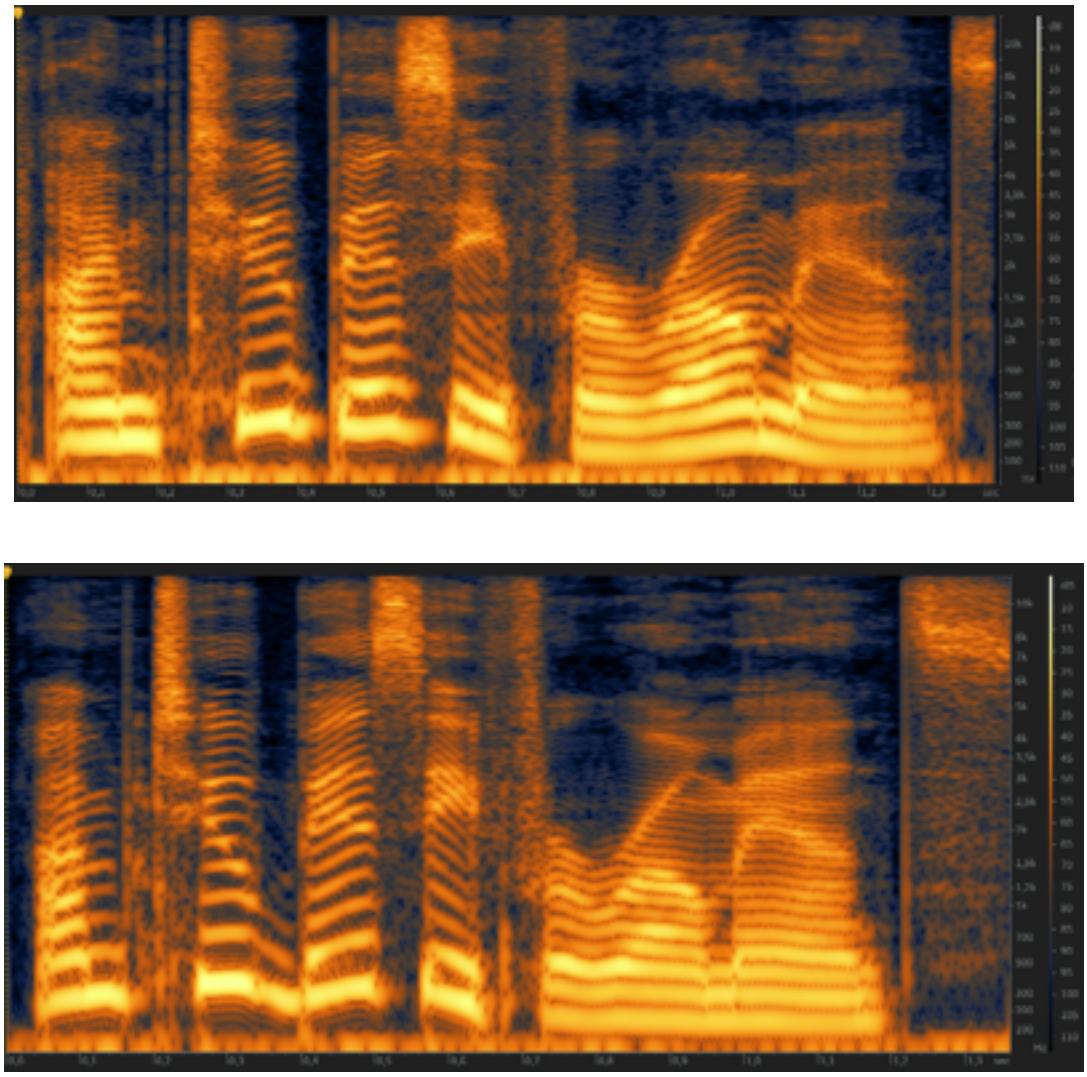


Figure 12 : spectrogrammes des phrases « I'm too busy for romance » généré par l'algorithme *Tecotron 2* (au dessus) et le même sample audio enregistré par la comédienne (en dessous).

[54] SHEN Jonathan, et al., *Natural TTS Synthesis by conditioning Wavenet on mel spectrogram predictions*, Google, Inc., University of California, Berkeley, 2017.

Ci-dessus, deux spectrogrammes[55] de la même phrase : « I'm too busy for romance », le premier généré par l'algorithme *Tecotron 2* et le second étant l'original (même quantification pour les deux samples : 24 kHz - 16 bits). Nous pouvons observer que les deux figures comportent les mêmes informations fréquentielles (même timbre entre les deux voix), qu'une différence peut être visible sur les transitoires (début de l'articulation de certains phonèmes), et que c'est principalement la mélodie de la voix sur "romance" qui reste plus stable sur la voix originale que sur la voix synthétisée (qui comporte une variation mélodique). Les différents échantillons peuvent être comparés à l'écoute via ce lien : https://google.github.io/tacotron/publications/tacotron2/index.html.

Afin de mettre à l'épreuve la crédibilité de leur voix de synthèse, nous pouvons faire passer à ces voix artificielles un test de comparaison, semblable à un test de Turing sans sa dimension d'intéractivité. Décrit en 1950 par Alan Turing, ce test consiste à placer un interlocuteur humain en confrontation verbale, à l'aveugle, avec un ordinateur et un autre humain. Le test est réussi si la personne n'est pas capable de dire lequel de ses interlocuteurs est l'ordinateur. Par analogie à ce test, les chercheurs en synthèse *neural text-to-speech (NTTS)* peuvent évaluer la crédibilité de leurs voix de synthèses en mettant à disposition des échantillons pour comparer l'audio originale du comédien enregistré, avec sa voix de synthèse générée par apprentissage. Cette comparaison permet à l'auditeur de faire son propre avis sur le réalisme des voix générées, qui, en terme de qualité, sont (du moins pour mon propre avis), non distinguables des voix d'origine. Comme nous l'avons vu avec la Figure 12, la différenciation entre la voix organique et son clone numérique est rendue difficile par le réalisme de la voix de synthèse. À l'écoute, les différences se jouent dans la prosodie et en particulier dans l'expressivité

[55] Un spectrogramme est une représentation du son sur un graphique temps (axe des abscisses) et fréquences (axe des ordonnées), il permet une visualisation instantanée d'un son. Plus l'amplitude accordée à une fréquence est grande, plus cette dernière est colorée intensément.

de la voix. La voix enregistrée semble bien plus souriante que la voix de synthèse et c'est ce qui pour moi, la rend plus réaliste. Je pense que si je n'avais pas de comparaison à faire, la voix de synthèse pourrait être une véritable voix enregistrée. Cette similitude de timbre et de prosodie montre bien que la voix de synthèse est un clone de la voix de la comédienne et que ces deux portent alors la même identité vocale. Néanmoins, il semble important de noter que le support technique nécessaire à l'incarnation des voix de synthèse ne pourra jamais être aussi fidèle qu'un corps incarnant une voix organique. J'entends par là que même si un enregistrement peut être dans certaines conditions, utilisé comme preuve, il restera toujours un doute lié au support de la voix, ne permettant par une fidélité spectrale aussi fiable qu'une voix prononcée de manière « vive ». Aussi, le support de diffusion de la voix peut servir de filtre qui masque les défauts de synthèse, permettant à une voix artificielle d'avoir un timbre crédible car couvert par un filtre de téléphone, ou un haut-parleur de télévision.

Aujourd'hui, le réalisme de ces voix de synthèse rend la différence entre voix artificielle et voix originale difficile à cerner pour l'auditeur. La communication avec ces voix artificielles *via* les commandes vocales se font de plus en plus naturelles, mais les possibles détournements de cette technologie sont aussi importants. S'il est possible de cloner l'identité vocale d'un locuteur, comment identifier cet échantillon audio comme une « fausse » voix ? Comment protéger notre identité vocale ?

D'un point de vue audio-visuel, nous pourrions comparer cette problématique à la vidéo publiée en avril 2018 par Jordan Peele afin d'alerter sur les possibles manipulations des « fake news »[56]. Sur cette vidéo (*Figure 13*), nous pouvons voir l'ancien Président Barack Obama dire des phrases telles que « President Trump is a total and complete

[56] URL (mars 2019) : https://www.youtube.com/watch?time_continue=26&v=cQ54GDm1eL0 "You won't believe what Obama says in this vidéo! ;)"

deep shit » (le président Trump est une grosse merde). Cette manipulation d'identité n'est ici qu'un trucage d'image et pourtant, grâce au talent d'imitateur de Jordan Peele, à la première écoute, un auditeur lambda peut croire que c'est bien la voix de Barack Obama, car synchronisée à son visage.

Figure 13 : capture d'écran de la vidéo de Jordan Peele pour avertir sur le devenir des « fake news », *BuzzFeedVideo*, avril 2018

La synchronisation de la voix et de la bouche amène encore plus de crédibilité à la voix (et à son contenu). Cette simultanéité de la bouche et de la voix cherche, symboliquement, à faire appartenir une voix au départ acousmatique[57] (la voix de Jordan Peele existe sans son corps) à un nouveau corps (celui de M. Obama). La synchronisation peut provenir de deux techniques différentes : une première est de re-jouer les expressions faciales « en directes » d'une vidéo, la seconde est de créer un nouveau fichier image où la modélisation du visage permet de faire bouger les lèvres en fonction d'un fichier audio donné. Nous n'aborderons pas ici les détails de fabrication de ces vidéos, dont les différentes techniques de modélisation font appel à du calcul d'image et non de son mais je pense qu'il est intéressant de noter qu'il y a une transformation de « masque » entre deux identités, comme nous le montre la vidéo de démonstration

[57] "acousmatique" comme le définit Michel CHION dans *La voix au Cinéma*, est une voix dont le corps, l'origine du son, n'est pas visible.

(*Figure 14*) de *Face 2 Face.* Cette usurpation d'identité, ou même simplement d'expressivité va au delà d'une possible copie vocale par voix de synthèse, car elle utilise la force de la synchronisation entre image et son pour « faire croire ».

Dans le contexte audio-visuel, cette synchronicité apporte une certaine crédibilité au contenu et peut-être qu'une diffusion "en direct" aurait pu garantir de l'authenticité des propos du président. Plus les algorithmes fonctionnant avec des réseaux de neurones pourront progresser et créer des images et des voix toutes aussi réalistes que le contexte audio-visuel que nous diffusons, moins nous pourrons différencier le vrai, du faux. Plusieurs solutions existent alors pour parer à ce défaut de l'œil et de l'oreille humaine qui se feront dépasser par le réalisme grandissant de cette technologie. C'est ce que nous allons aborder dans la seconde partie de ce mémoire, consacrée au clonage de voix, au détournement possible de cette technique et aux solutions que nous pourrions envisager pour "démasquer" le faux, du vrai.

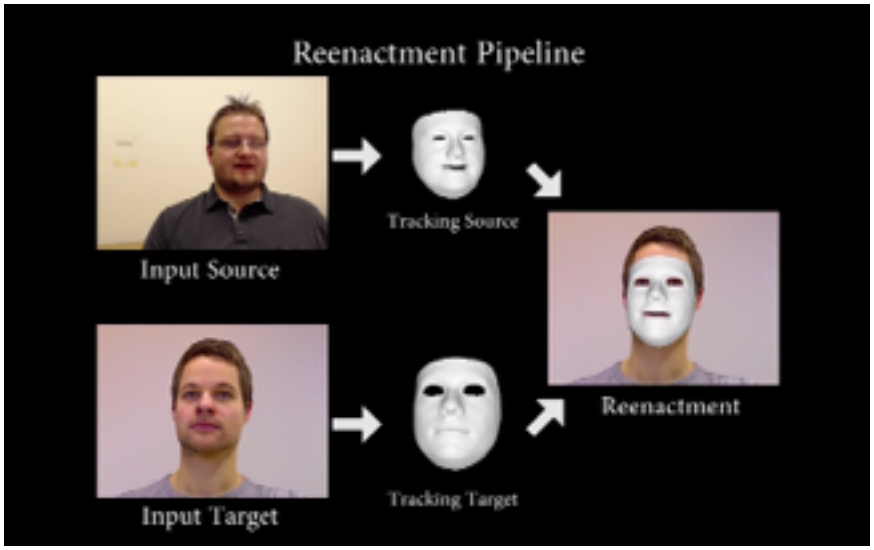


Figure 14 : instantané de la vidéo de présentation du projet *Face2Face: Real-time Face Capture and Reenactment of RGB Videos* (2016).[58]

[58] URL (février 2019) : projet *Face2Face :* http://niessnerlab.org/projects/thies2016face.html

PARTIE 2
La synthèse vocale détournée : pouvons-nous encore croire en la voix ?

Il semble important de redire que le clonage vocal permet de dissocier la voix de son corps d'origine, posséder cette voix revient à avoir un instantané d'un état de la voix d'un individu, sans être maître de toutes ses nuances et de ses changements avec l'âge etc. Les détournements possibles de cette voix de synthèse sont forcément en lien avec un support technologique (c'est-à-dire *via* un téléphone, une vidéo, un simple micro branché à des haut-parleurs par exemple). Cette mise à distance par la machine va dans le sens de la qualité sonore des clones vocaux, dont les fichiers générés ne pourront, par nature, être aussi précis qu'une voix organique. J'entends par là que la voix artificielle a besoin d'un support de reproduction sonore pour s'incarner et que par définition, le possible détournement d'un clone vocal se fera dans ce contexte numérique, de voix portée par un support électro-acoustique.

II / 1. La voix clonée

La synthèse vocale a su se développer pour arriver à copier l'identité vocale d'un comédien afin de rester fidèle à ses qualités d'articulation, de timbre et d'expressivité. Nous avons vu précédemment qu'aujourd'hui, la distinction entre la voix enregistrée et la voix générée pouvait être difficile, rendant le paysage de ces voix artificielles de plus en plus réaliste, facilitant la communication entre l'humain et la machine. Copier l'identité vocale d'une voix source revient à faire un clone de cette voix et aujourd'hui, la recherche et les compagnies qui en découlent tentent de démocratiser cette pratique en permettant à n'importe quel particulier d'effectuer son propre clone vocal. Dans cette partie, nous allons voir quelles pourraient être les applications possibles des clones vocaux et en quoi cette technologie pose également des questions légales de protections d'identité vocale et de manipulations.

a. Cloner la voix : applications directes

Cloner une voix permet en quelque sorte de la détacher de sa source, de son corps d'origine. La recherche en synthèse vocale a une application directe dans le domaine médical, notamment pour les patients atteints de SLA (Sclérose latérale amyotrophique, plus connue sous le terme de « maladie de Charcot »), qui les prive de l'usage de la parole. Pour s'exprimer, les patients, comme le célèbre chercheur Stephen Hawking, doivent faire recourt à une synthèse *text-to-speech* : c'est-à-dire écrire à travers une machine qui « prononce » les mots à leur place. Stephen Hawking utilisait une identité vocale dépendante de l'algorithme de synthèse employé[59]. Mais si cette même voix de synthèse est aussi utilisée par une petite fille de huit ans, atteinte de la même maladie, alors ces individus ne peuvent donc pas se démarquer par leur timbre unique, leur identité vocale étant complètement confondue.

La chercheuse Dr. Rupal PATEL[60] a contribué dans cette recherche à la singularisation de la voix de synthèse à travers le projet *VocalID*[61], qui permet de créer une transformation entre le son de la voix d'un patient et le spectre complet de celle d'un donneur, qui doit enregistrer entre 1000 et 3000 phrases afin que sa voix soit utilisable pour celle d'un patient. Les voix de synthèses générées sont ainsi en mélange entre le timbre du patient (qui correspondrait à son amplification par le larynx) et l'articulation du donneur.

---

[59] URL (mars 2019) : https://www.youtube.com/watch?v=6DHTh0nTo0w : dans cette vidéo, on peut entendre la voix de synthèse qu'utilise Stephen Hawking.

[60] PATEL Rupal, TEDTalk, *Synthetic Voices, as unique as fingerprints,* 2013, avec des exemples de voix de synthèses créées à écouter : https://www.ted.com/talks/rupal_patel_synthetic_voices_as_unique_as_fingerprints#t-161245

[61] URL (mars 2019) : https://vocalid.ai/about-us/

L'association *ALS Association* a également développé le projet *Revoice*[62], utilisant la création de synthèse vocale proposée par l'algorithme de neural *text-to-speech* de la société *Lyrebird AI*, pour créer les clones vocaux de patients n'étant plus en capacité de parler ; ils collectent des enregistrements et des vidéos des patients où leur voix peut être extraite afin d'établir un clone vocal par individu, fidèle à au timbre d'origine. Cette application médicale du clone vocal permet aux patients de garder leur identité vocale et facilite leur communication avec leur entourage, même via l'intermédiaire de l'outil *text-to-speech*.

Figure 15 : Stephen Hawking, atteint de *SLA*, s'exprimant avec un synthétiseur *text-to-speech* avec la voix « *Perfect Paul* »

Notre voix, comme celle de nos proches, tient le rôle principal dans notre environnement sonore quotidien, et notre oreille entretient un rapport affectif avec elles, bien plus qu'avec la voix de synthèse d'un comédien inconnu. Cloner l'identité vocale permet d'ancrer sa voix dans

[62] URL (mars 2019) : https://www.projectrevoice.org/ ; écouter un exemple de voix re-créée via ce lien : https://www.youtube.com/watch?time_continue=29&v=-pXSRjo7Czw

un autre corps que le sien, par exemple dans un objet technologique. Une application directe du clonage vocal serait de pouvoir implémenter la voix de nos proches dans le paysage des voix de synthèses qui nous entourent. La voix du GPS pourrait être celui d'un parent, le livre audio lu avant de dormir, pourrait être interprété par la voix de notre enfant. Cette utilisation revient à augmenter le rapport de familiarité entre un objet technologique et son utilisateur.

Le clonage vocal, qui dissocie la voix de son corps, permettrait donc de conserver cette part d'identité après un décès et donc, en quelque sorte, de faire parler les morts. Aujourd'hui, la SNCF a fait appel à l'entreprise *Voxygen* pour créer l'avatar *E-Mone*, le clone vocal de Simone Hérault, afin que cette voix reste l'identité sonore de la compagnie dans les prochaines centaines d'années à venir. Cet avatar est aussi incarné dans le « *chatbot* » (le robot de discussion) qui permet de répondre en ligne aux questions des utilisateurs. En 2013, la série anglaise *Black Mirror* consacre d'ailleurs le premier épisode de la saison 2 « *Be right back* » (*Bientôt de retour*) sur un clone d'un amant décédé, mettant en scène l'apprentissage vocal de l'intelligence artificielle qui clone le personnage.

Figure 16 : *E-Mone*, avatar de Simone Hérault pour la SNCF
(capture d'écran *BFM TV*) - on notera que l'avatar est
plus jeune que la comédienne -

Réaliser un clone vocal d'un acteur décédé, c'est ce qu'a réalisé le département « Analyse et Synthèse de la Voix » de l'IRCAM, en faisant jouer Louis de Funès, décédé en 1983, dans le film d'animation *Pourquoi j'ai pas mangé mon père* (DEBOUZE Jamel, 2015). Ré-entendre la voix d'un individu décédé pourrait donc donner l'impression de lui rendre la vie. La voix redonne la dimension qu'un corps physique est présent, elle redonne pouvoir au tangible. C'est ainsi que l'artiste français Philippe PARRENO a réussi à ramener à un semblant de vie l'actrice Marilyn Monroe dans son installation *Marilyn* (2012). À partir d'enregistrements de l'actrice, la voix d'une comédienne ayant un timbre similaire a été modifiée pour employer les mêmes intonations et ainsi correspondre à l'empreinte vocale d'origine. L'installation numérique *Marilyn* employait également une intelligence artificielle qui pouvait reproduire l'écriture manuscrite de l'actrice. La sensation du fantôme était grandement dûe à cette voix flottante, souvenir réaliste d'un corps disparu. Comme nous l'avions vu précédemment avec la voix du peintre hollandais Rembrandt, reconstituée à partir de ses autoportraits et d'informations sur sa personnalité, recréer l'identité vocale d'un individu permet, d'une certaine façon, de le ramener à la vie.

Cloner la voix d'un individu permettrait également de copier une des caractéristiques principales d'un comédien afin d'avoir ses qualités de jeu à disposition sans nécessiter sa présence physique. Nous pourrions imaginer dans un futur proche que le clonage vocal d'un comédien pourrait lui éviter des heures de post-synchronisation pour un film où la prise de son direct ne permettait pas une bonne intelligibilité ou que son jeu n'était pas assez bon. La société *Adobe* a présenté en 2016, une démonstration de *Adobe VOCO*, un projet de recherche autour d'un logiciel qui permet de monter de la voix, par le texte. L'algorithme permet de modifier ou d'ajouter des mots à l'écrit afin que la voix générée en *text-to-speech* propose une phrase réaliste au niveau du timbre mais également au niveau de la prosodie, permettant à l'utilisateur entre 8 intonations différentes pour chaque mot généré.

Ici la vidéo de lancement, où on change les mots d'un acteur en modifiant le scripte : https://www.youtube.com/watch?v=I3l4XLZ59iw

Depuis trois ans, *Adobe* n'a donné aucune nouvelle de l'avancement du projet mais celui-ci a soulevé de nombreuses questions, notamment sur les possibles détournement du discours d'un individu que cela pourrait amener. Changer la prosodie d'un individu, changer donc son intonation et donc son jeu est toutefois possible avec la nouvelle version du *plug-in* d'*Izotope, RX 7*, qui propose un « *dialogue contour* » et permet donc de rectifier la mélodie d'une voix, afin de corriger un problème de jeu par exemple.

