**VOTRE CENTRE D'EXPERTISE EN TI**

Avec ses équipes d'experts et son important réseau, le CRIM offre aux entreprises québécoises une expertise scientifique en TI actualisée dans des domaines variés et complémentaires qui permettent un éventail d'applications dans différents secteurs. Au fil des années, le CRIM a poursuivi sans relâche son rôle de levier économique en développant des outils spécialisés, en livrant des technologies structurantes à ses clients et en diffusant de manière proactive les meilleures pratiques et les dernières innovations en TI.

# BIOMÉTRIE VOCALE

## VERS UNE IDENTIFICATION INCONTOURNABLE

**Le CRIM participe à des compétitions et campagnes d'évaluations technologiques en reconnaissance de la parole et du locuteur depuis 1992, et son expertise dans ce domaine est l'une des plus avancées au monde.**

## COUP D'ŒIL SUR L'ÉVOLUTION DES TECHNOLOGIES DE BIOMÉTRIE VOCALE AU CRIM

### Qu'est-ce que la biométrie vocale ?

La biométrie vocale est un domaine scientifique et technologique qui vise à développer des applications permettant de vérifier l'identité d'une personne seulement grâce à sa voix. Si la reconnaissance de la parole sert à déchiffrer « ce qui est dit » dans un enregistrement sonore, la reconnaissance du locuteur (ou biométrie vocale) cherche à savoir « qui l'a dit ».

Par ailleurs, le terme « biométrie » est quelque peu trompeur. En effet, la voix n'est pas une caractéristique physique aussi fixe et mesurable que les empreintes digitales ou les motifs de la rétine, par exemple. La voix est dynamique et peut varier selon le comportement, l'âge, la situation et l'état du locuteur.

C'est justement cette grande variabilité de la voix d'un individu qui pose les plus grands défis pour arriver à une technologie de biométrie vocale efficace. Personne ne prononce le même mot ou la même phrase exactement de la même façon chaque fois. De plus, la présence de bruit de fond et la qualité de l'appareil d'enregistrement peuvent rendre l'identification encore plus ardue.

La reconnaissance du locuteur est l'exemple par excellence d'une tâche que l'humain réalise sans grand effort (reconnaître la voix d'un ami peu importe les mots qu'il prononce), mais qui est extrêmement complexe à maîtriser pour une machine. Les développements de ces technologies sont donc directement liés à certaines avancées en intelligence artificielle, notamment en apprentissage automatique.

### Pourquoi la développer ?

Pourquoi continuer de tenter d'améliorer les techniques actuelles de reconnaissance par la voix, si les défis sont si grands et que la voix n'est même pas une mesure d'identification biométrique aussi fiable que les empreintes digitales ou rétiniennes ?

La réponse est simple : parce que la voix est l'outil principal dont se sert l'être humain pour communiquer. Il s'agit d'une manière naturelle de s'identifier. De plus, cette technique a l'avantage d'être non intrusive, elle peut s'effectuer à distance et ne nécessite pas de contact physique ni d'équipement spécialisé. Bien que les techniques physiologiques de biométrie telles que l'analyse des empreintes digitales ou des motifs de l'iris demeurent beaucoup plus précises, le taux de fiabilité de la biométrie vocale et sa simplicité d'implantation en font une mesure d'identification parfaitement acceptable dans de nombreux contextes d'application.

### Applications possibles de cette technologie

- Authentification : accès à un site web, à un compte-client, etc.
- Sécurisation des transactions.
- Réinitialisation d'un mot de passe, pour remplacer les « questions secrètes ».
- Preuve de vie : Dans certains pays, les personnes âgées ou invalides recevant une rente de l'état doivent se déplacer périodiquement jusqu'aux bureaux du gouvernement pour prouver qu'elles sont toujours en vie et continuer de recevoir leurs allocations. La biométrie vocale permettrait de fournir ladite « preuve de vie » sans se déplacer hors de chez soi.
- Surveillance/enquête: confirmer l'identité d'un suspect mis sous écoute.
- Structuration multimédia ou analyse audio et vidéo : classifier les divers intervenants dans une émission radio ou télé par leur voix, afin de retrouver du contenu facilement.

## Comment ça fonctionne ?

1. Lors de son inscription, l'individu fournit un échantillon de sa voix (échantillon d'inscription).
2. Puis, chaque fois qu'il désire s'identifier, l'individu doit fournir un échantillon test (soit un mot de passe précis ou simplement un échantillon de sa voix sans contenu spécifique).
3. Le système compare l'échantillon test à un modèle entraîné à partir d'information provenant d'échantillons vocaux de nombreux autres locuteurs. La machine ne compare donc pas directement l'échantillon test avec l'échantillon d'inscription de la personne qui tente de s'authentifier. Le système cherche plutôt à mesurer à quel point la voix du test ressemble plus au modèle de voix lié à l'utilisateur tentant de s'identifier qu'au modèle de référence du système (modèle universel).
4. De plus, les développeurs doivent déterminer le niveau de sévérité de leur système de biométrie vocale. C'est-à-dire qu'ils doivent choisir la « note de passage », le score de similarité à partir duquel la voix test sera considérée comme celle de l'individu inscrit. Ce niveau de sévérité dépend souvent du coût d'une erreur : il sera évidemment plus élevé pour autoriser des transactions bancaires que pour avoir accès à votre compte de bibliothèque!
5. Enfin, selon le score de similarité obtenu et le niveau de sévérité choisi, le logiciel fournit une réponse sous forme binaire : il accepte ou il rejette l'identification.