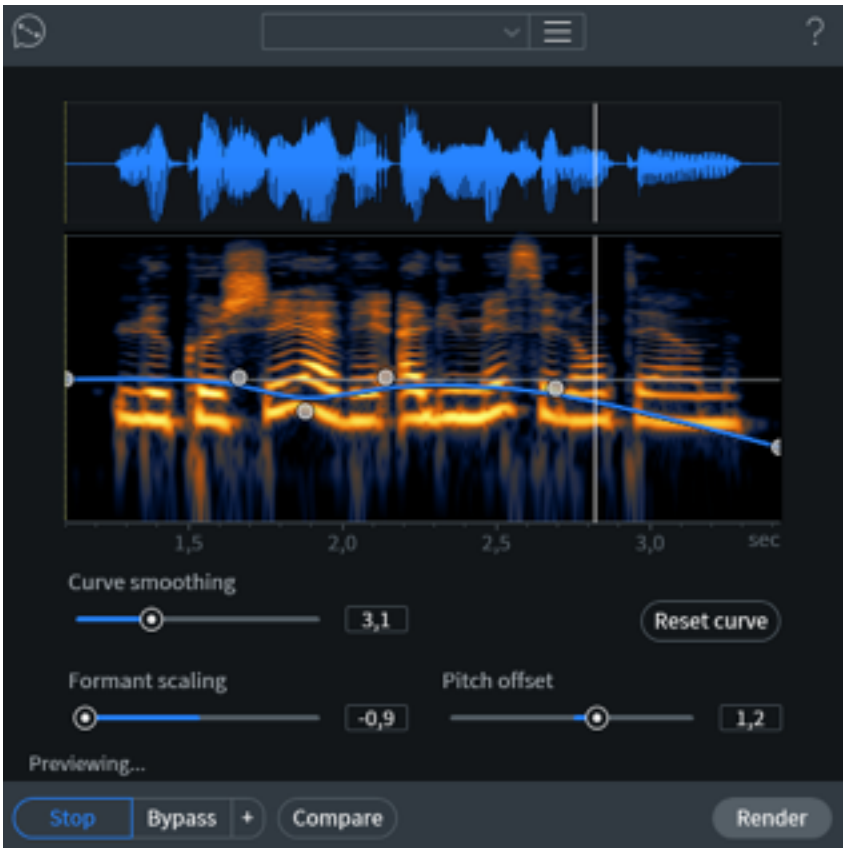


Figure 17 : capture d'écran de l'outil *dialogue contour* d'*Izotope RX 7*, la ligne bleue étant la ligne de pitch à appliquer, changeant au plus bas de 2 semi-tons et rendant ainsi la fin de la phrase descendante, conclusive.

b. Cloner la voix : détourner l'identité

Il ne faudrait pas voir la théorie du complot dans ce paragraphe mais, en essayant simplement de pousser les possibilités de cette technologie, il semble important d'énoncer les problèmes et les dérives d'une telle technologie. Nous l'avons vu, les possibilités qu'offrent le clonage vocal sont grandes et les détournements possibles d'une telle science sont aussi vastes. Comme toutes les nouvelles technologies, les utilisateurs repoussent les limites qui leur sont offertes. Notamment parce que la définition même de cette voix artificielle a besoin d'un support de reproduction pour s'incarner, les dérives *en ligne*, à travers un écran, sont directement concernées. L'utilisation d'un avatar, c'est-à-dire la mise à distance du « soi connecté » et de sa véritable identité pourrait être une des raisons de ces manipulations visant à modifier les images des autres.

Figure 18 : à gauche, l'acteur jouant la doublure de Paul Walker et à droite, le résultat après synthèse pour recréer le visage du comédien le plus réaliste possible

Au départ de ces recherches sur le lien entre visage et voix, je me suis aperçue que le détournement d'identité, ou du moins, le clonage d'une identité passait d'abord par l'image. Des technologies coûteuses et demandant un nombre d'heures conséquent permettent par exemple, à l'industrie du cinéma, de cloner le visage d'un acteur décédé ou d'un

acteur « rajeuni » sur le corps d'un autre comédien, créant alors un *morphing* entre les deux identités pour ainsi faire réapparaître à l'écran le visage d'un acteur qui n'était pas présent sur le tournage. Cette utilisation du clone sert le divertissement et l'industrie du film, sans porter préjudice à quiconque si les accords de droit à l'image sont respectés. Ainsi, c'est le frère de Paul Walker qui incarne la doublure de l'acteur décédé dans *Fast and Furious 7*, pour que l'image de synthèse sur son visage soit la plus réaliste possible (voir *Figure 18*). Le détournement d'identité sur le web a été récemment mis en avant par l'utilisation des « *deep fakes* »[63], c'est à dire des images créées à partir d'un apprentissage profond basé sur des réseaux de neurones d'une IA qui permettent un calcul précis d'une image de synthèse crédible. Ainsi, les images créées paraissent si réalistes, qu'il est difficile de distinguer à l'œil nu, devant son écran d'ordinateur, les vraies images des fausses. Celui qui accède à ces images est donc confronté à un détournement d'identité.

Figure 19 : « Deep fake » du visage de Barack Obama sur un corps de modèle, lors d'un défilé en maillot de bain, extrait de la vidéo « *Obama Swimsuit Competition Deepfake* » de la catégorie « Divertissement » de Youtube (février 2019)

[63] Littéralement "profond faux"

Ces usurpations d'identités via ces « *deep-fakes* », même si elles sévissent en ligne, peuvent avoir un réel impact dans la vie réelle des victimes de ces manipulations numériques. Nombreuses victimes se sont retrouvées humiliées en public après que leur visage ait été placé sur des vidéos pornographiques et de nombreux témoignages existent en ligne pour lutter contre cette pratique[64]. Les personnes visées sont principalement des personnalités publiques, dont leur image est facile d'accès sur internet (et permet un apprentissage profond plus réaliste car le nombre d'images est plus grand), mais elles peuvent aussi être des individus inconnus, du moment que celui qui crée le « *deep-fake* » possède assez de matière pour créer un rendu visuel qui soit assez crédible.

Outre l'usurpation d'identité, ce détournement d'image peut également partir d'un fichier audio pour créer une synchronisation labiale réaliste, comme le présente le chercheur Supasorn Suwajanakorn dans son projet *Synthesizing Obama: Learning Lip Sync from Audio*[65] (voir *Figure 20*). Le système développé par l'équipe du chercheur Supasorn Suwajanakorn convertit premièrement un fichier audio en un mouvement labial, qui servira de modélisation pour générer une texture photographique réaliste de la bouche, afin de la placer sur le visage voulu. Avant le rendu, la texture de la bouche et la vidéo cibles sont assemblées et ré-agencées afin que les mouvements de têtes apparaissent naturels, correspondant à l'audio de départ

---

[64] ici le témoignage de Noelle Martin, victime de l'utilisation de son image sur des vidéos pornographiques : https://www.youtube.com/watch?v=PctUS31px40

[65] SUWAJANAKORN Supasorn, SEITZ Steven M., and KEMELMACHER-SHLIZERMAN Ira, *Synthesizing Obama: Learning Lip Sync from Audio,* University of Washington, 2017.

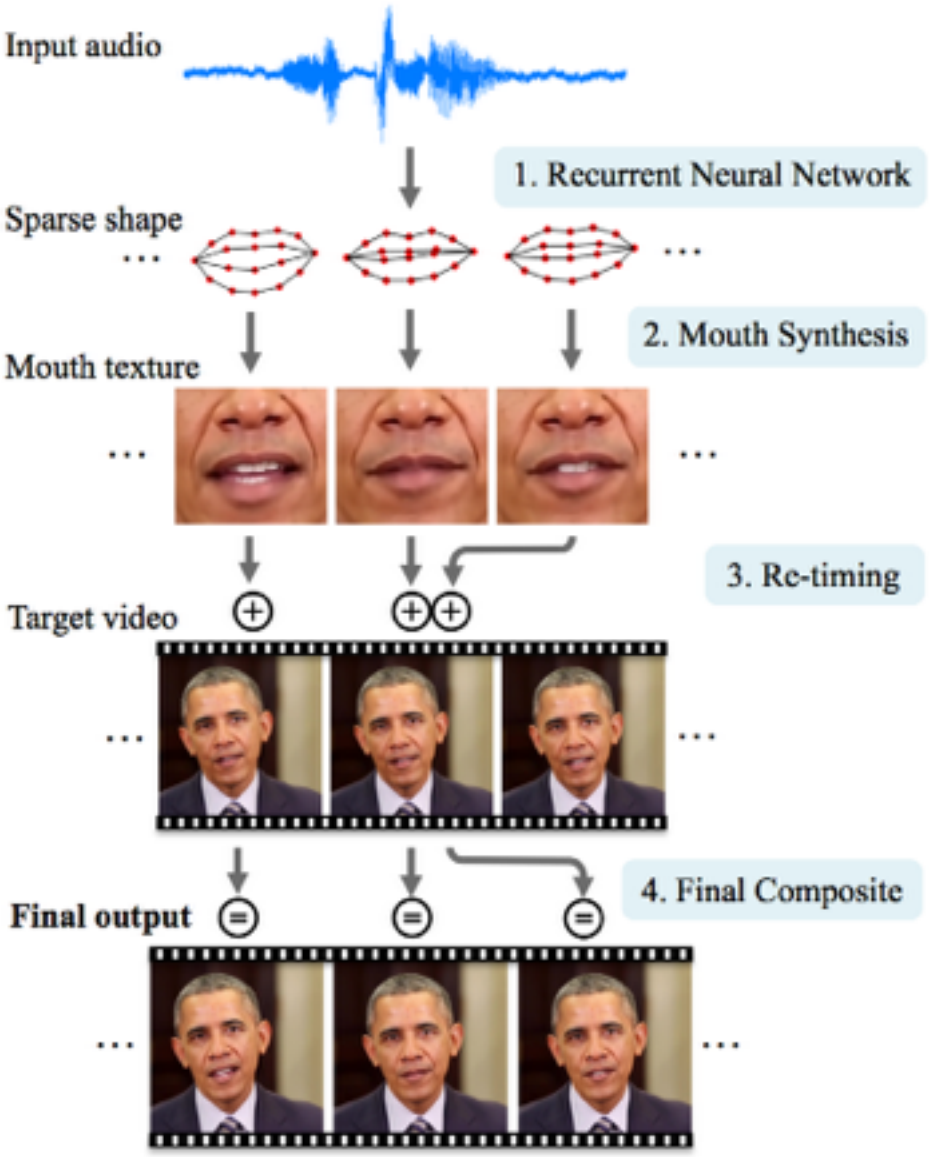


Figure 20 : principe du projet *Synthesizing Obama: Learning Lip Sync from Audio*[66]

[66] *Learning Lip Sync from Audio :* exemple de l'audio à écouter ici : https://www.youtube.com/watch?v=AmUC4m6w1wo

Une fois que le mouvement labial correspondant à l'audio a été créé, il est donc possible de placer ce même audio « dans la bouche » de n'importe qui, en choisissant différentes images de bouches, correspondant au visage cible.

Figure 21 : instantané de la vidéo de présentation de Supasorn Suwajanakorn sur TEDTalks, faisant écouter un même échantillon audio prononcé par différents visages, synchrones[67].

De la même manière que ces images synthétisées demandent de posséder un large nombre de photogrammes, faire la démarche de cloner une voix implique de posséder des enregistrements sources, témoignages de notre identité. À partir de ces enregistrements (plus la quantité est grande, plus le clone sera fidèle à la voix), il serait donc possible de réaliser un clone vocal d'un individu *à son insu*. L'importante

[67]Supasorn Suwajanakorn sur TEDTalks URL (avril 2019) : https://www.ted.com/talks/supasorn_suwajanakorn_fake_videos_of_real_people_and_how_to_spot_them?language=fr#t-421489

banque du Royaume-Uni HSBC[68] a pris le parti de passer par la reconnaissance vocale pour permettre à ses clients d'avoir accès à leur compte ; un des commentaire des consommateur en ligne revient souvient : « ça marche, même avec un rhume! ». Un clone vocal permettrait donc d'avoir accès à des données uniquement protégées par la voix d'un utilisateur.

La communication téléphonique permettant une mise à distance du corps et de la voix, il est donc possible de faire parler une voix de synthèse à la place d'un individu, sans que celui qui soit au bout du fil puisse s'en rendre compte. De plus, la bande passante du téléphone étant réduite (de 300 Hz à 3400 Hz, alors que la voix française a un spectre généralement concentré entre 50Hz et 7000Hz), ce moyen de communication peut notamment masquer certains artefacts qui seraient la preuve que la voix est bien générée par une machine et non produite par un corps humain.

c. Le masque vocal : manipulation du contenu par l'émotion à l'ère de la post-vérité

Aujourd'hui, 68% de la population française s'informe en ligne (incluant les réseaux sociaux)[69]. Le numérique embrasse l'idée libérale d'un espace public appréhendé comme un vaste marché d'idées où la vérité est la simple résultante d'une mise en concurrence des informations[70]. Ce lien entre la libre circulation des idées et l'économie d'un marché capitaliste encourage la consommation d'une multitude de

---

[68] Site officiel de HSBC faisant la promotion de VoiceID "bye bye passwords", "All you need is your voice" (Tout ce dont vous avez besoin est votre voix) : https://ciiom.hsbc.com/ways-to-bank/phone-banking/voice-id/#tab-2

[69] NEWMAN Nick, « Reuters Institute Digital News Report 2018 »

[70] LOVELUCK Benjamin, *La démocratie au prisme du numérique* (2017), publication sour la direction de TROUDE-CHASTENET, P., Edition, CLASSIQUES GARNIER.

sources d'informations, aussi fiables que mal informées. Au moment de l'élection de Donald Trump et du référendum de l'appartenance du Royaume-Uni à l'Union Européenne en 2016, le dictionnaire d'Oxford fait entrer dans son registre le terme de « *post-truth* » (mot à mot : post-vérité), qu'il défini ainsi : « Relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief ». Une définition française pourrait être : « qui fait référence à des circonstances dans lesquelles les faits objectifs ont moins d'influence pour modeler l'opinion publique que les appels à l'émotion et aux opinions personnelles ». Écrire aujourd'hui sur le concept contemporain d'ère de post-vérité serait un autre sujet de mémoire, mais il me semble important de mentionner ce concept, où l'émotion et en particulier, la voix d'une personnalité lors d'un discours devient plus importante que le contenu en lui-même.

Comme nous l'avons vu précédemment, la voix est l'un de nos principaux outils de communication. Elle manipule l'opinion, que ce soit en politique mais aussi dans la publicité, dans le monde du travail, ou dans un contexte privé. Un discours, par son émotion, touche à l'instinct corporel de l'auditeur avant de toucher sa raison. L'individu moderne aurait-il besoin d'être interpelé émotionnellement (dans son inconscient), au détriment d'être sollicité par une rhétorique construite sur un raisonnement fiable ? De la même manière, il serait important de mentionner la figure d'autorité dans l'accès à la vérité que représentent les intelligences artificielles qui investissent notre espace particulier (dans un contexte familial, via des utilitaires de domotique chez les particuliers). Ces outils ne peuvent mentir, ni entretenir de mythe (posez à *Siri* la question : est-ce que le Père Noël existe ?) et ont donc une place non négligeable dans un foyer, car ils peuvent manipuler l'opinion à la fois par le contenu de leurs propos mais aussi, potentiellement, par le ton de leur voix.

Comme nous le montre la vidéo de Jordan Peele imitant Barack Obama : l'auditeur peut croire au discours car le visage est synchronisé à la voix. Cette synchrèse apporte de la crédibilité au contenu, mais aussi à l'identité de la personne clonée. La synchronisation des lèvres et de la voix, peut être également détournée, comme sur cette chaîne youtube humoristique « *Bad Lips Reading* »[71] qui double des situations médiatiques (remontées) afin de décrédibiliser les discours et les personnalités publiques. Sur une de leurs vidéos[72], nous pouvons écouter le Président actuel des Etats-Unis dire : « You know, there's a few things you should know about me. I'm pretty sexy. I don't walk in nature. I hate shy kids. (…)»[73]. Le contexte de parodie et l'explicitation du processus de fabrication de ces vidéos permet de mettre en évidence l'importance du son direct comme preuve d'un discours. Le travail de montage son et de mixage de ces vidéos doublées jouent, comme au cinéma, au réalisme de l'image, pour un gain de vérité et de crédibilité.

---

[71] littéralement : mauvaise lecture sur les lèvres

[72] URL (12 février 2019) : https://www.youtube.com/watch?v=066WAeG5muE Bad Lips Reading remaniant le discours de Donald Trump du 5 février 2019, *STATE OF THE UNION*

[73] phrases traduisibles par : « Vous savez, il y a quelques choses que vous devriez savoir sur moi. Je suis assez sexy Je ne marche pas dans la nature. Je déteste les enfants timides. »

II / 2. Sécurité et protection des données vocales

« Je suis le service de réservation automatique de Google, je vais donc enregistrer l'appel »
Google Duplex avant chaque prise de contact au téléphone[74]

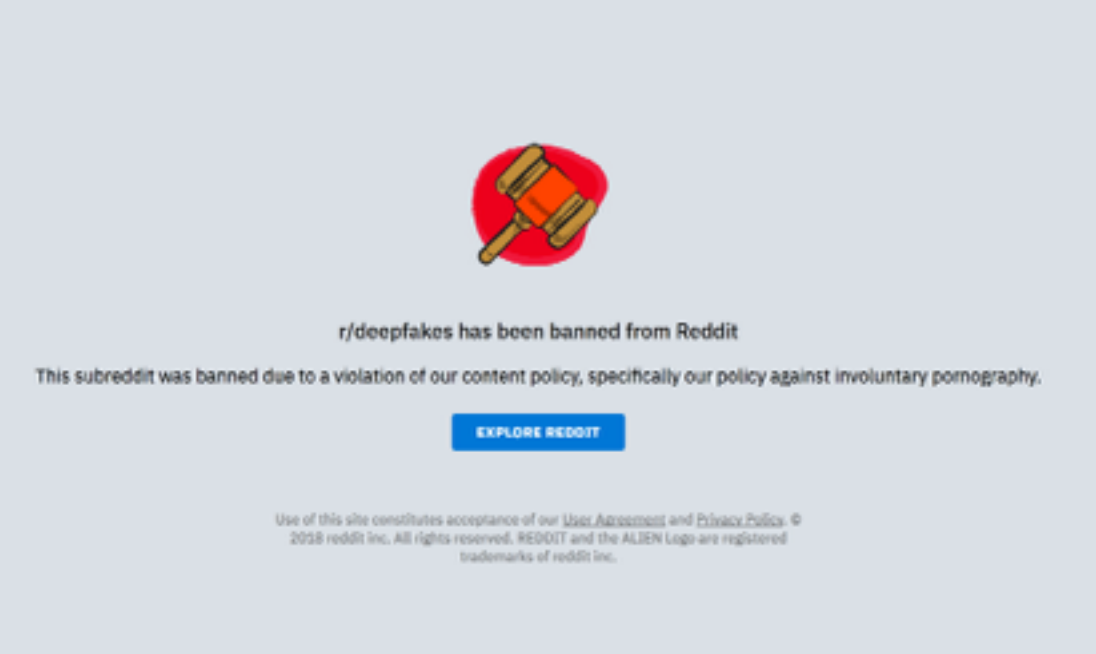


Figure 22 : capture d'écran (18 décembre 2018) du message qui annonce que l'application *deep-fake* a été bannie de son hébergeur pour violation de leur politique, notamment pour motif de pornographie involontaire

Comme le montre la *Figure 22*, une application en ligne comme *Deepfakes*, qui permettait de réaliser un deepfake vidéo en ligne, c'est-à-dire d'utiliser l'image d'un individu, en particulier son visage, pour le mettre à la place d'un visage d'une autre vidéo (malheureusement, le plus souvent pornographique), a été fermée par l'hébergeur. Dans cette partie, nous allons parler de la protection du clonage vocal. La synthèse

[74] URL (mars 2019) : https://www.bloomberg.com/news/articles/2018-05-18/google-s-duplex-ai-robot-will-warn-that-calls-are-recorded : "Google's Duplex AI Robot Will Warn That Calls Are Recorded"

vocale sait reproduire le timbre singulier d'une voix depuis plus de 10 ans mais aujourd'hui, cette technologie se démocratise pour rendre cette synthèse vocale accessible à tous. Il est désormais possible, pour un particulier, de cloner son identité vocale en ligne, en quelques phrases, avec une qualité de plus en plus grande, au fur et à mesure que les intelligences artificielles mises en jeu pour la *neural text-to-speech* continuent leur apprentissage. Aujourd'hui, la reconnaissance à l'oreille entre la voix acoustique et sa jumelle synthétique devient et continuera à être difficile, en tendant vers une différenciation qui se voudra impossible.

Dans cette partie, je vais évoquer les différentes questions qui font appel à la protection de l'identité vocale. Les parallèles que je ferai avec la détection d'image, la reconnaissance faciale ou même les lois existantes en matière d'image servent ici d'exemples pour questionner la démocratisation du clonage vocal.

a. Reconnaissance vocale : la voix démasquée

Si nous considérons donc les voix en dehors des archétypes physiologiques, nous pouvons dire qu'elles possèdent un certain nombre de marqueurs physiques qui ont très peu de chance de varier. L'identité vocale peut ainsi avoir pour fonction l'identification. Comme nous l'avons vu dans les différentes techniques de synthèse vocale, il est possible à l'aide de réductions statistiques, de produire des marqueurs relatifs à une voix unique pour permettre ensuite l'identification vocale de cette même voix.

En 2017, la CNIL (Commission nationale de l'informatique et des libertés) a autorisé 9 établissements financiers français à expérimenter pour une durée de 12 mois la reconnaissance vocale comme moyen d'identification des utilisateurs. Pour la protection des données vocales des utilisateurs, elle préconise alors *de privilégier les dispositifs qui garantissent à la*

personne concernée de garder la maîtrise de son gabarit. Cela suppose de stocker le gabarit biométrique :

- *soit sur un support détenu par la seule personne concernée,*
- *soit en base de données sous une forme inexploitable car illisible sans un secret détenu par la seule personne concernée.*[75]

Un système de reconnaissance vocale fonctionne sur la base d'une empreinte vocale faite à partir de l'enregistrement de la voix d'un individu. De cet enregistrement est extrait une quantité de données définissant avec exclusivité le locuteur[76]:

- une estimation de la fréquence moyenne de la voix à travers une analyse de la densité spectrale
- une réduction statistique à travers le modèle de Markov caché (*HMM, Hidden Markov Model*)
- une analyse de probabilité à travers un mélange de Gaussienne (courbe permettant d'effectuer une moyenne)
- une recherche de répétition en utilisant des algorithmes de réseaux de neurones
- une série de prédictions organisées autour d'un « arbre de décision »
- des représentations matricielles et vectorielles

Si la voix de synthèse a besoin d'un support technique pour être diffusée, la reconnaissance vocale doit pouvoir déceler si la voix est artificielle ou non malgré ce support. La bande de fréquences du haut-parleur du téléphone, réduisant le spectre de la voix, ne doit pas empêcher la

---

[75] extrait du site officiel du CNIL : URL (avril 2019) : https://www.cnil.fr/fr/la-cnil-autorise-lexperimentation-de-dispositifs-biometriques-de-reconnaissance-vocale-par-des

[76] DESPRATS, Pierre, *Recherche sur l'identité vocale dans la synthèse vocale et sa relation à la disfluence*, mémoire (sous le direction de Thierry Coduys et Greg Beller), Son, ENS Louis-Lumière, 2014.

reconnaissance de cette dernière. Cette analyse statistique des descripteurs de la voix permet de faire une réduction des données pour définir une unique empreinte vocale, correspondant à la voix d'un individu.

La reconnaissance vocale permet donc de s'identifier via notre empreinte vocale mais nous verrons par la suite qu'elle permet également de différencier la voix organique de son clone numérique. À la manière dont des intelligences artificielles peuvent reconnaître de « faux visages », générés eux-même par d'autres intelligences artificielles, nous pourrions imaginer entraîner des réseaux de neurones en reconnaissance vocale pour déceler les « vraies » voix des « fausses ». La première intelligence artificielle fabrique une voix de synthèse, la seconde décèle les défauts et artefacts de la voix, la première apprend de ses erreurs et construit une nouvelle voix de synthèse encore plus réaliste, qui entraîne une nouvelle fois la seconde intelligence artificielle à déceler les défauts de la nouvelle voix etc. Cette boucle d'apprentissage permettrait aux algorithmes de synthèse de tendre vers des voix générées, indifférenciables de leur voix acoustique d'origine. Cette technologie pourraient donc franchir ce que nous avons appelé plus haut, la vallée de l'étrange de la voix. J'invite le lecteur qui aimerait en savoir davantage sur le sujet, à lire : « *A comparison of features for synthetic speech detection* », Publication de l'*INTERSPEECH 2015*, M. Sahidullah, T. Kinnunen et C. Hanilci. Ce texte explique différentes techniques pour détecter une voix artificielle, mais ce sujet est encore une fois si vaste qu'il ferait tout à fait l'objet d'une recherche approfondie.

À titre d'exemple, un prototype d' « *AI anti-AI* » *(mot à mot : Intelligence Artificielle anti Intelligence Artificielle)* a été développé par une société australienne « *DT R&D* »[77] où un détecteur de voix artificielle peut être placé sur l'oreille, ou bien proche d'un haut-parleur, afin d'avertir de la nature synthétique de la voix entendue. Dans leur

[77] https://rnd.dt.com.au/anti-ai-ai-a-wearable-ai-device-244900e4d71c

prototype, leur système reçoit un corpus de « voix organiques » et de « voix artificielles » (dans leur exemple, les enregistrements vocaux sont de Donald Trump) et apprend a les différencier. Lorsqu'un nouvel échantillon audio est présenté à l'Intelligence Artificielle, elle est donc en mesure de reconnaître si la voix perçue est artificielle ou non. Le capteur placé sur l'oreille prévient l'auditeur à l'aide d'une fine plaque thermoélectrique qui devient froide si une voix artificielle est détectée. Bien sûr, ce gadget est seulement un exemple d'application privée d'une « *AI anti-AI* » et la compagnie qui le développe utilise une application extérieur appelée *TensorFlow*[78] pour faire l'analyse par réseaux de neurones entraînés à reconnaître les voix artificielles. Elle ne met donc en lumière que le fait que des utilisations de cet apprentissage profond sont possibles, notamment pour lutter contre l'usurpation d'identité (« *spoofing* » en anglais).

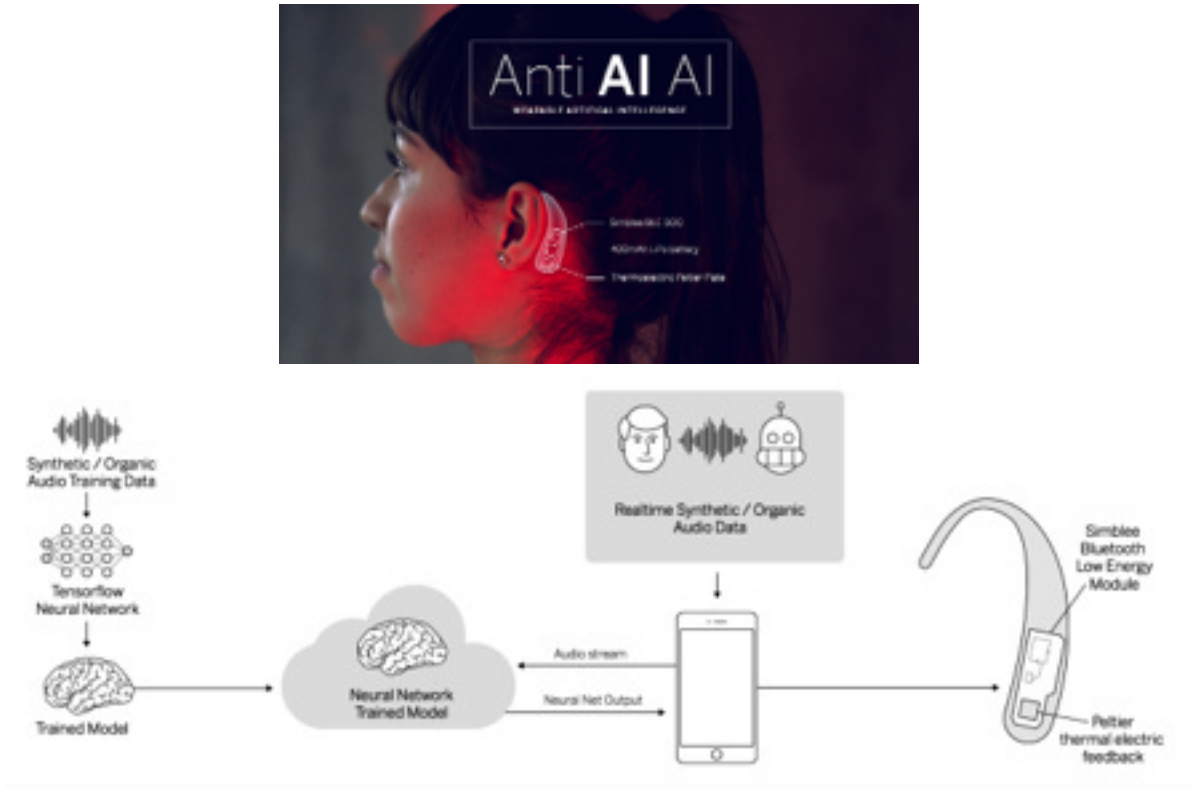


Figure 23 : photographie publicitaire et schéma de fonctionnement du prototype d'*AI anti-AI* développé par *DT R&D* (Melbourne)

[78] https://www.tensorflow.org/learn

b. Transparence des voix artificielles

Depuis le 1er octobre 2017, le Ministère de la Santé français impose par l'article « L. 2133-2 » du code de la santé publique la mention « photographie retouchée » aux images à usage commercial montrant des mannequins dont le traitement visuel vise à affiner ou épaissir leur silhouette. Lorsque le contenu est ambiguë sur la nature des modifications visuelles, afin de ne pas « duper » celui qui regarde, les publicités se doivent donc être explicites et transparentes sur leurs modifications visuelles.

Figure 24 : gros plans sur des messages publicitaires
portant la mention « photographie retouchées »,
photographies prises dans le métro parisien, août 2018

J'utilise volontairement ici un premier parallèle avec la culture visuelle, car il n'existe aujourd'hui pas encore de système législatif obligeant les entreprises ou institutions qui génèrent ou diffusent du contenu audio (par exemple celles qui permettent de cloner une voix en ligne, ou qui font simplement du *text-to*-speech) de « marquer » le contenu audio créé pour préciser que la voix entendue a été synthétisée. Dans une situation où il pourrait avoir un doute sur la nature de la voix, cette précision pourrait lever l'ambiguité de la nature de l'interlocuteur (humain ou non).

Les voix de synthèse générées aujourd'hui seraient-elles trop humaines ? Le but d'une telle voix est-il de se faire passer pour humaine ou simplement de faciliter la communication orale ? Les compagnies qui développent leurs propres avatars vocaux, qui prennent de plus en plus la place de voix humaines sur nos plateformes numériques, doivent être transparentes sur l'identité de leur avatar. *Google* avait été critiqué au moment de la présentation publique de son assistant *Google Duplex* (l'Intelligence Artificielle qui permet à *Google Assistant* passer des appels à la place du consommateur), car la personne qui avait été contactée au téléphone n'avait aucune idée que son interlocuteur n'était pas humain. En réponse à cette critique, *Google Duplex* qui effectue un réservation par téléphone, se présente aujourd'hui comme la voix d'un algorithme, sans prétendre être la voix d'un humain. De plus, elle doit préciser que la conversation est enregistrée, ce qui implique que la voix de l'interlocuteur soit aussi récupérée. Libre à l'interlocuteur d'accepter ou non la conversation avec cette IA.

La perception de la voix dépend également de son statut : connu ou non. Si je sais que la voix que j'entends est une voix de synthèse, je vais la percevoir différemment que si je ne sais rien et pense avoir affaire à une voix organique. Des tests perceptifs seraient à mener sur la reconnaissance ou non des voix artificielles mais l'annonce formelle et transparente lorsque nous écoutons une voix artificielle change notre rapport à l'écoute. Si la voix de *Google Duplex* doit annoncer qu'elle est de toute façon artificielle, pourquoi chercher à reproduire parfaitement une parole humaine ? L'interaction avec ces assistants par la voix (*Vocal Command Interface*) permettra une fluidité d'échange entre l'homme et la machine, partageant alors le même canal de communication. Cependant, leur statut de « machine » ne sera pas remis en question, car ces voix resteront incarnées par des corps numériques, des téléphones, des tablettes, des enceintes, et leur nécessité de support technique ne pourra jamais remplacer un corps vivant. L'utilisation de cette communication orale avec les machines rendra floue cette barrière, cette

distinction entre un être humain derrière une machine ou bien une simple machine.

Cette ambiguité de relation est ce qu'évoque Spike Jonze dans son film *Her* (2014), où un homme tombe amoureux d'une Intelligence Artificielle (IA) qui lui sert d'assistante. La voix de cette IA est incarnée par l'actrice Scarlett Johansson (et non par une voix de synthèse). La personnalité de l'intelligence artificielle passe donc entièrement par la voix de l'actrice : jeune, sensuelle, amusante etc. L'ambiguité entre l'Homme et la machine est ici questionnée, car le personnage principal éprouve de « vraies » émotions vis à vis de cette IA, qui a de son côté des relations avec plus d'un millier d'utilisateurs à la fois. Même si le personnage a conscience que cette voix n'est pas celle d'un être humain, il ne peut s'empêcher des émotions qui seraient habituellement éprouvées pour d'autres êtres humains. La fonction de ces voix de synthèse n'est pas uniquement d'informer, elle touche le sensible, fait appel aux émotions avant la raison.

Figure 25 : l'acteur Joachin Phoenix dans *Her*, écoutant dans son oreillette la voix de l'Intelligence Artificielle dont il tombe amoureux

c. Tatouages numériques pour marquer les voix de synthèse ?

Le tatouage numérique, aussi appelé filigrane, est un élément permettant de marquer un fichier digital afin de traquer un fichier, de notifier un droit d'auteur, de vérifier l'authenticité d'un contenu particulier ou de toute autre information dépendante du fichier d'origine. Le tatouage numérique est conçu de telle sorte qu'il soit impossible à retirer du fichier. Dans ce paragraphe, je ne ferai qu'énumérer différentes techniques de tatouages numériques sonores car ce sujet de protection de données audio pourrait faire l'objet d'un mémoire entier. Pour plus d'informations sur le sujet, je pourrais recommander la lecture de l'article *Overview of Audio Watermarking Techniques* issu de International Journal of Emerging Technology and Advanced Engineering, 2012.

Figure 26 : image soumise à des droits d'auteur et donc protégée par un tatouage numérique (« watermark » en anglais) empêchant toute exploitation de l'image d'origine

Dans le cadre d'un fichier audio, le tatouage numérique peut être audible ou non. Afin de protéger un fichier audio de droit d'auteur, nous

pouvons superposer au fichier un élément audible (par exemple, une voix prononçant le nom de celui qui possède les droits d'auteur, ou un clic rendant inutilisable le fichier audio). Mais les tatouages numériques se veulent la plupart du temps inaudibles, afin que l'utilisateur puisse tout de même écouter le contenu du fichier hôte. Différentes façons de créer des tatouages numériques ont été proposées ces dernières années : l'ajout de fréquences en dehors du spectre audible (20Hz - 20kHz), la substitution du *LSB (Least Significant Bit)*, le codage de phase (*Phase Coding*) et l'ajout d'écho caché *(Echo Hiding)*[79]. La technique la plus répandue pour tatouer les fichiers audio reste néanmoins le tatouage à spectre étalé : le *SSW (Spread Spectrum Watermaking)*. Cette technique de tatouage numérique « inaudible » consiste à marquer le fichier hôte d'une séquence définie de bruit (appelée pseudo-noise en anglais et plus connu sous l'acronyme *PN*) sur une large bande de fréquences. L'utilisateur qui veut retirer le tatouage, doit connaître le « *PN* » correspondant. Ce tatouage numérique est utilisé pour la haute sécurité, car l'utilisateur non autorisé ne peut pas identifier l'information cachée dans le fichier hôte.

Dans le cadre de la synthèse vocale, un tatouage numérique implémenté dans le fichier audio généré par la synthèse *text-to-speech* pourrait permettre d'identifier et donc de différencier une voix de synthèse d'un enregistrement acoustique. Cependant, cette identification reste dépendante de l'utilisation d'un outil numérique, capable de récupérer cette information cachée, incorporée dans le fichier audio de synthèse. L'oreille humaine resterait donc dépendante de la confiance accordée à la détection numérique. La responsabilité des compagnies qui développent les outils publics de clonage vocal ont donc la lourde responsabilité de protéger les voix de synthèse des utilisateurs.

---

[79] DAVARYNEJAD M., SEDGHI S., BAHREPOUR M., WOOK AHN C., AKBARZADEH M., COELLO COELLO C. A., Article : *Detecting Hidden Information from Watermarked Signal using Granulation Based Fitness Approximation*, dans Applications of Soft Computing: From Theory to Praxis, Springer, Series: Advances in Intelligent and Soft Computing, Volume 58/2009.

d. Anonymisation de la voix

Par extension à la reconnaissance de voix artificielles, l'apprentissage profond a permis, du côté de la reconnaissance faciale de générer de nouveaux visages réalistes de personnes inexistantes[80]. Dans le domaine de l'audio, une transformation de la voix permet déjà de la déformer pour lui ôter ses caractéristiques identitaires. Cependant, c'est dans le contenu (habitudes orales, emploi de certains mots ou expressions, disfluences) que l'identité vocale peut-être reconnue. La voix modifiée serait donc une voix anonyme, comme le visage généré par une intelligence artificielle ci-dessous.

Figure 27 : visage synthétisé par une intelligence artificielle extrait du site « this person doesn't exist »

[80] URL (mars 2019) : https://www.thispersondoesnotexist.com/ : "This person doesn't exist" invente une nouveau visage toutes les 2 secondes, en se basant sur l'apprentissage des intelligences artificielles appelé GAN (Generative Adversarial Networks, introduit en 2014), leur but étant de synthétiser des images artificielles, qui ne sont pas distinguables de visages authentiques.

Quelles sont les modifications vocales qui permettent à une signature vocale d'être anonyme ? À quel point une modification vocale ne permet-elle plus une identification vocale ?

e. Cloner une voix : quelles lois pour réguler l'utilisation ?

Nous l'avons vu plus haut, notre voix, comme notre image, sont des marques de notre identité. S'il est possible de cloner notre voix, à quel moment notre voix digitale nous appartient-elle ? Qui contrôle l'utilisation des clones vocaux ? Est-ce que je peux cloner la voix de n'importe qui ? À qui appartiennent légalement ces voix ? Pour réaliser mon clone vocal, je dois réaliser plusieurs enregistrements : où sont-ils conservés, et pour combien de temps ? Mes enregistrements sont-ils protégés ?

En lisant les conditions générales d'utilisation du site de la start-up Canadienne *Lyrebird AI*, qui permet de réaliser son propre avatar vocal anglais en ligne, gratuitement, nous pouvons noter que leur onglet « *Biometric* » rend explicite que (mars 2019)[81] :

- « Please note that Biometric Data may include, without limitation, data that may allow someone to identify/contact you and non-public data. »
  Veuillez noter que les données biométriques peuvent inclure, sans limitation, des données pouvant permettre à une personne de vous identifier / vous contacter et des données non publiques.
- « We store Biometric Data that we collect on a server hosted by our third-party service provider, Amazon (https://aws.amazon.com/). »
  Nous stockons les données biométriques que nous collectons sur un serveur hébergé par notre fournisseur de services tiers, *Amazon* (https://aws.amazon.com/).

---

[81] LYREBIRD AI, *« Lyrebird, Inc. Agreement and Written Release Regarding Collection, Storage, and Disclosure of Biometric Data »*, site officiel, URL (consulté en mars 2019) :
https://about.lyrebird.ai/terms/biometrics

- « Texas and Washington residents : We will permanently delete your Biometric Data after it is no longer needed for the Purpose for which we collected it. Please note that the Purpose may last for an indefinite time period. »
  Résidents du Texas et de Washington : Nous effacerons vos données biométriques de manière permanente, lorsque qu'elles ne serviront plus aux fins pour lesquelles nous les avons collectées. Veuillez noter que ces raisons peuvent avoir une durée indéterminée.
- « Lyrebird's current purpose in collecting Biometric Data is to allow us to provide, maintain, improve, and enhance our technology, including our Services and machine learning models. »
  L'objectif actuel de *Lyrebird AI* en matière de collecte de données biométriques est de nous permettre de fournir, maintenir, améliorer et augmenter les performances de notre technologie, y compris nos services et nos modèles d'apprentissage automatique.

Les conditions ne sont pas encore claires sur l'appartenance juridique de la voix de synthèse. Pouvons-nous s'opposer à la diffusion de sa voix artificielle ? En créant son propre clone vocal, nous donnons toutefois à une entreprise privée des enregistrements de notre voix et nous leur confions alors notre identité vocale. La protection du clone reste cependant la responsabilité de l'entreprise qui l'a créé. D'un point de vu légal, comme la voix comporte des données personnelles (qui permettent l'identification), son clone qui comporte (en théorie) les mêmes données, reste protégé par le Code Pénal français et est également soumis au droit d'auteur. Un « copyright » vocal pourra donc être à attribuer lors de l'utilisation de la voix d'un individu par un tiers.

Même si le clonage vocal tend à se démocratiser, les données vocales qui servent à la fabrication du clone restent encore propriétés des entreprises qui proposent ce service. Cloner sa voix revient à confier son identité vocale à une entreprise, responsable de la protection de cette donnée.

PARTIE 3

Partie pratique de mémoire : THE VOICE IS VOICES - *installation sonore autour du clonage vocal*

Dans cette partie je vais me concentrer sur mon utilisation de la synthèse vocale pour créer mon propre clone. Cette approche relevant au départ d'un travail d'autoportrait vocal, s'est ensuite transformée en une expérience sensorielle autour de notre capacité d'écoute critique face à la rencontre d'une voix de synthèse et de sa jumelle organique. J'ai dans un premier temps essayé de cloner ma voix via des applications trouvées sur internet, gratuites, qui m'ont permis de créer des clones approximatifs en anglais et français. La crédibilité de mon clone vocal français a été possible avec l'aide du chercheur Nicolas Obin de l'IRCAM, et avec l'utilisation de l'algorithme de synthèse vocale *ircamTTS*. J'ai ainsi pu converser avec ma propre voix de synthèse et écrire l'installation sonore *THE VOICE IS VOICES*, dont les étapes de création ainsi que les retours du publics seront également présentés dans cette dernière partie.

III / 1. Création de "ma propre" voix de synthèse

Les outils publics pour créer sa propre voix de synthèse

Une variété d'entreprises en ligne proposent de réaliser la synthèse vocale de sa propre voix. En fonction de la langue, nous pouvons utiliser différentes sociétés comme *Candyvoice*[82], *Cereproc*[83] ou *Lyrebird AI*[8] (liste non exhaustive, de nombreuses entreprises étant en train d'éclore dans le domaine, proposant une synthèse vocale gratuite ou non). Dans cette partie, je vais me concentrer sur la réalisation de ma propre synthèse vocale à partir de l'algorithme de la *start-up* canadienne *Lyrebird*

---

[82] URL (février 2019), Site officiel de Candyvoice https://candyvoice.com/demos/voice-conversion?target=jeanne#studio et leur slogan "Be a voice, not an echo"

[83] URL (mars 2019) : https://www.cereproc.com/products/cerevoiceme

[84] (février 2019), Site officiel de *Lyrebird AI* https://beta.myvoice.lyrebird.ai/

AI. À la suite de la création de son avatar vocal, l'utilisateur peut générer un ficher audio en *text-to-speech* et télécharger librement l'audio créé.

a. Clone vocal anglais : *Lyrebird AI*

Essais de ma propre synthèse vocale

Lyrebird AI propose de créer son propre avatar vocal anglais en quelques minutes d'enregistrement. Afin de créer son propre avatar vocal anglais, il est conseillé d' « enregistrer sa voix dans un endroit calme, avec un bon micro et d'effectuer le plus d'enregistrements possibles »[85] (temps d'enregistrement conseillé de 5 minutes). Les enregistrements doivent correspondre à des phrases proposées par leur algorithme, qui visent à couvrir l'ensemble des consonnes et des voyelles en anglais. Sur leur site, nous pouvons apprendre que la synthèse vocale qu'ils proposent « fonctionne particulièrement pour l'accent anglais américain ». Ils recommandent de "jouer" les phrases à enregistrer, c'est-à-dire d'ajouter de l'expressivité dans nos enregistrements, en ayant un rythme de lecture assez rapide.

Exemple de phrases à enregistrer sur le site de *Lyrebird AI* : « It was a tough battle. Our brigade had suffered terrible losses. »

Lyrebird AI est un réseau de neurones qui progresse avec l'ensemble des enregistrements qui lui sont fournis. Sa synthèse source-filtre permettrait donc de synthétiser sa propre voix avec un nombre très réduit d'enregistrement ; notre timbre jouant le rôle de source et les autres voix appartenant à *Lyrebird AI* jouant le rôle de filtre. Plus l'algorithme apprend de voix, plus il grandit et permettra de synthétiser le timbre de n'importe quel utilisateur avec un nombre de plus en plus réduit d'enregistrements.

---

[85] conseils tirés du site officiel de *Lyrebird AI*

Afin de comprendre comment l'algorithme pouvait progresser en ajoutant des enregistrements, j'ai dans un premier temps comparé deux échantillons générés à partir de deux sessions d'enregistrements distinctes.

Méthode

- J'ai enregistré 148 phrases avec le micro de mon ordinateur dans un endroit calme et j'ai créé mon avatar vocal pour lui faire dire : « *This voice has been generated after 148 recordings with my computer's microphone. What do you think ? Is it relevant ? Can you believe it's a real recording or can we hear it is an artificial voice ?* »[86]

Le résultat de cette expérience s'écoute ici :
https://soundcloud.com/m-lia-roger/melias-artificial-voice-148-samples-lyrebird

Après avoir supprimé tous les anciens enregistrements effectués avec mon micro d'ordinateur, j'ai ré-enregistré 800 phrases avec un microphone *Neumann TLM 103*, relié à une interface audio *Focusrite 2* entrées, dans un lieu calme et mat. J'ai re-créé mon avatar vocal pour lui faire dire : « *This voice has been generated after 800 recordings with a good microphone. What do you think ? Is it relevant ? Can you believe it's a real recording or can we hear it is an artificial voice ?* »[87]

Le résultat de cette seconde expérience s'écoute ici :
https://soundcloud.com/m-lia-roger/melias-artificial-voice-800-samples-lyrebird

---

[86] traduction : « Cette voix a été générée après 148 enregistrements avec le microphone de mon ordinateur. Qu'en pensez-vous ? Est-ce pertinent ? Pouvez-vous croire que c'est un vrai enregistrement ou peut-on entendre que c'est une voix artificielle? »

[87] traduction : « Cette voix a été générée après 800 enregistrements avec un bon microphone. Qu'en pensez-vous ? Est-ce pertinent ? Pouvez-vous croire que c'est un vrai enregistrement ou peut-on entendre que c'est une voix artificielle? »

Résultats

Les fichiers audio générées par mes avatars vocaux ont :

- une fréquence d'échantillonnage de 22 050 Hz
- une quantification par échantillon de 16 bits
- un canal audio (mono)

La fréquence d'échantillonnage étant de Fe = 22 050 Hz, elle ne permet de numériser une fréquence maximale que de 10 025 Hz (théorème de la fréquence de Nyquist où f_{max} = Fe / 2), soit une fréquence bien en dessous des transitoires générées par une voix enregistrée pleine bande (20-20kHz). Cette contrainte ne permet donc pas de synthétiser un fichier vocal de qualité suffisante pour être comparée à l'enregistrement source (fréquence d'enregistrement de Fe = 44.1 kHz).

La différence entre les deux exemples de qualité d'enregistrement est quand même sensible, notamment dans la fidélité du timbre, bien plus proche du mien après tous les enregistrements effectués au *Neumann TLM 103*.

De plus, dans les phrases à enregistrer, je suis retombée sur les mêmes exemples à prononcer, comme si j'étais arrivée au terme de ce que l'algorithme pouvait réellement intégrer dans le calcul de la voix de synthèse. Cependant, le temps de calcul de la voix a varié de 2 heures de calcul pour les 148 enregistrements réalisés avec le microphone de mon ordinateur jusqu'à 8 heures de calcul pour les 800 réalisés avec le *Neumann TLM 103*.

Expressivité

Sur le timbre et l'articulation

La voix synthétisée me semble proche de mon timbre. Ma langue maternelle étant le français, j'entends mes défauts de prononciation sur les échantillons générés en anglais, ce qui est en soit fidèle à ma manière d'articuler.

Sur la vitesse

Les phrases synthétisées sont « lues » à vitesse normale. La phrase 2 (« This voice has been generated after 800 recordings with a good microphone. What do you think ? Is it relevant ? Can you believe it's a real recording or can we hear it is an artificial voice ? ») a été générée en un fichier de 13 secondes alors que mon rythme naturel de lecture prend 14 secondes à lire cette même phrase. Sur d'autres exemples, avec davantage de ponctuations, le rythme de lecture peut sembler rapide mais cela demanderait de réaliser une étude spécifique pour ce paramètre.

Sur la proximité

La voix de synthèse est générée avec une sensation de proximité dûe à sa matité. Aucune réverbération n'est présente pour lisser l'ensemble, la voix est brute, au plus proche de sa source.

Sur les émotions perçues

Il est possible d'ajouter un « *smiley* »[88] au texte généré par la voix. Cependant, ce dernier n'influence pas « la manière » dont serait prononcée la phrase mais *Lyrebird AI* a, pour l'instant, fait le choix de sonoriser les *smileys*. Par exemple, à un visage souriant sera associé un rire. À une image de pomme est associé un son de quelqu'un qui croque dans un fruit. Cette association ne permet donc pas de modifier l'expressivité de la voix mais rajoute un élément ponctuel de sonification d'émoticône. Nous pourrions imaginer par la suite que l'ajout d'un « :) » (émoticône qui sourit)à la fin d'une phrase permette de générer une phrase dont l'émotion perçue est la joie.

[88] Un *smiley* est un petit dessin figurant une expression faciale et peut se décliner en différents "émoticônes" qui permettent d'incruster des dessins et symboles de toutes sortes à un fichier texte.

La prosodie, c'est-à-dire la mélodie de la phrase est cependant influençable au moment de l'écriture par la ponctuation. En écrivant « ! » pour une phrase exclamative, la hauteur est augmentée d'un demi-ton. Les questions marquées de « ? » sont aussi entendues comme telles. Les points de suspension « … » permettent de marquer un temps avant le début d'une autre phrase. La dernière phrase du fichier texte à générer est généralement marquée par un *pitch* descendant à la fin, comme pour clore l'ensemble des phrases. Toutefois, lorsque nous écrivons plusieurs fois les mêmes mots (par exemple : « I don't know. I don't know. I don't know. »), les premiers qui se suivent sonnent exactement de la même façon, sans variabilité dans le prosodie.

Tentative d'ajout d'expressivité par D.A.V.I.D, « IRCAM Tool »

Après avoir importé un échantillon vocal généré par *Lyrebird AI*, j'ai voulu essayer d'ajouter une expressivité artificielle en utilisant l'outil D.A.V.I.D (Da Amazing Voice Inflection Device) développé par le groupe de recherche *Cream* à l'Ircam. Cet outil fonctionne sur un environnement *MaxMSP* (testé sur les versions du logiciel *Max6* et *Max7*) et peut travailler en « temps réel » avec un délai inférieur à 15 ms. En utilisant les pré-réglages de cet outil, j'ai tenté de faire sonner l'extrait audio « This voice has been generated after 800 recordings with a good microphone. What do you think ? Is it relevant ? Can you believe it's a real recording or can we hear it is an artificial voice ? » de manière joyeuse, apeurée et triste.
Ces essais sont disponibles ici :
https://soundcloud.com/m-lia-roger/sets/emotional-artificial-voice-neutral-happy-afraid-sad

Même si l'on peut effectivement reconnaitre des attributs caractéristiques des émotions synthétisées, mon intention d'ajouter *a posteriori* de l'émotion, de l'expressivité à la voix s'est avérée rendre un résultat encore plus artificiel que l'original. Cette artificialité provient des figures mélodiques et rythmiques que nous pouvons entendre dans la voix, comme la régularité du vibrato appliqué pour la voix « apeurée ».

b. Clone vocal français : *CandyVoice*

CandyVoice est une compagnie française qui propose de créer son avatar vocal en ligne. Un nombre de « tickets » permet ensuite de générer et télécharger l'audio généré. Il est conseillé d'enregistrer au minimum 80 phrases, sachant qu'une première voix de synthèse peut-être réalisée dès 40 phrases enregistrées. L'algorithme de synthèse vocale utilisé pour générer le clone vocal gratuitement sur leur site est la synthèse *text-to-speech* de *Microsoft*.

Après 275 phrases enregistrées en studio, voici l'exemple que j'ai pu générer : https://soundcloud.com/m-lia-roger/vocal-clone-french

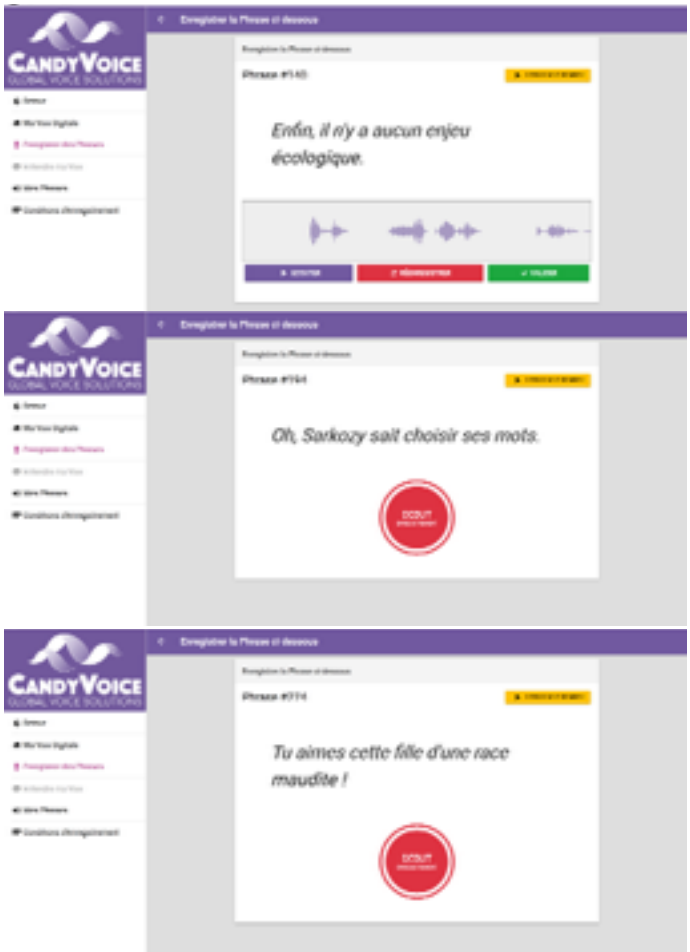
Le texte lu par la voix artificielle était : « Cette voix a été générée via l'application de *text-to-speech* appelée Candyvoice, après 275 enregistrements de bonne qualité. Qu'en pensez-vous? Est-ce que cette voix ressemble à la mienne ? Est-elle pertinente? Est-ce une question ? Ceci est une affirmation. Ceci est une phrase en suspension… »

Pour l'enregistrement du corpus vocal, j'ai utilisé un microphone *Neumann TLM 103* dans studio insonorisé. Le microphone entrant dans une interface numérique *Focusrite Scarlett* ayant une fréquence d'échantillonnage dépendante du site sur lequel l'enregistrement est stocké (même si je pouvais régler ma fréquence d'échantillonnage à 44.1 kHz, mes fichiers étaient dépendant de l'enregistrement *via* une plateforme en ligne).

L'enregistrement des phrases en français était simplifié par la possibilité de pouvoir écouter un exemple de prononciation avant d'effectuer l'enregistrement. L'exemple permettait de relever quelles liaisons devaient être faites, ou bien la manière dont certains chiffres devaient être prononcés (par exemple : « 1915 » était prononcé « mille neuf cent quinze » et non « dix neuf cent quinze »).

Cependant, à titre personnel, je fus très surprise du contenu des phrases proposées pour l'enregistrement. Certaines phrases avaient un contenu très orienté politiquement et contenaient parfois des propos misogynes. Je me permets donc ici de fournir quelques captures d'écran des exemples que j'ai trouvé les plus choquants. Je tiens à rappeler ici que les phrases sont normalement choisies pour leur contenu syllabique, comprenant une large variété de phonèmes qui visent à couvrir le spectre de la langue française. À la lecture de ces phrases, il m'est arrivé de douter de la bienveillance de ce site quant à l'utilisation des données personnelles, notamment de mes enregistrements. Suite à ce doute, j'ai contacté le service de communication de *CandyVoice*, qui a voulu me préciser que « les phrases utilisées ont été sélectionnées parmi 175 000 autres car elles sont phonétiquement équilibrées ».

Figure 28 : captures d'écran tirées des phrases à enregistrer sur le site de *CandyVoice*



Résultats

Les fichiers audio générées par mes avatars vocaux ont :

- une fréquence d'échantillonnage de 32 000 Hz
- une quantification par échantillon de 32 bits flottants
- un canal audio (mono)

Expressivité

De nombreux artefacts sont audibles, compromettant l'intelligibilité de la parole. De plus, je ne reconnais pas mon timbre ou une manière personnelle d'articuler. Cependant, les paramètres de vitesse, hauteur, formants et expressivité sont ajustables pour modifier « manuellement » la qualité de notre voix générée afin qu'elle soit plus fidèle à notre timbre. Même en essayant de modifier ces paramètres, je ne parviens pas à retrouver mon identité vocale. De plus, peu importe si la phrase écrite est une affirmation, une question, une phrase exclamative ou en suspension, l'intonation reste la même.

La voix générée est si loin de mon identité vocale qu'il m'est impossible d'émettre un avis sur d'autres paramètres tels que la proximité, la vitesse, ou les émotions perçues.

c. Utilisation de la synthèse *text-to-speech* de l'IRCAM

Recommandations pour l'utilisation d'IrcamTTS

Afin de réaliser un clone vocal crédible et semblable à mon timbre « organique », le chercheur en synthèse vocale Nicolas Obin m'a permis d'utiliser la synthèse *text-to-speech* de l'IRCAM (Institut de Recherche et Coordination Acoustique/Musique), dans les locaux de l'Institut, avec leur programme codé sur *Matlab*.

Cet outil n'est donc pas public, la synthèse a été réalisée dans un cadre académique, me donnant accès de manière individuelle à ce programme. Afin de réaliser mon clone vocal, il a fallu que je fournisse un nombre

d'enregistrements assez conséquent à l'algorithme afin qu'il ait accès à tous les phonèmes que je voulais synthétiser. Pour que le système puisse associer mes phonèmes à un alphabet phonétique (*XSAMPA*[89]), je devais également fournir les fichiers texte correspondants à chaque phrase du corpus vocal.

Pour l'enregistrement du corpus vocal, j'ai utilisé un microphone *Neumann TLM 103* dans studio insonorisé. Le microphone entrant dans interface numérique *Focusrite Scarlett* ayant une fréquence d'échantillonnage de 44.1 kHz. Après un essai avec 6 phrases (6 fichiers audio et 6 fichiers texte correspondant) qui a permis de remettre le programme en état d'être utilisé, j'ai donc fourni 1 171 phrases (ainsi que 1 171 fichiers texte correspondants) au *ircamTTS* afin qu'il synthétise ma voix.

Vous pouvez entendre un exemple de la voix ici :
https://soundcloud.com/m-lia-roger/french-artificial-voice

La phrase synthétisée est : « Bonjour, je suis la voix de synthèse de Mélia Roger ».

Les indications pour obtenir une synthèse semblable au corpus proposé était d'avoir une articulation parfaite, correspondante exactement au texte écrit (chaque pause entendue dans la phrase doit être marquée par une virgule à l'écrit). De plus, la prosodie doit rester assez neutre : ne pas jouer les mots entre parenthèses, avec des guillemets et garder un débit et volume assez stable dans l'élocution.

---

[89] XSAMPA : eXtended Speech Assessment Methods Phonetic Alphabet (alphabet phonétique étendu) utilisé au moment de la traduction du texte en phonèmes à concaténer

Choix du corpus de textes pour les enregistrements

Sachant que j'allais lire pendant plus de six heures, j'ai commencé par enregistrer des textes liés à mon sujet de mémoire (articles scientifiques, essais sur la linguistique, mémoires de fin d'études). J'imaginais que *Le plaisir du texte* de Roland Barthes allait me permettre une lecture fluide, objective et neutre, mais ses jeux de ponctuation et son utilisation des point-virgule ne me permettait pas d'obtenir une lecture à voix haute qui définissait des phrases de manière claires et audibles. Les nombreux termes en italique et entre parenthèse ne me facilitaient pas la lecture (et c'est d'ailleurs bien de ce plaisir du texte lu intérieurement dont il est question dans le texte). Ainsi, la concentration et l'adaptation de ma lecture au fur et à mesure des problèmes rencontrés me demandaient une trop grande attention et cela s'entendait dans les enregistrements. En effet, ma voix était hésitante, marquant des pauses là où il n'y en avait pas, se reprenant, la parole devenant presque « orale » au lieu d'être « lue ».

Mon second choix a été de me tourner vers un texte « pour enfant », afin de me concentrer sur une prononciation parfaite. J'ai donc choisi *Le petit prince* d'Antoine de Saint-Exupéry. Le texte contient beaucoup de dialogues et de phrases très courtes. Comme je devais découper mes fichiers audio par phrase (au sens grammatical du terme, c'est-à-dire une unité composée d'éléments ordonnés, capable de porter l'énoncé complet d'une proposition), je me retrouvais avec un grand nombre de phrase ayant peu de contenu (exemple : « Ça suffit. » qui doit être interprété comme une seule phrase, et donc générer un fichier audio et un fichier texte indépendant). Les dialogues entre les différents protagonistes étaient également difficiles à ne pas « interpréter », et ma voix n'était alors plus « neutre » dans son jeu. De plus, le texte comprenait beaucoup de répétitions et n'avait donc pas une large palette de différents phonèmes. Je cherchais alors un contenu à lire qui contienne des phrases plus conséquentes, sans que j'ai besoin de les remanier.

Mon troisième choix s'est porté vers la lecture de *Boule de Suif* de Guy de Maupassant. Cherchant des phrases assez longues, j'ai choisi les descriptions détaillées naturalistes de l'auteur pour soutenir ma lecture. Cependant, de nombreuses dates, abréviation (Mme pour Madame) et de noms propres ont empêché une lecture fluide, ne sachant pas toujours choisir leur prononciation. J'ai essayé d'adapter la lecture en remplaçant alors les noms propres par des pronoms personnels mais cela me demandait alors de ré-écrire les phrases *a posteriori* et donc m'a fait perdre beaucoup de temps de ré-écoute et de re-écriture.

Le choix le plus pertinent en terme de facilité de lecture et de richesse phonétique a été la lecture de pages Wikipédia. Les articles objectifs sur les couleurs, les nuages, les amphibiens et la lumière m'ont permis d'avoir un corpus audio conséquent avec des phrases souvent longues comprenant des termes scientifiques pouvant être difficiles à articuler mais dont le sens restait simple et qui m'ont donc permis de garder un rythme de lecture fluide et stable.

Exemple de phrase prononcée qui a servi à la synthèse de mon clone vocal : « Les équations vues précédemment comportent certaines hypothèses qui prennent pour acquis que les mouvements de l'air et la condensation se produisent assez lentement pour que la pression, la température et le contenu en eau s'adaptent graduellement. »

Performance de la voix

J'ai essayé de ne pas trop espacer les sessions d'enregistrement afin que mon timbre reste stable en fonction des jours : mêmes horaires, repos, alimentation saine, absence d'alcool et de cigarettes durant la période d'enregistrement.

Le choix de garder un timbre neutre au moment des lectures avait pour moi deux utilités. La première était d'avoir une stabilité d'élocution au

moment de la synthèse, réalisée à partir d'un corpus d'expressivité homogène. La seconde raison était une intention de simuler différentes expressions en « post-production », à l'aide de transformation vocale, comme l'utilisation de D.A.V.I.D pour modifier les émotions entendues par la voix, impliquant de partir d'une base neutre pour modifier les émotions.

L'écoute de ma propre voix au moment de l'enregistrement puis par la suite, lors du découpage des fichiers audio par phrase m'a été douloureuse. Ce ton « neutre » me paraissait assez « ennuyé » et a largement affecté mon humeur lors des étapes de découpage. J'ai retrouvé ce dont m'avait parlé Jean-Julien Aucouturier (chercheur de l'équipe Cream de l'IRCAM, travaillant sur D.A.V.I.D) : il y a une corrélation entre l'émotion exprimée par le locuteur et celle perçue par l'auditeur (Cf. présentation de la recherche PARTIE I/c).

Afin d'avoir une parfaite corrélation entre l'audio prononcé et la phrase écrite, j'ai effectué une dernière écoute de vérification de ponctuation, afin que chaque pause marquée à l'oral soit aussi présente à l'écrit.

Découpage des fichiers audio et texte

Afin de faire correspondre mes phonèmes à l'alphabet phonétique qui allait ensuite servir à la synthèse, il fallait que je fournisse un fichier audio par phrase (marquée par un point), associée à son fichier texte correspondant. Les échantillons audio devaient comprendre du silence au début et à la fin du fichier, pour éviter tout problème de lecture « trop courte » (échantillon inférieur à 1 seconde). Les fichiers audio ont une longueur variant de 1 à 36 secondes.

Dans les silences des débuts et fins de fichiers, j'ai laissé des respirations, qui peuvent ensuite être associées à un silence au moment de la synthèse, rendant ma voix artificielle encore plus réaliste, car respirante. Les phonèmes associés pour la synthèse sont prélevés dans les fichiers

audio générés et comportent donc les mêmes bruits de bouche et précisions d'articulations. La fréquence d'échantillonnage utilisée pour la synthèse est de 44.1 kHz, comme pour le corpus vocal de telle manière qu'il n'y ait aucune conversion au moment de la synthèse. Les fichiers textes à fournir au programme *IrcamTTS* sur *Matlab* demande un format *UTF-8*, à définir lors de l'export du fichier texte sur le logiciel *TextEdit*. J'ai donc recopié tous les textes sur des fichiers *.txt* par phrase, exportés dans ce format.

Analyse du corpus audio pour ircamTTS

J'ai réalisé les enregistrements et le découpage des fichiers audio / textes et les ai ensuite fournis à Nicolas Obin afin qu'il puisse faire les premiers tests avec ce corpus vocal. Le résultat de la voix est pour moi une expérience assez étrange, car les fichiers audio générés ont été écrits par lui et lorsque j'ai reçu les fichiers audio, j'ai eu la surprise d'entendre une voix portant ma signature prononcer des mots que je n'avais jamais dits. Ainsi, j'ai pu entendre ma propre voix prononcer « Bonjour, je suis la voix de synthèse de Mélia Roger » et « Tu m'avais beaucoup manqué, et toi ? ».

Afin de générer la voix, l'audio s'appuie sur le corpus vocal que j'avais fourni et les données acoustiques de ma voix ont dû être analysées. Ainsi, sa fréquence fondamentale moyenne, ses paramètres glottiques (à quel point la glotte est en tension ou relâchée), la répartition des fréquences dans le temps, la variation de sa hauteur ont été analysées. Le réseau de neurones va ensuite associer chaque phonème à une portion d'audio donnée, en tenant compte de ces paramètres pour ainsi les utiliser au moment de la synthèse vocale : il choisira *via* une analyse statistique les portions d'audio servant à la synthèse en fonction de la ligne mélodique principale de la phrase. En effet, plus le corpus vocal servant à la synthèse vocale est large, mieux l'algorithme pourra choisir les échantillons qui correspondent à la prosodie voulue.

Par la suite, nous pourrions imaginer, pour synthétiser différentes expressions dans la voix, que les corpus soient répartis en fonction de l'émotion jouée. Ainsi, un corpus « neutre » pourrait être associé à un corpus « joyeux », ou « triste » afin de moduler les expressivité de la voix. La synthèse par concaténation d'unité viendrait alors piocher dans différents corpus en fonction de l'expression voulue à différents moments de la phrase. Cependant, cela demanderait une performance d'acteur conséquente, car il faudrait réussir à enregistrer au moins six heures de phrases « joyeuses », six heures de phrases « tristes » etc. Nous avons vu que la modification de l'expressivité de la voix de synthèse (comme par l'utilisation de D.A.V.I.D sur une voix de synthèse basée sur un corpus vocal « neutre ») *a posteriori* n'avait pas amené, dans notre cas, un résultat naturel. Si rendre la synthèse vocale expressive n'a pas un résultat crédible en « post-production », on pourrait penser la rendre expressive dès l'enregistrement du corpus vocal, tout en permettant à la synthèse d'être cohérente dans la construction prosodique de ses phrases.

III / 2. Partie Pratique : Réalisation de l'installation sonore "THE VOICE IS VOICES"

THE VOICE IS VOICES est une installation sonore autour du clonage vocal. L'installation a été montrée les 12 et 13 avril 2019 dans les locaux de l'ENS Louis-Lumière. En *annexe 5*, vous trouverez le carton de présentation de l'installation, que le public était invité à lire avant d'entrer dans l'espace sonore. Ici, un extrait de la présentation du travail au public :

« Vous entrez dans un espace à deux voix : l'une organique et l'autre, son clone numérique. Cette voix de synthèse a été réalisée à partir d'un corpus vocal de plusieurs heures d'enregistrements qui ont permis de générer des phrases en synthèse *text-to-speech*. Ainsi chaque phrase artificielle générée a été écrite, puis lue par une machine. Circulez

librement dans tout l'espace sonore et prêtez l'oreille pour discerner la vraie voix de son *fake*[90]. »

Dans cette partie, je présenterai les différentes étapes de travail qui ont conduit à la réalisation de *THE VOICE IS VOICES*, les enjeux artistiques, notamment vis-à-vis du contenu de la voix et du parcours scénographie proposé, ainsi que les questions techniques. J'analyserai également les retours d'expérience demandé au public qui a participé à l'installation.

Pour réaliser ce travail, j'ai collaboré avec Grégoire BÉLIEN (ENS Louis-Lumière, Cinéma 2020) à l'image, Adrien ZANNI (ENS Louis-Lumière, Son 2020) à la réalisation informatique et Salomé Oyallon (ENS Louis-Lumière, Photographie 2018) en direction artistique.

THE VOICE IS VOICES est aussi présentée à la ZHdK (Zürcher Horschule der Kunste) du 3 au 19 juin 2019, dans le cadre des projets de diplôme du Master Transdisciplinaire. Pour écrire le contenu de la voix, j'ai notamment eu le soutien de deux séminaires proposés par le programme Transdisciplinaire de l'école : *Spoken Words* par Delphine Chapuis Schmitz et *Nicht Schreiben, Ein Schreibworkshop* par Dominic Oppliger qui m'ont permis de présenter mon travail et mes essais vocaux au sein des conférences qu'ils proposaient.

[90] L'emploi du terme « fake » fait ici référence à « fausse copie »

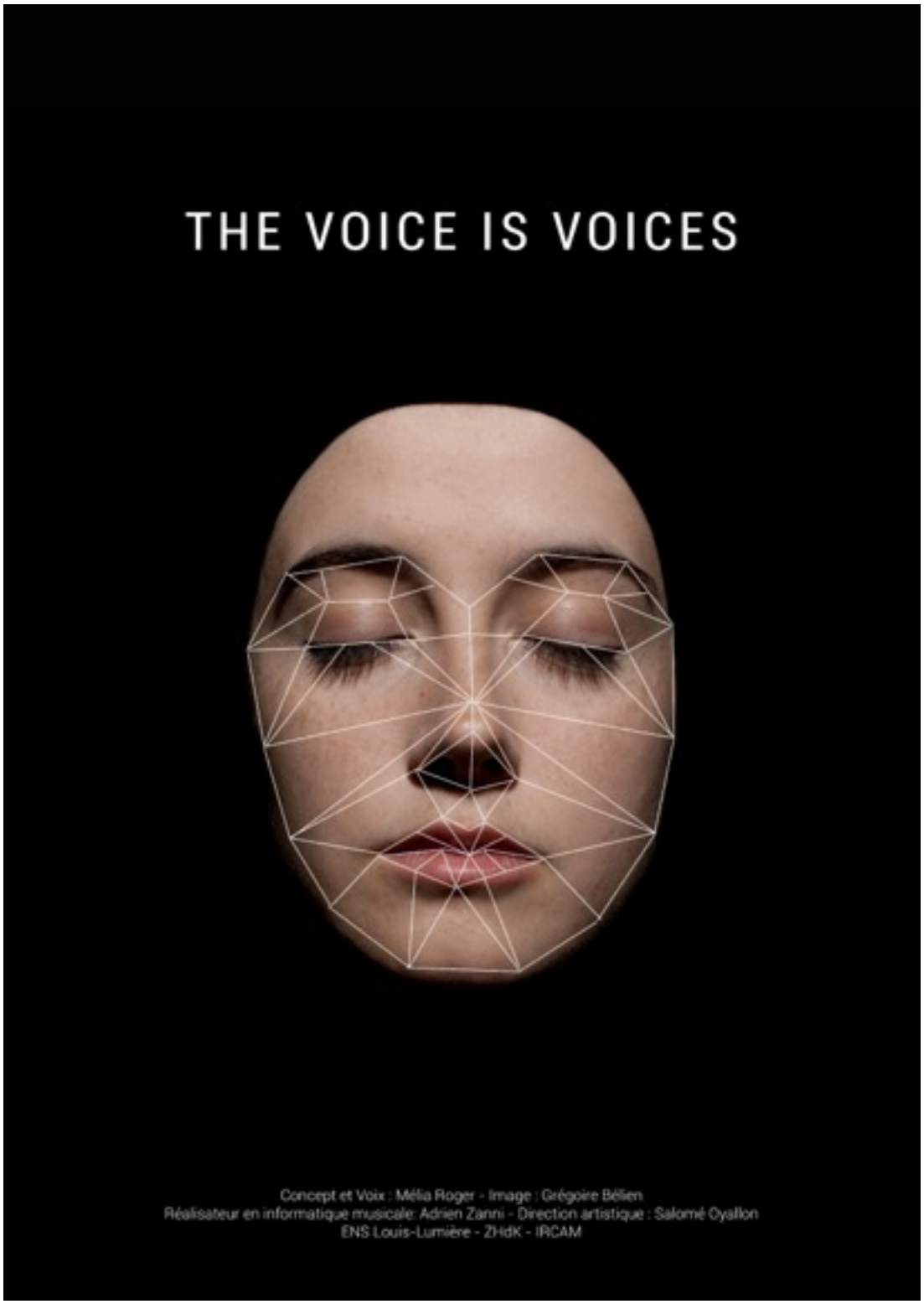


Figure 29 : affiche communication de l'installation sonore, réalisée par Salomé Oyallon

a. Corps utopique et choix du contenu

La voix a besoin d'un corps pour s'incarner. Cette voix jeune, dynamique, neutre mais déterminée est la mienne. Pourtant, elle s'ancre dans une machine, dans cet ordinateur du premier étage du laboratoire de recherche, elle peut me faire dire n'importe quoi, elle peut prendre ma place en réunion, elle peut passer des appels téléphoniques à ma place, elle pourra un jour me permettre de converser avec un « moi » du passé. Elle est déjà le passé. Elle va à l'encontre de ce que décrivait Michel Foucault dans *Le Corps utopique*[91], car ma voix, mon lien entre mon esprit, mon corps et l'extérieur se téléporte dans un objet inerte, alimenté par du courant, passant par l'écrit pour parler. Mon corps s'est transporté et ma voix m'est devenue élément contemporain. Dans son chapitre *L'art des arts* dans *Les Aveux de la chair*[92], Foucault définit l'aveu comme un mouvement de l'intériorité vers l'extérieur. Dans cette situation qui est la mienne, ma voix de synthèse représente ce mouvement extérieur, de mise à nu de la pensée, or cette pensée est écrite par autrui et ma signature, mon identité est ainsi détournée par la personne qui écrit les phrases prononcées par la synthèse.

Le choix du contenu de cette voix est donc fondamental dans la perception du « je » caché derrière la voix. Cette voix me dépossède-t-elle de mon corps ? Me permet-elle des confessions écrites, extériorisées non par mon propre aveu mais par celui de ma voix de synthèse ? Que dit cette voix ? Que vit elle ? Que rêve-t-elle ? A-t-elle son propre corps pour s'incarner ? À qui s'adresse-t-elle ? En reprenant mes lectures autour du corps, j'ai été marquée par un texte lu par Roland Barthes, une liste de « j'aime, je n'aime pas », où il arrivait à la conclusion que peu importe la

---

[91] FOUCAULT Michel, Le corps utopique ; suivi de Les hétérotopies, Paris, Éditions Lignes, 19 juin 2009, transcription radiophonique 1966.

[92] FOUCAULT Michel, *Les Aveux de la chair*, Histoire de la sexualité, parution posthume 2018, Ed. Gallimard.

nature de ses goûts, si chaque individu aimait des choses différentes, c'est bien parce que « mon corps n'est pas le même que le vôtre »[93].

Alors j'ai voulu que cette voix parle de mon corps, tel un récit autobiographique extériorisé, avoué par un *double*. Quelle dise à voix haute un texte que j'aurais écrit sur mes « j'aime, j'aime pas », qu'elle parle de chaleur, de sensation physique, qu'elle parle de ce corps qu'elle n'a pas et qu'elle aimerait avoir pour s'incarner, vraiment. Le premier texte écrit pour la voix a donc été :

« J'aime… regarder la pluie depuis l'intérieur, l'odeur des vêtements de Gabriel, le goût de la Moussaka de ma mère, le solo de cor anglais dans la Damnation de Faust de Berlioz, manger avec mes doigts, même la salade, ouvrir une nouvelle confiture, sentir la gravité au décollage d'un avion, le rire de mes amis, ne pas comprendre une langue, l'odeur d'une bougie éteinte, pleurer lorsque je fais parti d'une foule qui applaudit, l'humidité de la terre lorsque j'arrose mes plantes, le structuralisme, mettre du rouge à lèvre pour aller seule au cinéma, écouter de la poésie, la chaleur d'une bouillote, les citrons.

Je n'aime pas… sentir que je m'habitue, avoir la mâchoire qui craque, réciter des poèmes, m'épiler, devoir m'essuyer les mains parce qu'elles sont trop humides, les longues formules mathématiques, les gens qui parlent fort, les bruits de bouche, le goût des câpres, ne pas arriver à éternuer, la texture des huitres, sentir de l'eau froide pendant une douche brûlante, avoir soif, dormir avec de la lumière sur mon visage, sentir qu'un homme a du désir à côté de moi, la lenteur, le poids des sacs plastiques, savoir que je procrastine, oublier comment jouer du piano, les maladies, l'ordre, tomber. »

---

[93] https://www.youtube.com/watch?v=dYaimZKnQyE Roland BARTHES, Le Théâtre du langage, documentaire, 2015, extrait.

Ecrit comme des confessions, futiles, universelles, sans importance pour personne, mais qui cherchent à dire « j'existe », « je suis ancrée », quelque part, dans un corps, un visage, la voix s'adresse ici à celui qui possède un corps.

Mais ce texte, peut-être trop personnel, m'a fait me sentir dépossédée de cette voix, de ce corps et j'ai voulu recentrer le discours de cette voix sur elle-même. Au lieu de placer une intention personnelle, autobiographique qui aurait fait appel à « ma » voix, au « je » qui l'incarne, j'ai voulu écrire sur « la » voix. L'autoportrait s'est transformé en portrait. L'installation *THE VOICE IS VOICES* parle de cette pluralité d'une même voix, peut importe à qui elle appartient. Néanmoins, présentée de manière simultanée à sa jumelle organique, la voix de synthèse parle de cette signature vocale à partir de laquelle elle a été créée. Ainsi, elle tente de définir le corps physique auquel elle appartient, tel un portrait. Un premier texte a alors été écrit pour cette voix de synthèse, gardant un rythme répétitif et lent.

Extrait de *THE VOICE IS VOICES* :

La voix est pleine.

La voix est mienne.

La voix est corps.

La voix est femme.

La voix est jeune.

La voix est blanche.

La voix est lourde.

La voix est neutre.

La voix est silence.

La voix attire.

La voix désire.

La voix respire.

La voix ennuie.

La voix tue.

La voix divertit.

La voix doute.

La voix argumente.

La voix trompe.

La voix persuade.

La voix ressent.

La voix est lente.

La voix meurt.

Cette anaphore a aussi pour rôle d'ancrer la voix dans une forme d'apprentissage, de rituel presque initiatique. La répétition fait déjà appel à la seconde partie du texte qui fera référence à l'impossibilité de reproductibilité de la parole.

Mais si cette voix lit, si cette voix ne peut pas improviser, alors elle ne peut pas douter ? Comment pourrais-je écrire le doute ? Comment est-ce que cette voix pourrait incarner un cheminement de pensée qui oscille, hésite, tranche et revient ? Est-ce que je pourrais lui faire dire seulement des disfluences, et pourquoi pas mes propres disfluences ? Pour ce texte sur le doute et la parole improvisée, j'ai voulu d'abord m'enregistrer et mettre à nu mes propres disfluences. Vous pouvez entendre ce que j'aurais envie d'appeler *Discours du rien*, simple monologue autour de ma recherche de texte, dont tout le contenu

« utile » a été supprimé pour ne garder que mes hésitations. Ecouter le résultat ici : https://soundcloud.com/m-lia-roger/discours-du-rien
Pour aller plus loin sur les disfluences et leur pouvoir théâtral, la performance *Critical Mass* de Kerry TRIBE (2010), inspirée du film expérimental de Hollis FRAMPTON (1971)[94], met en scène la fragmentation du discours de dispute. Le jeu de répétition des mots, la montée en volume et en tension reflètent les émotions qui déconstruisent la structure du dialogue.

La voix qui dit « je » tout en doutant, la voix qui songe et qui l'admet, c'est cette répétition de la phrase que le public de *THE VOICE IS VOICES* aura retenu : « je ne sais pas ». Liée à cette hésitation de quelle voix est artificielle et quelle voix est organique, le « je ne sais pas », répété, interprété, a permis a ma voix de synthèse de rester de marbre dans une itération effrénée, alors que ma voix organique ne pouvait répéter deux fois de la même manière le même « je ne sais pas ». De plus, l'anaphore est là pour faire oublier le contenu et l'adresse de la voix, afin de permettre à l'auditeur de se concentrer sur le timbre, le grain. Une seconde partie du texte écrit pour l'installation et prononcé par la voix de synthèse est :

Je ne sais pas.

Je ne sais pas.

Je ne sais pas.

Je ne sais pas.

Je ne sais pas.

Je ne sais pas.

Je ne sais pas.

[94] *Critical Mass* de Kerry TRIBE (2010) : http://www.kerrytribe.com/project/critical-mass/#video inspirée de *Critical Mass* de Hollis FRAMPTON (1971) : https://www.youtube.com/watch?v=q8g-Pf36Hxw (la vidéo commence seulement avec le son)

Je ne sais pas.

Je ne sais pas.

Je ne sais pas.

Je ne sais pas.

Je ne sais pas.

Je ne sais pas.

Je ne sais pas.

Je ne sais pas.

Je ne sais pas.

Je ne sais pas.

Je ne sais pas.

Je ne sais pas.

Je ne sais pas.

Je ne sais pas.

Je ne sais pas.

Cette phrase est présente pour incarner le doute, mais aussi pour incarner la reproductibilité de la voix de synthèse qui, lorsque je lui écris deux fois la même phrase, n'apporte pas de variation prosodique dans la répétition, tel un copier-coller, un duplicata d'un même mot. Cette répétition d'une même phrase est une référence à l'œuvre sonore *For Children* de Bruce Nauman (2015)[95] autour de l'apprentissage. Or, ma voix organique, ne peut par nature répéter deux fois la même chose. Elle ne peut être fidèle à elle-même, car elle avance dans le temps, car l'air qui remplit mes poumons n'est pas le même, car mes muscles se fatiguent, car mon corps est vivant.

---

[95] NAUMAN Bruce, For Children/Pour les enfants (2015), article au moment de son exposition à la Fondation Cartier (2015) :
http://www.sonore-visuel.fr/evenement/exposition-de-bruce-nauman

Avant d'avoir la possibilité de créer mon clone vocal en français, j'ai écrit le texte pour ma voix de synthèse anglaise, en lui donnant les mêmes phrases traduites en anglais. La voix disait alors : « The voice is full. The voice is mine. The voice is body. (…) », ce qui m'a conduit au choix du titre anglophone, gardé pour l'installation avec la voix française. Par *THE VOICE IS VOICES*, j'ai voulu parler de la pluralité de la voix, en utilisant un jeu phonétique entre « voice is » et « voices », jeu impossible à faire en français car le nom commun « voix » est invariable. De plus, je voulais rendre l'installation accessible à un public non francophone, en lui donnant un aperçu du contenu répétitif et concentré sur la parole. Je voulais être sûre d'intégrer une partie « phonétique » dans le titre, sans pour autant choisir une écriture utilisant l'alphabet *XSAMPA*. Cet alphabet international permet d'écrire en phonétique n'importe quelle langue et il est utilisé pour générer les phonèmes de ma voix de synthèse. Ainsi, dans cet alphabet, la voix s'écrit « la vwa », pour être correctement lu par la voix de synthèse. En visitant l'exposition « Artistes & Robots »[96] au Grand Palais (mai 2018), j'avais été surprise et déçue du choix de typographie utilisant le langage de programmation « Python », amenant une dimension technologique « de surface » au contenu présenté dans l'exposition. Dans *THE VOICE IS VOICES*, la notion de *text-to-speech* est implicite, elle est contenu dans ce titre fait pour être "lu" afin d'en percevoir les subtilités.

Afin d'adapter la prononciation de certaines de mes phrases pour que l'intelligibilité soit bonne, j'ai par exemple dû écrire « la voix re sang » au lieu de « la voix ressent » au moment de la synthèse. Ce jeu entre écrit et oralité m'a permis de comprendre la manière dont s'articule la synthèse.

[96] URL (avril 2019) https://www.grandpalais.fr/fr/evenement/artistes-robots

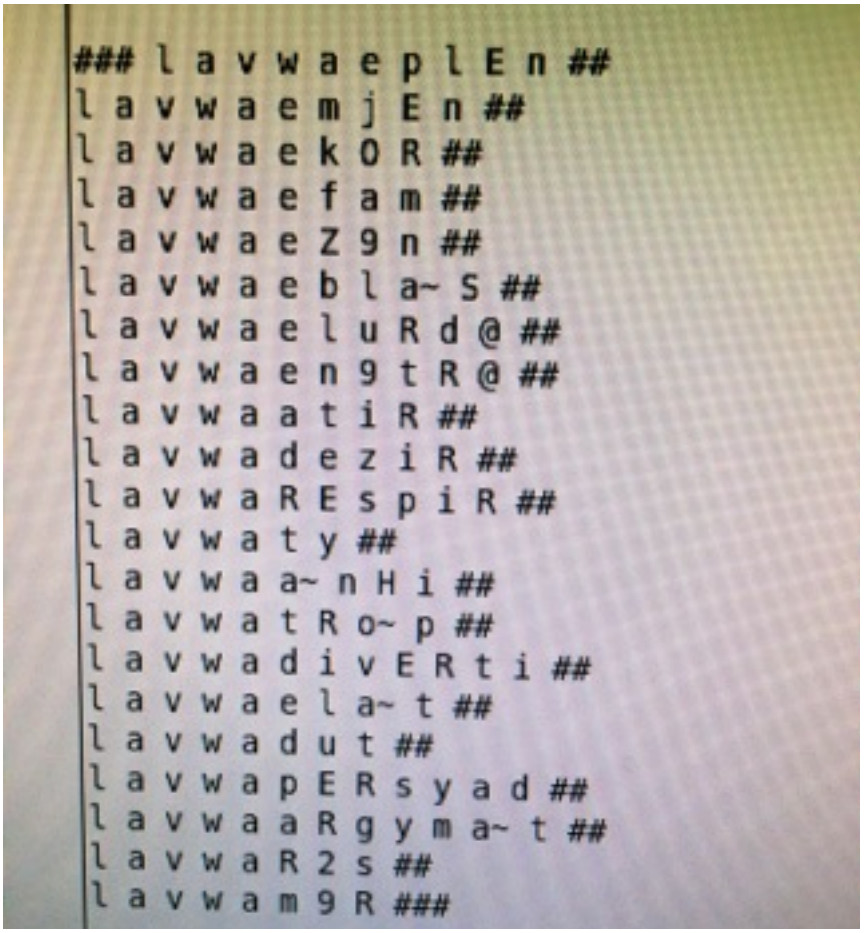


Figure 30 : Séquence phonétique du texte « La voix est pleine »
(les « ## » représentent des silences et « ### » les début et fin de phrases)

b. Imitation de la voix de synthèse

En me rendant à l'IRCAM pour générer mes phrases, j'ai été surprise de voir qu'elles n'avaient pas toutes la même réponse à la ligne prosodique implémentée dans l'algorithme, qui correspond à une phrase affirmative française. Certaines phrases écrites comme affirmatives ressemblaient à des questions (fin de la phrase qui monte), ou des phrases en suspension. Ces défauts donnaient une vie au discours qui était sensé avoir toujours la même forme prosodique. Afin de comprendre en quoi les phrases générées par la synthèse étaient proches de ma manière de parler, j'ai d'abord essayé d'enregistrer ces mêmes

phrases de la manière qui m'était la plus naturelle à prononcer. Mais, pour rendre cette voix de synthèse totalement jumelle de ma voix que j'appelle « organique », il a fallu que je l'imite. Ainsi, les deux mélodies se ressemblent parfaitement, les deux voix fonctionnent comme une symétrie, un miroir l'une de l'autre. Ici, écouter le texte « La voix est pleine » prononcé simultanément par la voix organique et la voix de synthèse :
https://soundcloud.com/m-lia-roger/la-voix-est-pleine-voix-artificielle-et-voix-organique-simultanees

Le processus d'imitation était assez spontané : j'écoutais la voix plusieurs fois, et comme si je voulais réaliser une post-synchronisation au cinéma, j'ai tenté de "jouer" de la même manière. Comme si je me retrouvais à l'écoute d'une autre personne, j'ai voulu rendre ma voix organique malléable par cette voix de synthèse afin qu'elle deviennent indiscernables l'une de l'autre. Mais pourtant, même si la précision de l'imitation permettait une synchronisation *quasi* parfaite, c'est bien dans ce *presque*, cette inexactitude que réside dans ce qui les différencie et ce qui fait de ce clone un clone et non ma voix. À quel point cette voix de synthèse est-elle humaine ? Un léger décalage à la fin d'un mot, une respiration un peu plus large, un bruit de bouche plus naturel, un sourire sous-entendu, une tension glottique plus marquée ? Les deux timbres sont jumeaux et pourtant, la différence est là, résidant dans notre perception de l'humain, dans ce ressenti « qui ne s'explique pas » (extrait d'un questionnaire du retour du public) ou du moins, que les mots peuvent difficilement exprimer. Dans la figure ci-dessous (*Figure 31*), on peut observer la même phrase prononcée par la voix organique et le clone synthétique. Nous pouvons remarquer que la prononciation de « La voix » n'a pas exactement le même rythme, et que l'enregistrement de la voix organique contient davantage de graves (autour de 100Hz) que la voix de synthèse. De plus, nous pouvons observer que la voix de synthèse a une respiration au début de la phrase (respiration créée de manière aléatoire dans les « silences » enregistrés, symbolisés par « ### » en langage phonétique), alors que la voix organique ne respire pas.

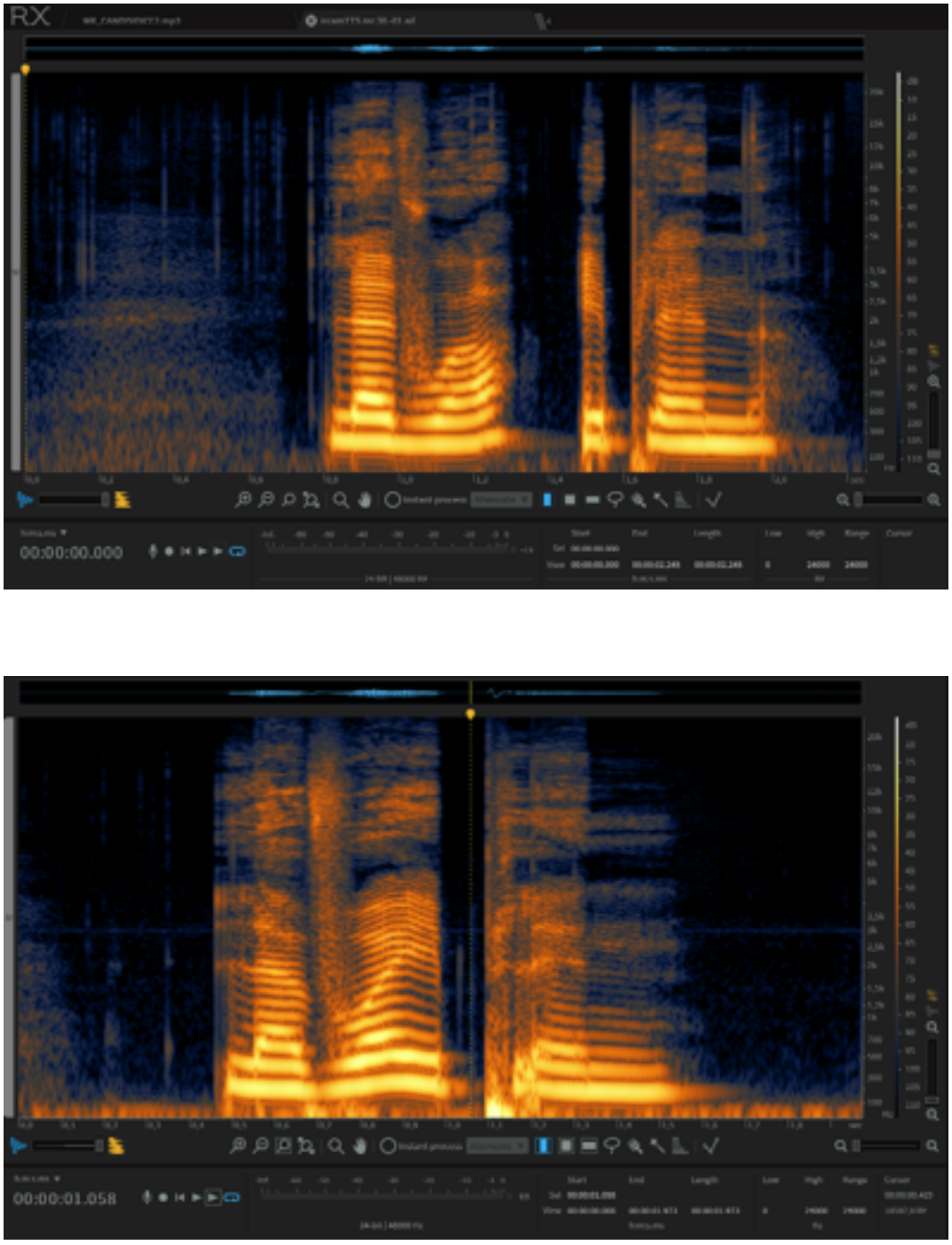


Figure 31 : spectrogrammes (représentation temps - fréquences obtenue avec l'outil *RX 7* d'*Izotope*) de « La voix est pleine » prononcé par la voix de synthèse (en haut) et son originale organique (en bas)

c. Scénographie (cheminement vers le sonore)

Une première étape de travail autour de cette voix de synthèse a été de comprendre en quoi elle pouvait être incarnée. Le travail de portrait sonore est passé par différentes étapes, différentes influences, que je vais essayer d'expliquer dans cette partie.

Une référence forte dans ce projet autour de la voix comme masque, comme identité, a été le travail de la photographe américaine Gillian Wearing. Ses portraits aux masques réalistes apportent un décalage entre la représentation de l'individu et sa propre image. Au travers de *THE VOICE IS VOICES*, c'est un parallèle sonore à son travail que j'ai voulu réaliser.

Figure 32 : Gillian Wearing, autoportrait, 2000

Le travail de Wearing autour du portrait et du masque m'ont aussi conduite vers ses films expérimentaux autour de la voix, où elle synchronise l'image d'une mère avec la voix de l'un de ses fils et inversement.

Figure 33 : image extraite de *2 into 1*, vidéo, 5 min (1997)[97]

Ce décalage créé par la distorsion entre visage et voix a d'ailleurs été repris dans l'une des installations vidéo d'Omer Fast intitulée *Looking Pretty for God (After G. W.)*[98]. C'est le rapport entre corps, identité et incarnation qui rend ces films expérimentaux étranges et dérangeants.

La simultanéité du visuel et du sonore m'avait alors marqué dans l'interprétation de la voix, regardant le visage. Incarner une voix de

---

[97] WEARING Gillian, *2 into 1* (1997) : https://www.youtube.com/watch?v=36WUgFMDY-M

[98] FAST Omer, *Looking Pretty for god (After G. W.)*, 2009 : https://vimeo.com/239612588

synthèse dans un corps organique était pour moi une manière de mettre en avant cette inquiétante étrangeté qui était le cœur du sentiment que je voulais faire ressentir dans *THE VOICE IS VOICES*. J'ai alors réalisé un autoportrait vidéo, synchronisant ma propre image à ma voix de synthèse (d'abord avec celle réalisée en anglais, sans grand succès, avec *Lyrebird AI*). À ce moment là, je n'avais pas encore le contenu de la voix, et les mots ainsi prononcés par la voix de synthèse mentionnaient une réflexion sur la confiance et sur le passé. L'idée de pouvoir synthétiser ma propre voix m'a conduite vers celle de vouloir synchroniser cette voix à mon visage synthétique, réalisé à partir de différentes photographies de mon visage, où les artefacts amenés par l'algorithme de synthèse, auraient formé une plasticité étrange mais proche d'un visage réaliste.

Dans cette démarche incluant de la vidéo, je voulais sentir que la voix de synthèse associée à un corps organique pouvait donner un décalage étranger (*the uncanny*). Pourtant, comme la voix de synthèse a besoin d'un support technologique (haut-parleur) pour se faire entendre, l'image de mon visage avait besoin d'un support pour être capturée et diffusée. J'ai voulu alors aller plus loin dans l'incarnation de cette voix et rendre le corps physique du spectateur moteur de la voix qu'il entend. Par un moteur de reconnaissance faciale, il serait à même de contrôler les émotions perçues dans la voix de synthèse et ainsi la rendre vivante par la simple présence de son corps et de ses mouvements. Cependant, je voulais encore que la voix de synthèse soit synchronisée à mon visage et que le décalage d'émotion ne soit rendu audible que par la présence du corps du spectateur, entendant les modifications sur la voix, mais ne voyant aucun changement sur mon visage.

Afin de réaliser cette commande par reconnaissance faciale, il fallait que j'ai : la voix de synthèse prononçant mon texte écrit pour elle, une vidéo de mon visage prononçant ce même texte en *play-back*, une webcam qui capturerait les expressions faciales de chaque visiteur (un par un). Pour comprendre que leur visage était à la commande la voix, je voulais

installer un système optique comprenant un miroir semi-aluminé[99] afin que leur reflet soit superposable à mon visage.

Voici quelques photographies de la mise en place de ces étapes de travail, qui ont permis d'aboutir à la scénographie finale de *THE VOICE IS VOICES*.

Je joins en *annexe 2* le synoptique de la première version de l'installation.

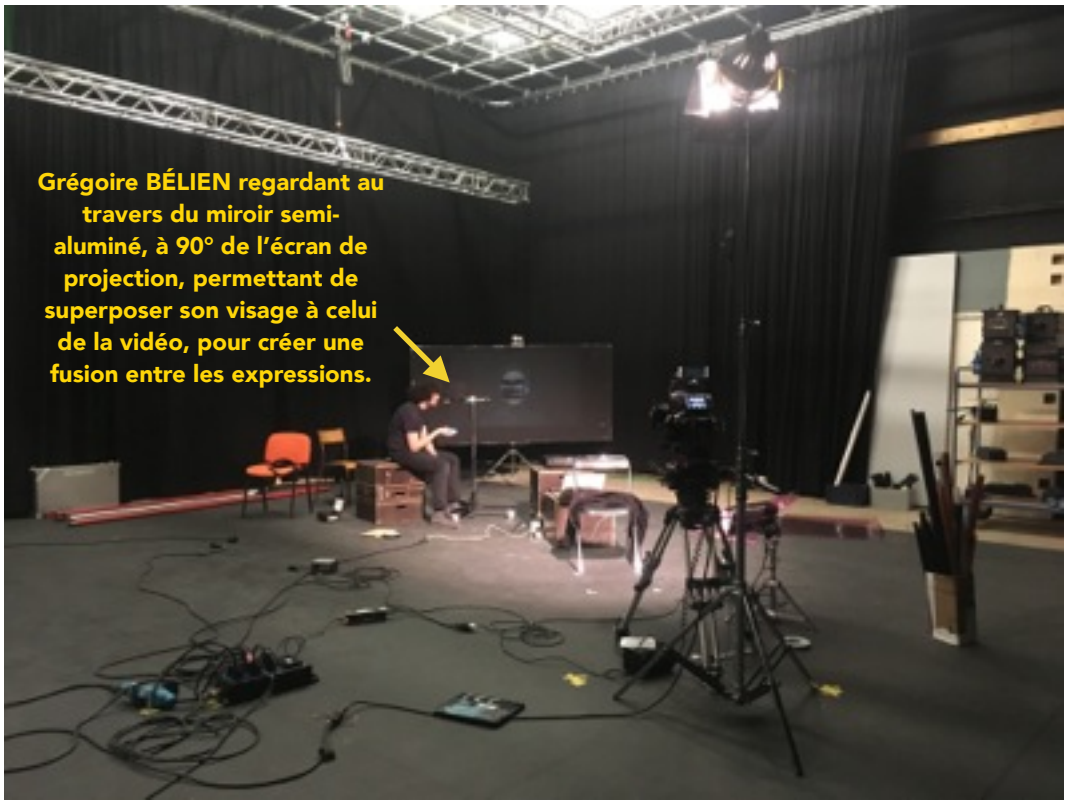


Figure 34 : tournage de la vidéo de *THE VOICE IS VOICES* - on peut voir un écran de projection noir au fond de la salle ainsi que le miroir semi-aluminé dans sa boîte, qui permettait de tester la bonne superposition du portrait filmé et celui du spectateur.

[99] un miroir semi-aluminé est un miroir qui reflète 50% de la lumière et qui en laisse passer les 50% restant. Placé à 45° d'une source visuelle, il permet sa visualisation « fantomatique » depuis un point de vue à 90° de cette source. Intégré dans un jeu optique, il permet donc des superpositions visuelles, créant une impression fantomatique.

Figure 35 : Dans la salle d'exposition, essais de superposition du visage avec celui du spectateur par le miroir semi-aluminé. La webcam qui filme les expressions du spectateur est fixée derrière le miroir, et est reliée à *FaceOSC*, un logiciel qui permet de récupérer le positionnement des différents points du visage.

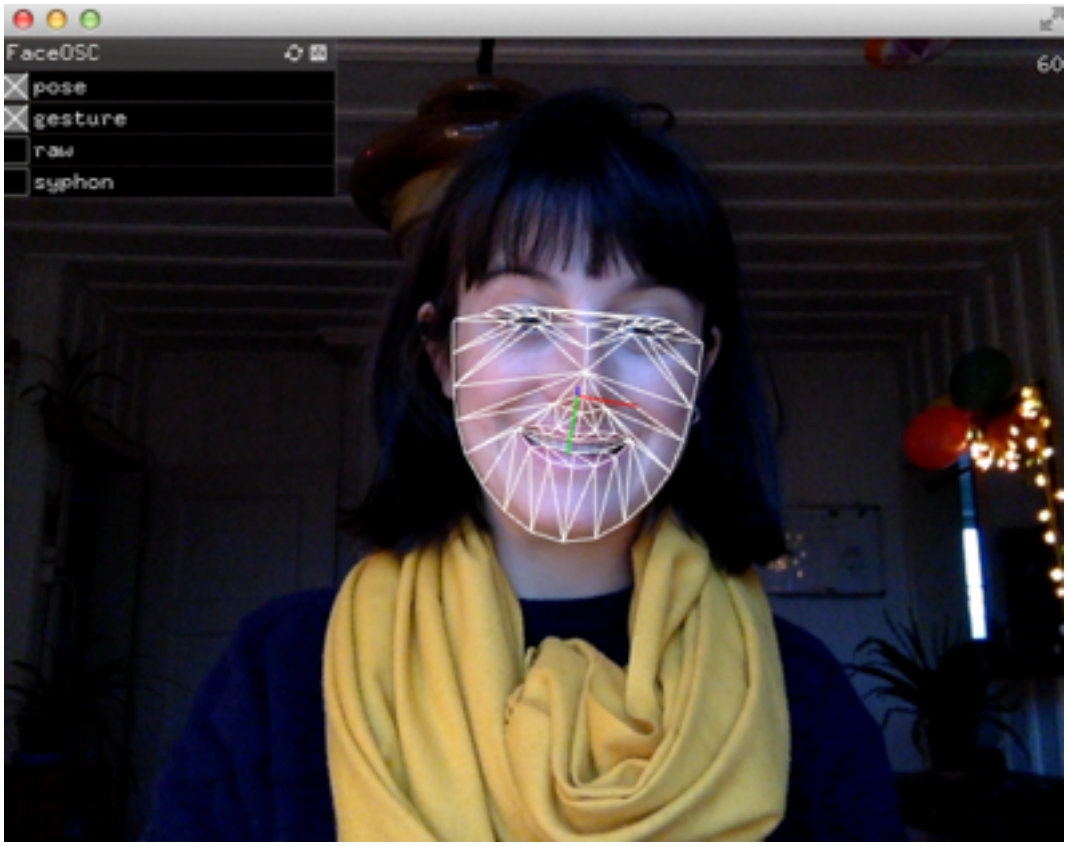


Figure 36 : capture d'écran de *FaceOSC* permettant de mesurer la position des différents points sur un visage afin d'envoyer des informations en *OpenSoundControl*[100] (récupérées dans *MaxMSP*). Ici, le positionnement de la bouche et des sourcils permettait de reconnaître trois émotions : joie, peur, tristesse. Ces points permettaient alors de commander les émotions entendues dans la voix, expressions créées par D.A.V.I.D. en amont (fonctionnant également sous *MaxMSP*).

[100] *Open Sound Control* est un format de transmission de données entre ordinateurs, synthétiseurs, robots ou tout autre matériel ou logiciel compatible, conçu pour le contrôle en temps réel.

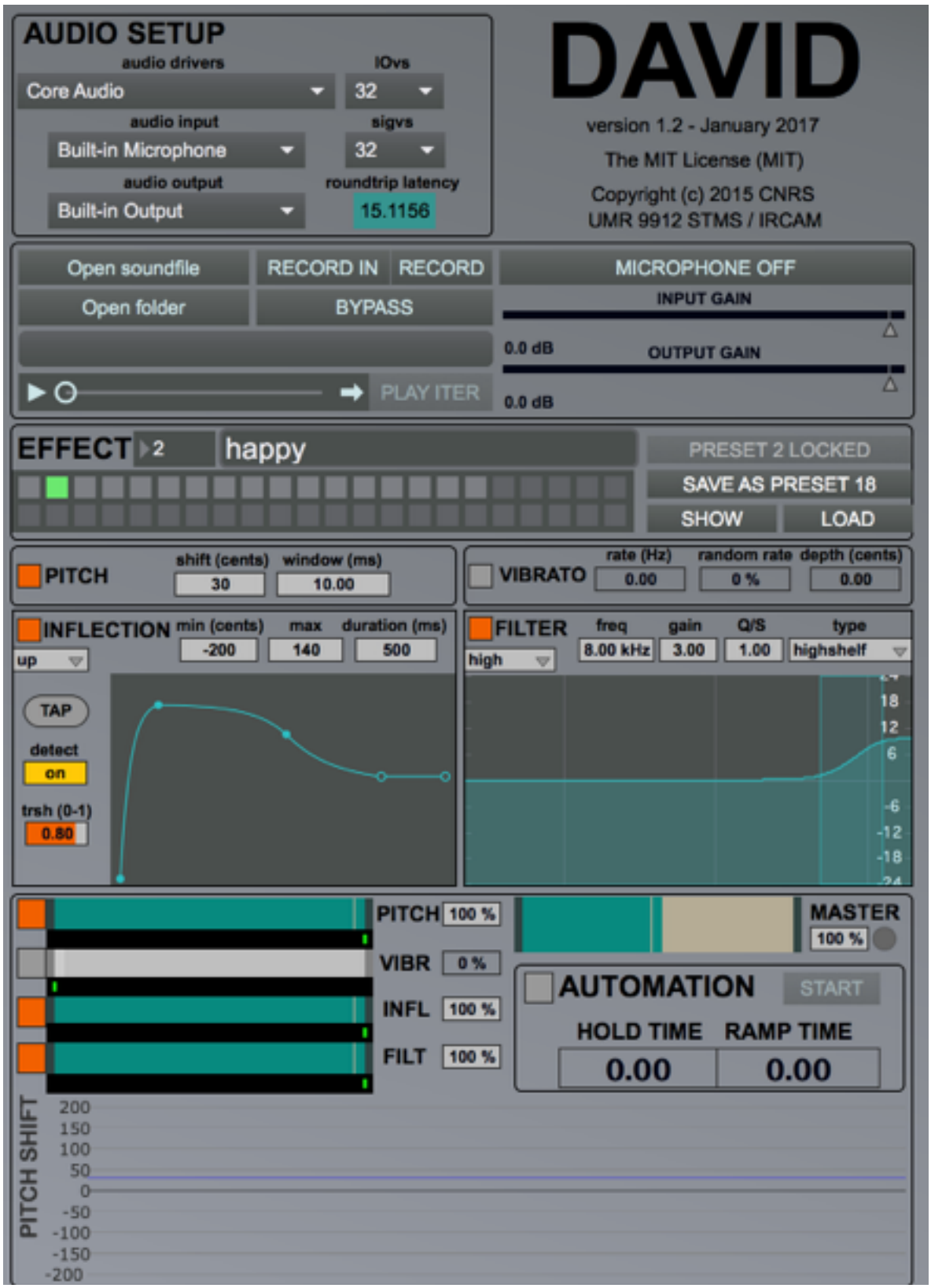


Figure 37 : capture d'écran de l'interface D.AV.I.D. (sous *MaxMSP*) avec l'application des paramètres par défaut de l'émotion « joie »

Cette corrélation entre l'expression du visage et l'émotion vocale entendue crée un lien physique entre le corps et la voix de synthèse. Le corps du spectateur incarne l'émotion de cette voix et sans lui, la voix reste neutre, inexpressive. La démarche d'incarnation de la voix de synthèse par un corps « organique » m'a cependant montré, en la réalisant, que le cœur de mon sujet était davantage porté sur la ressemblance entre la voix de synthèse et la voix organique. Les étapes de travail étant bousculées par le fait d'avoir le texte avant la voix, la voix avant l'image pour faire le *play-back* : il a fallu aller au bout de ce concept pour me rendre compte que la reconnaissance faciale rendait *THE VOICE IS VOICES* une démonstration technique et non une expérience sensible autour du clonage vocal. De plus, la reconnaissance faciale commandant la voix aurait dû commander un contenu vocal bien plus dense, plus linéaire, avec moins de pauses comme celles qui étaient écrites pour la voix de synthèse. En effet, les effets de « joie », « peur » et « tristesse » sont presque inaudibles et c'est dans les différences que l'effet sonore s'entend, il aurait fallu les exagérer et cela aurait davantage servi l'artificialité de la voix que son penchant expressif réaliste. J'ai dû ainsi repenser la scénographie de l'installation, en la rendant uniquement sonore. Mettant en avant le fait que la voix (de synthèse comme organique), s'incarnait dans les haut-parleurs, j'ai compris que ce corps technique qui incarnait la voix, pourtant si réaliste, permettait également une distorsion entre humain et machine. L'inquiétante étrangeté de la voix, *the uncanny valley of speech*, réside dans l'écoute simultanée d'une voix artificielle et de sa jumelle naturelle.

THE VOICE IS VOICES est devenue une installation sonore où il a fallu ré-écrire la scénographie pour mettre en avant ce jeu de doute entre la voix organique et son clone numérique. J'ai cependant gardé le visuel du suivi facial (comme sur le logiciel *FaceOSC* qui traque la position du visage) pour l'affiche (voir affiche sur la *Figure 29*), proposant un portrait dont nous essayons de cerner les traits. Si le public s'approche de l'affiche en taille réelle, il peut également apercevoir certains artefacts de la reconnaissance faciale, symbolisant les erreurs de la synthèse vocale, pourtant si proche de la voix d'origine.

d. Réalisation et réception

En *annexe 3* et *annexe 4*, le schéma de l'installation dans l'espace d'exposition, ainsi que le parcours le plus fréquent des spectateurs au sein de l'installation.

THE VOICE IS VOICES est donc devenue sonore et les haut-parleurs ont endossé le corps de la voix. Elle comporte 5 haut-parleurs, une boucle sonore de 2 min 33 sec et est accompagnée d'un jeu de lumières qui indique le parcours sonore que suivent les haut-parleurs. L'installation est pensée en deux parties : une première comportant le texte « La voix est pleine » (voir III/2.a.) et la seconde l'itération « Je ne sais pas » (voir III/2.a.). Dans toute l'installation, chaque haut-parleur incarne soit la voix organique, soit la voix artificielle. Le spectateur est invité à lire le carton de présentation, qui propose de se balader librement dans l'espace et suggère de prêter l'oreille aux différences entre la voix organique et son clone numérique.

« La voix est pleine » permet aux deux voix d'être entièrement jumelles : la voix de synthèse a le même timbre que la voix organique mais la voix organique cherche également à imiter la voix de synthèse dans sa prosodie. Dans la scénographie, ce sont les haut-parleurs diffusant « La voix est pleine » qui ouvrent l'espace d'exposition, ils se font face et permettent une écoute simultanée, où l'auditeur est amené à alterner entre l'un (voix artificielle) et l'autre (voix organique), parfaitement synchronisés. Il peut isoler les voix en s'approchant physiquement des haut-parleurs, quitte à coller son oreille à la membrane.

Les répétitions de « Je ne sais pas » mettent en avant la reproductibilité de la voix de synthèse qui n'a aucune variation dans sa prononciation alors que la voix organique va jusqu'aux pleurs, se désynchronisant totalement de son clone. Les deux haut-parleurs qui incarnent ces voix sont dos à dos, ne permettant pas une écoute simultanée des deux voix.

Ils marquent leur opposition mais également la « froideur » du « Je ne sais pas » artificiel qui reste inchangé par l'émotion de la voix organique.

Le dernier haut-parleur marque un point de symétrie dans l'espace. Il est à la pointe de ces deux triangles (voir schéma) et a pour rôle ce point de pivot entre l'artificiel et l'organique. Il diffuse les bruits de bouche de la voix organique, synchronisés à celle-ci. À première écoute, ces bruits peuvent s'apparenter à des artefacts numériques, des erreurs de numérisation ou bien être la preuve d'un haut-parleur défectueux. Pourtant, ces bruits de bouche sont la preuve que notre voix est incarnée par un corps, ils font entendre la salive, les lèvres, la bouche, la respiration. Obtenus en les isolant de la voix organique, ils s'écoutent de manière synchronisés à celle-ci et peuvent alors indiquer qu'ils appartiennent à la voix organique.

Figure 38 : Les deux haut-parleurs incarnant les « Je ne sais pas » et le dernier, diffusant les bruits de bouche synchronisés à la voix organique (lumière générale, espace d'exposition)

Les projecteurs qui permettent une lecture de l'espace mettent en avant ce dernier haut-parleur, à la fois avec « La voix est pleine » et les « Je ne sais pas ». Entre les deux boucles, une respiration vient éteindre les lumières ; cette respiration est à la fois organique mais est aussi synchronisée aux respirations artificielles (c'est-à-dire tirées des silences de la voix de synthèse). Les enceintes mettent en avant *l'objet* qui contient la voix. Cette dernière demeure acousmatique, concentrant l'écoute sur les résidus du corps présents dans ses bruits de bouche, respirations et articulations. La lumière synchronisée aux différentes parties de l'installation permet toutefois une personnification du corps hypothétique de la voix par le haut-parleur, en soulignant les enceintes qui « parlent ».

Figure 39 : Spectateur écoutant l'un des haut-parleurs dans *THE VOICE IS VOICES* (lumière d'exposition)

Le mixage *in situ* m'a permis de rendre plus perméable l'espace entre les différents haut-parleurs. Ainsi, en jouant une réverbération de « pièce » sur les haut-parleurs qui ne diffusaient pas la source directe (et étant donc dans l'ombre), j'ai essayé de faire entendre un léger écho de la voix. Cela permettait aussi de rendre l'espace sonore plus homogène, sans pour autant ajouter une réverbération sur le son direct, car il m'était inconcevable d'ajouter de la réverbération sur une voix de synthèse, qui est par nature sèche, dépendante de la réverbération du corpus de voix à partir duquel elle est créée.

À la suite du parcours dans l'installation, j'ai proposé aux spectateurs de répondre à un retour d'expérience afin de comprendre leur choix de déambulation et leur sentiment au sein de l'installation. Dans le public qui est venu visiter l'installation, 51 ont répondu au questionnaire (à lire en *annexe 6)*.

J'ai fait le choix de demander au public de lire mon carton de présentation (voir *annexe 5*) afin que leur écoute soit influencée par la recherche de cette voix artificielle. Dans la première partie « La voix est pleine », comme la voix organique imite la voix de synthèse, la différence entre les deux voix permet un jeu d'écoute entre les deux haut-parleurs, oscillant tel une hésitation. Le fait de savoir que l'une des voix entendue est artificielle a également emmené quelques spectateurs sur la piste que la voix la plus réaliste, notamment celle qui exprime différents « je ne sais pas » était donc la voix artificielle. Mais finalement, de savoir laquelle est laquelle importe peu, tant que dans l'écoute, nous avons eu ce mouvement d'aller-retour, d'hésitation et d'écoute critique. Le parcours demande donc une attention particulière portée sur les éléments qui pourraient permettre de les différencier. Certains spectateurs ont d'ailleurs eu du mal à « discerner la vraie voix ».

La difficulté de la question « avez-vous distingué deux voix différentes dans l'installation ? » est que j'avais peine à parler de ma « vraie » voix et de « son clone », sachant qu'elles représentent la même signature vocale mais, comme elles sont entendues de manière simultanée et parfois désynchronisées, nous pouvons percevoir la question comme « avez-vous entendu deux voix simultanément ? » ou « avez-vous entendu deux signatures vocales différentes ? ». Ainsi, les réponses à ces questions sont ambigües, car ne répondant pas toujours à la même question. Certains visiteurs sont allés jusqu'à entendre plus de deux voix différentes, ce qui montre bien que ma question n'est pas claire. *A posteriori*, j'aurais aimé poser la question : « avez-vous douté ? » car c'est la remarque qui m'est le plus revenue dans mes entretiens à l'oral avec le public. Le fait « de ne pas savoir laquelle est laquelle » a été le principal commentaire de *THE VOICE IS VOICES*.

J'ai demandé le nom et prénom des participants afin de pouvoir savoir si le spectateur connaissait ma voix ou non. Et en effet, j'ai pu constaté qu'une personne qui m'était inconnue avait beaucoup plus de peine à distinguer les deux voix que mes amis ou même ma mère, pour qui il a été très facile de faire la distinction entre la voix organique et son clone. La connaissance de la voix influence donc sur la perception du clone et les artefacts qui peuvent passer inaperçus pour d'autres spectateurs, paraissent flagrants pour un public qui a déjà une connaissance de cette voix en particulier. En effet, 27 participants sur 51, soit presque 53 % du public ont répondu qu'ils n'auraient pas remarqué une voix de synthèse dans l'installation s'ils ne l'avaient pas su avant. Sur les 8 participants ayant répondu qu'ils auraient perçu la présence d'une voix de synthèse (16 % du public), 6 personnes font partie de mon entourage proche et ont donc une bonne connaissance de ma voix.

Ce qui a permis à la plupart du public de distinguer la voix de synthèse était un artefact laissé sur la phrase « la voix est neutre ». La voix organique était quant à elle davantage perçue sur l'impression d'émotion

qui s'en dégageait, même si elle cherchait à être d'expressivité neutre. Les adjectifs choisis par le public pour la définir ont été : réelle, intrigante, troublante, claire, émotionnelle, autoritaire, étrange, enthousiaste, jeune, froide, bluffante, étrangère, métallique, émouvante, féminine, neutre, hésitante, véritable, robotique, hypnotisante, solennelle, troublante, déterminée, convaincante, pleine, perdue, déshumanisée, perturbante, humaine, obsessionnelle. Le choix de ce vocabulaire correspond bien à la double identité de cette voix, à la fois réelle mais ayant une prosodie synthétique.

22 personnes sur 51 du public (43 %) ont répondu qu'elles n'auraient pas du tout imaginé l'installation en lumière naturelle, expliquant que le choix de l'obscurité apportait une concentration supplémentaire (et nécessaire) à l'écoute. 49 % ont répondu « oui » ou « peut-être » pour voir l'installation en lumière naturelle ; c'est sous ce dernier choix scénographique que l'installation sonore sera présentée à la ZHdK en juin 2019.

Le retour du public quant au cinquième haut-parleur (contenant les bruits de bouche), qui était le moins évident dans la compréhension de la scénographie, a d'ailleurs suscité de nombreuses réactions, soit enthousiastes soit négatives. Pour certains, il brouillait l'écoute et la possible distinction entre les deux voix. Pour d'autres, il servait de point de bascule entre les deux voix, et orientait quant au contenu artificiel de l'une d'elle.

Suite aux retours du public, pour la suite de *THE VOICE IS VOICES*, j'aimerais apporter davantage de contenu à la voix, jouer sur des phrases qui présentent plus d'artefacts, permettre une écoute enceinte par enceinte (tel un écho entre la voix organique et la voix artificielle), et pourquoi pas amener un contenu plus personnel, comprenant une narration plus linéaire et moins conceptuelle. Les remarques quant au contenu du texte ont fait référence à des poèmes surréalistes, à des auteurs de l'absurde, qui sont à la fois mes références mais également

quelque chose que je reproche aussi à certaines installations qui, par un contenu trop conceptuel, ne sont accessibles qu'à une élite qui possède la référence. À la manière des installations sonores vocales de Dominique PETITGAND[101], j'aimerais utiliser l'espace de diffusion dans l'écriture du contenu et maintenir une certaine vie dans le choix des coupes des voix, comme pour entretenir un suspend auditif, promesse d'une suite qui pourrait arriver. J'aimerais montrer l'installation dans un espace plus large, permettant une déambulation plus grande entre les haut-parleurs afin que le parcours ne soit pas aussi ramassé entre les deux parties et que l'isolation par enceinte soit davantage dépendante du rapprochement physique du spectateur vers le haut-parleur.

Figure 40 : spectateur déambulant dans
THE VOICE IS VOICES (avril 2019)

[101] PETITGAND Dominique, artiste sonore qui a notamment présenté *Les Heures creuses*, installation sonore pour théâtre vide au Théâtre de Gennevilliers (2018).

Conclusion

La voix est notre lien intime au social et notre signature vocale fait partie intégrante de notre identité. Depuis le XVIIIème siècle, l'humain a cherché à reproduire le mécanisme de la parole. Aujourd'hui, la synthèse vocale permet de créer un clone de sa propre voix, c'est-à-dire une voix artificielle portant notre timbre, singulier. Nous avons vu que les outils de synthèse accessibles aux particuliers (études de *Lyrebird AI* en anglais et de *CandyVoice* en français) n'ont cependant pas encore un résultat crédible. La création d'un clone vocal fidèle au timbre d'origine reste réservé à un contexte de recherche ou bien de commercialisation. Même si le clonage vocal présente une solution pour les patients atteints de *SLA* (Sclérose latérale amyotrophique), nous pouvons imaginer que cette technologie pourra avoir un impact sur les métiers liés à la performance vocale (doublage, lecture de livres audio ou travail des post-synchronisations au cinéma), où le clone devra être soumis à d'éventuels droits d'auteur. Dans la mesure où cette synthèse vocale personnalisée se démocratise, elle pourra présenter le risque d'être détournée pour « faire dire n'importe quoi à n'importe qui ». La protection des données vocales repose aujourd'hui sur la responsabilité des entreprises qui proposent ce service.

À travers l'installation sonore *THE VOICE IS VOICES*, j'ai voulu créer ce sentiment de « doute » entre une voix organique et son double artificiel. Le clone de la voix a été réalisé avec l'aide du programme de synthèse vocale de l'IRCAM, *ircamTTS*. Par la suite, la ressemblance entre les deux voix a été retravaillée par un processus d'imitation, brouillant les différences et posant la question de la singularité de l'identité vocale. La scénographie permet d'écouter la voix organique synchronisée à sa jumelle synthétique. Construite comme un portrait sonore, l'installation met en scène *the uncanny*, cette hésitation entre deux voix étrangement proches, qui questionnent la crédulité de notre oreille. Cependant, lors de l'installation, le spectateur connait la nature artificielle de l'une des

deux voix : cette information définit la posture d'écoute et influence l'attention que nous portons aux caractéristiques humaines qui se dégagent de la parole, plaçant notre oreille dans une position critique.

Si aujourd'hui un algorithme est capable de reproduire notre voix, créant à l'avenir une impossible distinction entre une voix humaine et une voix artificielle, le rôle de l'humain est peut-être de protéger ce bien si précieux et de faire confiance aux relations sociales de « vive voix », seuls liens qui peuvent encore nous permettre de croire en la voix.

Annexes

annexe 1 : *Extrait de « Quelques heures sur le bout de ma langue », travail de transcription orale pour la Jahrespublikation (publication annuelle) de la Zürcher Hochschule der Künst, ZHdK (Zurich, juin 2019)*

Mélia 51 Roger

XX
XXXXXXX

Quelques heures sur le bout de ma langue

Codes between writing process and real time talking:

Jumping one line = 1 min
Adding one space between words = 1 second
The size of words depends of the loudness of the voice
All the sounds coming from the mouth and recorded are written

Sept heures et cinquante-trois minutes, chambre en bois
T(ccc)est enregistrement binaural… Left. Right. (ccccccccccc)
Bon. (shhhhhhh) J'vais les accrocher comme ça… J'vais LOCK.

Sept heures et cinquante-quatre minutes, chambre en bois
Et ensuite, j'vais pouvoir me balader avec. eeeeeeeeeee

Sept heures et cinquante-cinq minutes, salle de bains, fenêtre ouverte
'Tain j'ai un ch'veu blanc.

Sept heures et cinquante-six minutes, salle de bains, fenêtre ouverte
Merde, c'est pas l'bon. Voilà. ca c'est le stress hein.

Sept heures et cinquante-huit minutes, cuisine
Va falloir que je décide si j'me censure ou pas…?

Huit heures et sept minutes, cuisine
Hmmmmmmmm. La carte d'identité. hhhhhhhhhhhhah que bon ! Allez, j'vais faire du café…
Ah putain on n'a plus de café ! haaaaaaan ? Ah si, elle est trop forte Klara, elle en a racheté.
Bust coffee !
Ok faut qu'j'trouve. Alors. J'vais mettre l'enregistreur dans ma poche…? Merde j'ai pas d'poches.
Ok j'vais l'coincer là…? Comme ça… Fermé… 'Tain j'ai l'air bien débile hein.

Huit heures et huit minutes, cuisine
(shhhhhh ccccchhh) Oh putain j'ai tellement pas d'force !
Huit heures et dix minutes, cuisine
Ça sent - trop bon ! Hmmmmmaaaaammmm
La cafetière… La cafetière, la cafetière… Elle est où ? Bah ! Trop bizarre ! Arrrrhhhaaaahhh !
Elle est là.

Huit heures et onze minutes, salle de bain
hhhhhhhhhh J'ai encore un autre cheveu blanc ? Ça va pas du tout !

Huit heures et vingt-deux minutes, chambre
Bon quel temps y fait ? Ahhh oui, un peu froid…?
hhhhhhhhhhh, faut qu'fasse une machine. OK.
Hum?

Huit heures et treize minutes, chambre
Faut des chaussettes. Hummm.
Nan mais en vrai c'est débile d'enregistrer tout ça ! J'suis pas un malemphmmmm !

Huit heures et quinze minutes, salle de bain
P'tet' qu'y faudrait que j'décrive tout c'que j'fais…? Alors éééééééh, là, voyez-vous… on met… on met les habits…éééééééééh, pour aller à l'ecoleeeeeéééé.

Là en particulier je mets les chaussettes. Voilà ! Ce sont plutôt des socquettes, blanches, avec uneuh petiteuh ligneuh bleue, je vous voue que j'avais les mêmeuh hier, voilà, c'est comme ça, tout l'monde le sait, y'a plus de mystère. Alors. Le petit T-shirt que ma offert ma mamhin, il y a de ça déjà… quatre hinn… cinq hinn… m'en souviens pô. Déodorant à la rôseh, poureh suentireh la rôseh ! Vooilàáá

Huit heures et vingt-et-une minutes, salle de bain
Eh bah les pot', les poches pour meufs, c'est, c'est vraiment… vraiment pas grand quoi. LES POCHES POUR MEUFS C'EST VRAIMENT PAS GRAND. Comme ça, on, on fait tout mettre dans le sac à main…
No shame, haha.
Non mais, ça s'v, ça voit que j'porte des micros quand même…? Nan ça va. T'façon j'sais pas si j'ai envie que ce soit discret ou si j'ai envie que les gens l'remarquent. C'est, c'est fou comme le fait d'avoir des micros sur soi, ça change notre manière de parler bon bah sinon le matin, là, il est quoi huit heures vingt ? J'parle jamais ! Ouala l'café ça doit être prêt !

Huit heures et vingt-deux minutes, cuisine
Ah bah juste à temps !
Merdeuh putain, chaque fois ! Alors… Ah ouais, la tasse… Mi-tropa ! 'Tain faut qu'on finisse le champagne. J'vais prendre du lait de soja… à la vanille. J'pense qu'à midi j'mangerai ici ? Allez, un p'tit verre d'eau ? (shhhhhhh) OK. Prendre un bol. Faut que j'finisse la tarte aux pralines aussi. J'voulais que Gabriel la goûte mais… bon s'est pas r'vus. voilà. Et ma petiteee toucheee personnelleeeee… un tout petit peu… de cacao. voilà… commeuh ça….
hmmmm.. Il nous reste les figues aussi. J'ai vraiment l'impression de parler toute seule comme une grosse débile… Allez, on va jouer ça plus réaliste, normalement je ne parle pas le matin ! Voilà…
hhhhoui

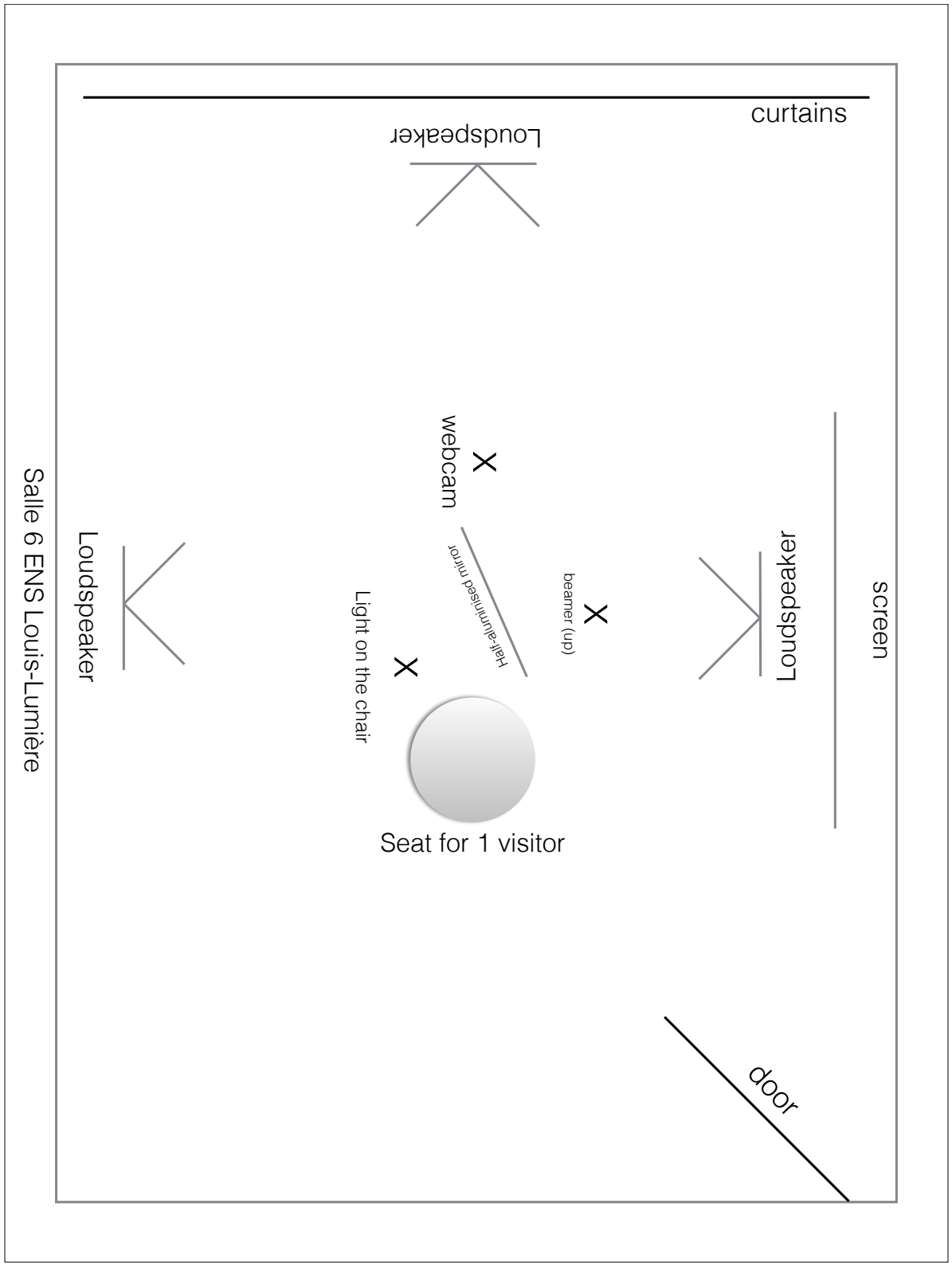
Huit heures et trente-et-une minutes, cuisine
Ahhhh moi qui déteste les bruits de bouche, j'vais être servie hein !

Huit heures et trente quatre minutes, cuisine
Ah, y'a déjà un citron d'ouvert.

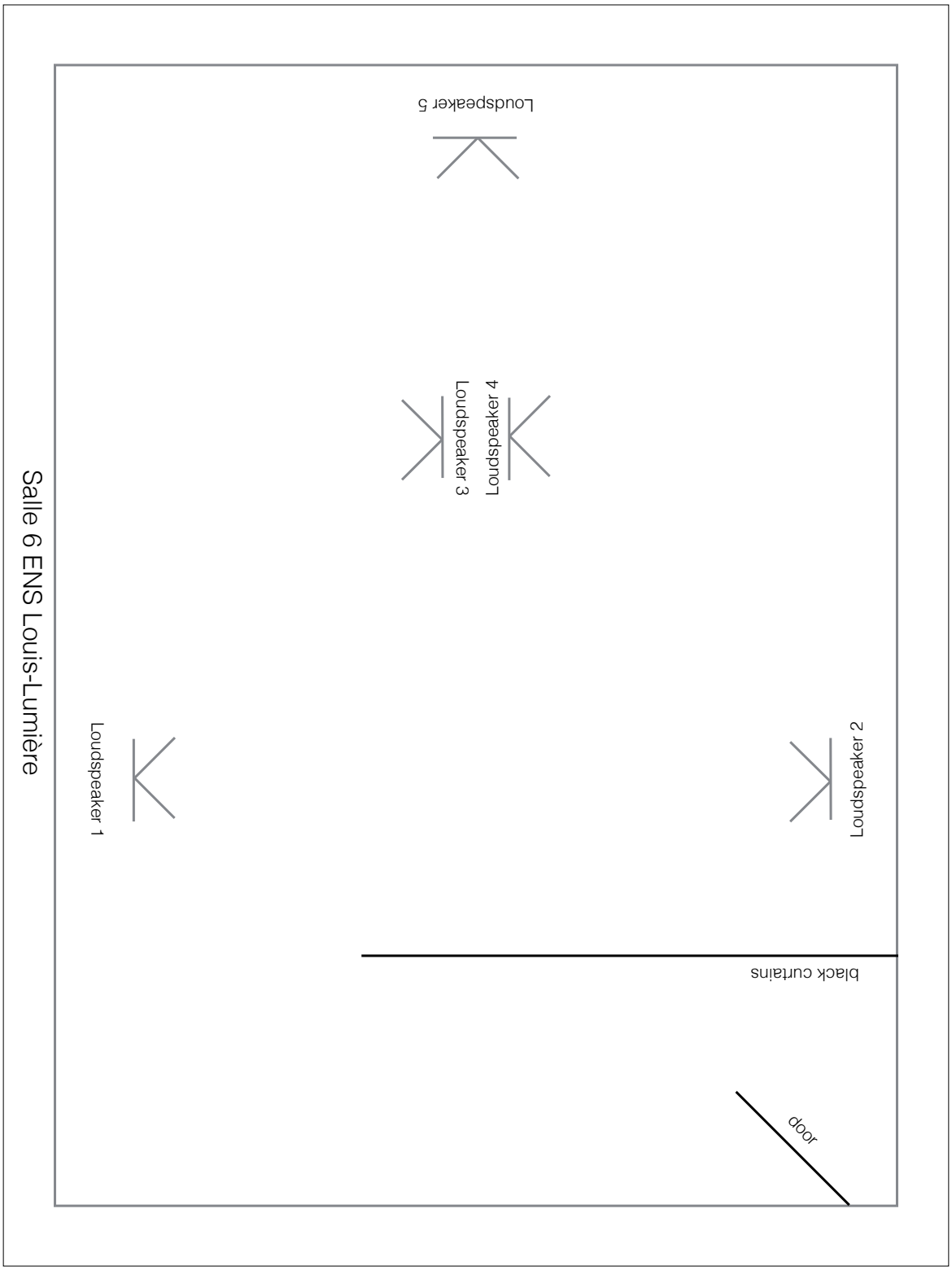
Huit heures et quarante minutes, chambre
Oh ! la belle araignée ! Bmmmmm, j'vais ramener une tasse à l'école, pour prendre les cafés. Laquelle j'peux prendre ? une tasse un peu moche…? Pmmmmr, elles sont toutes trop belles nos tasses… Celle là, la bleue là, ouais… mais c'est lourd ! Allez, c'est pas grave.
Hannnnn trop mignon ! Rohhhh, c'est trop choupi. petite plante… Elles font toutes des bébés.

Huit heures et quarante deux minutes, cuisine
N't.
(shhhhhhh)

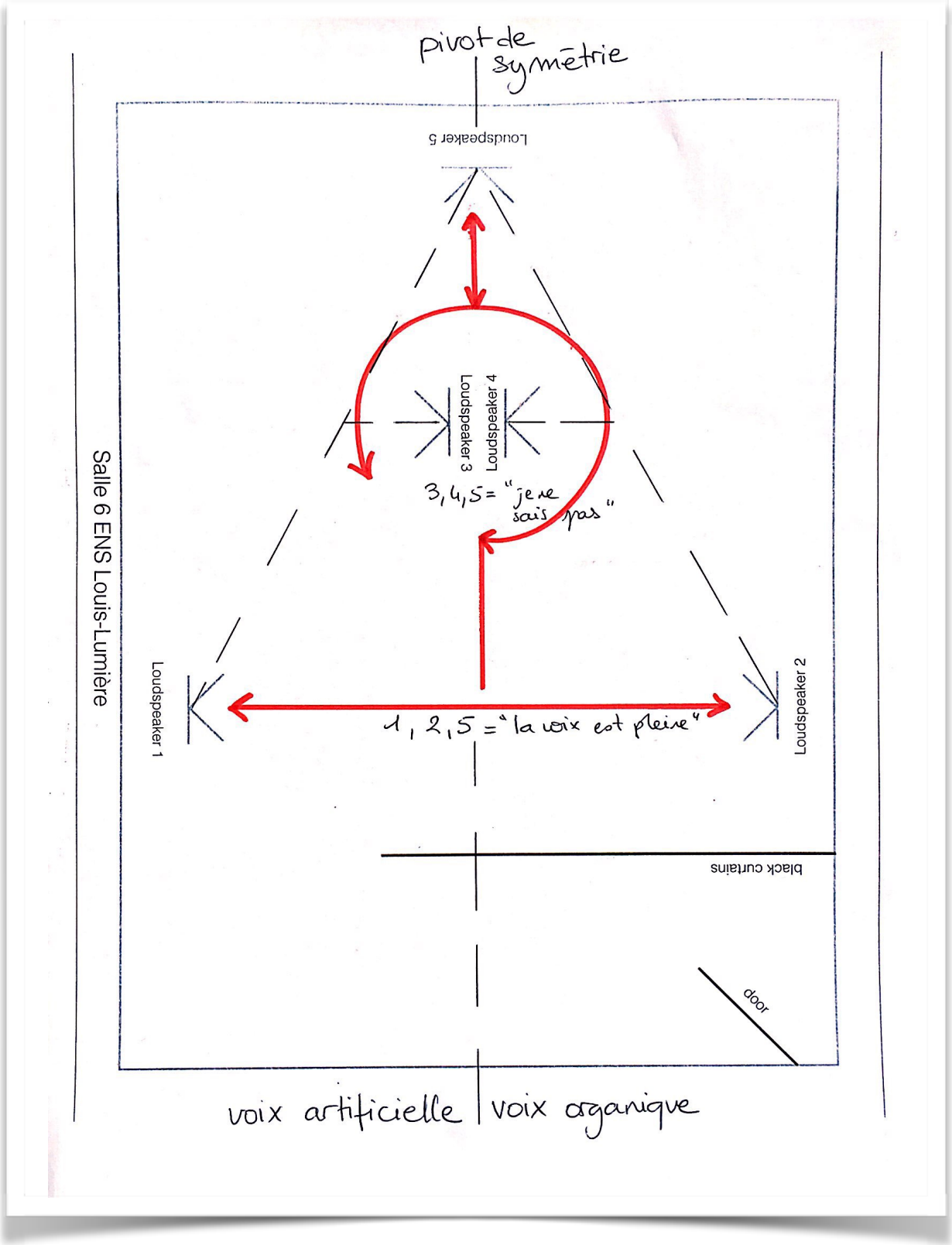
annexe 2 : schéma de la première version de *THE VOICE IS VOICES*, comprenant de l'image



annexe 3 : schéma de la seconde version de *THE VOICE IS VOICES*, installation uniquement sonore



annexe 4 : schéma de la seconde version de *THE VOICE IS VOICES*, installation uniquement sonore, avec (en rouge), le parcours de déambulation du public.



annexe 5 : carton de présentation à lire avant d'entrer dans l'installation sonore *THE VOICE IS VOICES*

THE VOICE IS VOICES

installation sonore autour du clonage vocal

Vous entrez dans un espace à deux voix : l'une organique et l'autre, son clone numérique. Cette voix de synthèse a été réalisée à partir d'un corpus vocal de plusieurs heures d'enregistrements qui ont permis de générer des phrases en synthèse "text-to-speech". Ainsi chaque phrase artificielle générée a été écrite, puis lue par une machine.

Circulez librement dans tout l'espace sonore et prêtez l'oreille pour discerner la vraie voix de son "fake".

Crédits

Conception et voix - Mélia ROGER

Image - Grégoire BÉLIEN

Réalisateur en informatique musicale - Adrien ZANNI

Direction artistique - Salomé OYALLON

Voix artificielle générée en collaboration avec l'Ircam

Direction de mémoire de fin d'études - Alan BLUM (ENS Louis-Lumière), Hanna Järveläinen (ICST, ZHdK)

annexe 6 : retour d'expérience proposé au public à la sortie de l'installation

THE VOICE IS VOICES

installation sonore autour du clonage vocal

retour d'expérience

Nom, Prénom :

Avez-vous distingué deux voix différentes dans l'installation ?

Si oui, quels sont les éléments qui vous ont permis de les distinguer ?

Quel adjectif utiliseriez-vous pour qualifier l'identité vocale que vous avez entendue ?

Vous-êtes vous beaucoup déplacé.e au sein de l'espace ? Si oui/non, pourquoi ?

Vous-êtes vous approché.e des haut-parleurs ? Si oui/non, pourquoi ?

Que vous évoque le texte prononcé par la voix ?

Parmi les phrases prononcées, la ou lesquelles retenez-vous ?

Que vous évoque le concept de voix de synthèse, artificielle ?

Selon vous, quel est le rôle du haut-parleur du fond ?

Auriez-vous vu cette installation sonore en lumière naturelle ?

Auriez-vous compris qu'il y avait une voix de synthèse si vous ne l'aviez pas su en entrant dans l'espace ?

Autres remarques :

Liste des figures

Bibliographie

ADOBE Creative Cloud, *Adobe VOCO,* présentation à *Adobe MAX 2016 Sneak Peeks,* 4 novembre 2016, URL : https://www.youtube.com/watch?v=I3l4XLZ59iw

ARNOLD, Aron, *La voix genrée, entre idéologies et pratiques – Une étude sociophonétique*. Linguistique. Université Sorbonne Paris Cité, 2015.

ALS association, *Revoice,* site officiel du projet, URL consulté en février 2019 : https://www.projectrevoice.org/

BARTHES, Roland, *Le grain de la voix*, entretien *De la parole à l'écriture*, La Quainzaine littéraire 1er-15 mars 1974, Essais, 1981.

BARTHES, Roland, *Le Bruissement de la langue*, Ed. Seuil, Paris,1984.

BERGEN Mark, article "*Google's Duplex AI Robot Will Warn That Calls Are Recorded*", *Bloomberg,* 18 mai 2018, URL : https://www.bloomberg.com/news/articles/2018-05-18/google-s-duplex-ai-robot-will-warn-that-calls-are-recorded

BOVE, Rémi, *Analyse syntaxique automatique de l'oral : étude des disfluences.* Aix: Université d'Aix-Marseille, (2008).

BRODEUR David, *Interaction humain-robot par la voix avec traitement des émotions du locuteur*, Mémoire de maîtrise sous la direction de MICHAUD François, 2016.

CALLIOPE, TUBACH J., *La parole et son traitement automatique* - Collection technique et scientifique des télécommunications (ENST), 1989, Paris: Masson.

CANDYVOICE, site officiel, URL consulté en février 2019 : https://webapp.candyvoice.com/#/connect

CHION, Michel, *La Voix au cinéma,* Ed. Les Cahiers du cinéma, Paris, 1982.

CNIL, *« La CNIL autorise l'expérimentation de dispositifs biométriques de reconnaissance vocale par des établissements bancaires »*, 23 mai 2017, URL site officiel :
https://www.cnil.fr/fr/la-cnil-autorise-lexperimentation-de-dispositifs-biometriques-de-reconnaissance-vocale-par-des

COPENHAGEN PRIDE, VIRTUE, EQUAL AI, KOALITION INTERACTIVE & THIRTYSOUNDSGOOD, « *Q, the genderless voice* », la voix de synthèse non genrée, site officiel, URL consulté en mai 2019 :
https://www.genderlessvoice.com

CREAM, équipe de l'IRCAM, site officiel, URL consulté en février 2019 : http://cream.ircam.fr/

DAVARYNEJAD M., SEDGHI S., BAHREPOUR M., WOOK AHN C., AKBARZADEH M., COELLO COELLO C. A., Article : *Detecting Hidden Information from Watermarked Signal using Granulation Based Fitness Approximation*, dans Applications of Soft Computing: From Theory to Praxis, Springer, Series: Advances in Intelligent and Soft Computing, Volume 58/2009.

D'ALESSANDRO, Christophe, TZOUKERMANN, Evelyne (sous la direction de), *Synthèse de la parole à partir du texte, numéro de Traitement Automatique des Langues (TAL)*, Hermès, Vol. 42, No 1, 2001.

DE BOYSSON-BARDIES Bénédicte, *Comment la parole vient à aux enfants*, Article, Les Cahiers du MURS, 1998.

DENIS, G.,*Transformation de l'identité d'une voix.* Rapport de stage DEA ATIAM, 2003.

DESPRATS, Pierre, *Recherche sur l'identité vocale dans la synthèse vocale et sa relation à la disfluence*, mémoire (sous le direction de Thierry Coduys et Greg Beller), Son, ENS Louis-Lumière, 2014.

DOLAR Mladen, *A Voice and nothing more*, Cambridge: MIT Press, 2006.

DUDLEY Homer W., *Bell System Technical Journal* 1940, p. 509, Fig.8 Schematic circuit of the *Voder*.

DT R&D, projet *« AI anti AI »*, 18 mai 2017, URL : https://rnd.dt.com.au/anti-ai-ai-a-wearable-ai-device-244900e4d71c

EDELIN, Thomas, *Signature : une expérience sociale et interactive autour de la voix humaine*, mémoire (sous la direction de Thierry Coduys et Nicolas Obin), Son, ENS Louis-Lumière, 2017.

ENCYCLOPÉDIE DE LA PAROLE, site officiel, *définition de timbre*, URL consulté en février 2019 : https://encyclopediedelaparole.org/fr/taxonomy/term/123#notice

FOUCAULT Michel, *Les aveux de la chair*, Histoire de la sexualité, parution posthume 2018, Ed. Gallimard.

FOUCAULT Michel, *Le corps utopique ; suivi de Les hétérotopies,* Paris, Éditions Lignes, 19 juin 2009, transcription radiophonique 1966.

HABERMANN, G. *Stimme und Sprache. Eine Einführung in ihre Psychologie und Hygiene*, Stuttgart, Georg Thieme Verlag, 1978.

HOREV Rani, *« Style-based GANs – Generating and Tuning Realistic Artificial Faces »*, 26 décembre 2018, article *LyrnAI* pour expliquer fabrication des images tirées du site *« this person doesn't exist »*, site officiel, URL : https://www.thispersondoesnotexist.com/

HUSSON, R. *La voix chantée*, Paris, Gauthier-Villars, 1960.

ICHER Bruno, « Les *Friends* perdent des voix », *Libération*, 30 août 2003, URL : https://www.liberation.fr/medias/2003/08/30/les-friends-perdent-des-voix_443347

JWT Amsterdam, *« Learn how to paint from the Master himself in The Rembrandt Tutorials »,* 28 février 2019, URL : https://jwt-amsterdam.pr.co/171584-learn-how-to-paint-from-the-master-himself-in-the-rembrandt-tutorials

LEOTHAUD Gilles, Théorie de la Phonation, Cours de DEUG 2ème année, 2004.

LORTIE Catherine, *Le vieillissement de la voix : De la production à l'évaluation*, Thèse sous la direction de TREMBLAY Pascale, GUITTON Matthieu, Université LAVAL, Canada, 2017.
LOVELUCK Benjamin, *La démocratie au prisme du numérique* (2017), publication sous la direction de TROUDE-CHASTENET, P., Edition, CLASSIQUES GARNIER.

LYREBIRD AI, *« Lyrebird, Inc. Agreement and Written Release Regarding Collection, Storage, and Disclosure of Biometric Data »*, site officiel, URL (consulté en mars 2019) :
https://about.lyrebird.ai/terms/biometrics

McGURK H., MACDONALD J. *Hearing lips and seeing voices*, Nature, vol. 264,1976.

NIEßNER Matthias, projet *Face2Face: Real-time Face Capture and Reenactment of RGB Videos*, juin 2016, URL :
http://niessnerlab.org/projects/thies2016face.html

PEELE Jordan, "You won't believe what Obama says in this vidéo! ;)" via *BuzzFeed*, avril 2018, URL :
https://www.youtube.com/watch?time_continue=26&v=cQ54GDm1eL0

PATEL Rupal, *VocalID*, site officiel, URL consulté en mars 2019 :
https://vocalid.ai/about-us/

RACHMAN, Laura, LIUNI Marco, ARIAS Pablo, LIND Andreas, JOHANSSON Petter, HALL Lars, RICHARDSON Daniel, WATANABE Katsumi, DUBAL Stéphanie, AUCOUTURIER Jean-Julien, *DAVID: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech*, publication 3 avril 2017.

FREUD Sigmund, *L'inquiétant familier (suivi de : "Le marchand de sable" de E.T.A. Hoffmann)*, Paris, Payot, coll. "Petite Bibliothèque Payot", 2012.

SCHRÖDER Marx, *Emotional Speech Synthesis: A Review*, DFKI Saarbrücken, Institute of Phonetics, University of the Saarland, 2001.

SHEN Jonathan, PANG Ruoming, WEISS Ron J., SCHUSTER Mike, JAITLY Navdeep, YANG Zongheng, CHEN Zhifeng, ZHANG Yu, WANG Yuxuan, SKERRY-RYAN RJ, SAUROUS Rif A., AGIOMYRGIANNAKIS Yannis and WU Yonghui, *Natural TTS Synthesis by conditioning Wavenet on mel spectrogram predictions*, Google, Inc., University of California, Berkeley, 2017.

SUARÈS A., *Remarques*, Paris, Gallimard, Les Cahiers De La Nrf, 2000.
SUWAJANAKORN Supasorn, SEITZ Steven M., and KEMELMACHER-SHLIZERMAN Ira, *Synthesizing Obama: Learning Lip Sync from Audio,* University of Washington, 2017.

TANK Nikita, *Voice User Interface (VUI) – A Definition*, Botsociety Blog, Avril 2018.

THAPEN Neil, *Pink Trombone,* site officiel, URL consulté en février 2019 : https://dood.al/pinktrombone/

Œuvres citées

CORBIAU Gérard, *Farinelli,* film 111 minutes, couleur (1994).

DEBOUZE Jamel, *Pourquoi j'ai pas mangé mon père*, film, 113 minutes, couleur (2015).

DE MAUPASSANT Guy, *Boule de Suif,* paru chez G. Charpentier (1880).

DE SAINT-EXUPÉRY Antoine, *Le Petit prince*, Ed. *Reynal and Hitchcock,* 93 pages (1943).

FAST Omer, *Looking Pretty for god, After G. W.)*, installation vidéo, *Manifesta 7* (2009).

FRAMPTON Hollis, *Critical Mass,* vidéo (durée non connue, 1971).

HARRIS Owen, *Black Mirror, S*aison 2, Episode 1 « *Be right back* », 50 minutes (2013).

JONZE Spike, *Her*, film 120 minutes, couleur (2013).

NAUMAN Bruce, *For Children,* installation sonore, *Fondation Cartier* (2015).

PARRENO Philippe, *Marylin*, installation vidéo (2012).

PETITGAND, Dominique, *Les Heures creuses*, installation sonore, *GB Agency*, création au Théâtre de Gennevilliers (2018).

THOMAS Thierry, *BARTHES, Le Théâtre du langage,* documentaire, 55 minutes, couleur (2015).

TRIBE Kerry, *Critical Mass,* performance pour deux acteurs, 25 min (2010).

WAN James, *Fast and Furious 7,* film 137 minutes, couleur *(2015).*

WEARING Gillian, *2 into 1*, vidéo 5 min, couleur (1997).

WEARING Gillian, *autoportrait*, photographie, couleur (2000).