## Types d'identification vocale

### Vérification active ou passive

**Vérification active** : un échantillon test ou mot de passe est comparé à l'échantillon d'inscription et aux autres échantillons de voix que le système possède.

**Vérification passive** : cette technique cherche à confirmer l'identité d'un individu en analysant sa voix au fil d'une conversation, par exemple lors d'un appel à un centre de service à la clientèle. Ceci permet à l'intervenant de valider l'identité et même de personnaliser son service de manière non intrusive. Cette méthode a un très bon taux de fiabilité, car l'échantillon analysé est beaucoup plus long qu'un mot ou qu'une phrase. Par contre, elle ne peut être utilisée que dans certains contextes où en complément d'autres facteurs d'identification, car l'individu a accès au service demandé dès le début de son appel, même si son identité n'a pas encore été vérifiée par la biométrie vocale.

### Identification dépendante ou indépendante du texte

Les systèmes de biométrie vocale dépendants du texte exigent que l'usager prononce **un mot ou une phrase précise** afin d'être identifié. D'autres types de systèmes peuvent identifier l'usager par sa voix, **peu importe les paroles prononcées**.

Afin d'éviter qu'un fraudeur en possession d'un enregistrement de la voix de l'usager puisse se faire passer pour lui, certains systèmes donnent à l'usager **un mot de passe différent** à chaque tentative d'identification (par exemple une série de chiffres). Un fraudeur qui ne possède qu'un mot de passe enregistré ne pourra donc pas accéder au compte.

## Défis à venir

Évidemment, les défis sont nombreux pour la reconnaissance du locuteur, domaine toujours à la fine pointe des développements dans les domaines de l'intelligence artificielle et de la science des données. Pour l'identification vocale indépendante du texte, les enregistrements d'authentification très courts ne permettent pas encore un bon taux de succès, et les systèmes sont souvent moins aptes à identifier une personne qui s'exprime dans une autre langue que celle de son enregistrement, ou dans plusieurs langues à la fois.

Du côté des systèmes à mot de passe fixe, les problèmes sont multiples. D'un côté, l'accumulation de données, car chaque utilisateur doit enregistrer un mot de passe différent pour chaque service pour lequel il veut utiliser la biométrie, et se souvenir de toutes ces infos, ce qui n'est pas beaucoup plus efficace qu'un mot de passe tapé sur un clavier. De plus, si un fraudeur réussit à obtenir un enregistrement de l'usager en train de prononcer son mot de passe, il sera difficile pour la majorité des systèmes de déceler la différence. Par contre, des recherches en cours pourraient contribuer à enrayer ce problème, notamment grâce à la détection d'artefacts imperceptibles à l'oreille humaine qui permettraient de différencier une voix provenant directement d'un humain d'une bande enregistrée.

Bref, les experts en biométrie du CRIM et du monde entier ne risquent pas de manquer de travail durant les années à venir! Qui sait, peut-être un jour passerons-nous les douanes avec comme seule preuve d'identité requise… notre voix !

### TAUX D'ERREUR

Deux types de mauvaise identification sont possibles avec un système de reconnaissance du locuteur :

- Le logiciel rejette un échantillon test fourni par le bon individu : **faux négatif**
- Le logiciel accepte un échantillon test fourni par un imposteur : **faux positif**

Ces deux types d'erreurs sont complémentaires, c'est-à-dire que lorsqu'on réduit la fréquence des faux négatifs, on augmente celle des faux positifs, et vice versa. Si le seuil d'acceptation du logiciel est très haut, les échantillons auront beaucoup de chances d'être rejetés, donc de créer des faux négatifs. Par ailleurs, si le seuil est très bas, beaucoup d'échantillons seront acceptés, ce qui augmente le taux de faux positifs.

Les systèmes sont donc définis par une «zone de tolérance» qui varie selon le type d'application.

Dans le domaine financier, on préfère conserver un haut seuil, en prenant le risque de rejeter des individus légitimes, pour éviter à tout prix les fraudes.

D'un autre côté, dans le domaine de l'écoute électronique pour le domaine policier ou judiciaire, on préfère les seuils plus bas, ce qui permet d'identifier toutes les conversations où une voix similaire à celle du suspect est entendue. Dans ce cas, il est mieux d'avoir trop de contenu que pas assez.

## Le saviez-vous ?

Actuellement, le taux d'erreur des systèmes de biométrie vocale oscille entre 1 et 10 % selon le type de système. Il s'agit d'un taux d'erreur élevé lorsque l'identification vocale est la seule manière de sécuriser un accès. Les développeurs combinent donc souvent plusieurs systèmes différents afin de réduire la marge d'erreur : par exemple, la biométrie vocale ainsi qu'une autre technique biométrique telle que la reconnaissance faciale ou les empreintes digitales, ou encore la combinaison de l'identification vocale et d'un mot de passe secret. En combinant deux processus d'identification, le taux d'erreur chute drastiquement. Évidemment, plus on intègre de niveaux d'identification, plus le taux descend près du 0 % !

# UN PEU D'HISTOIRE

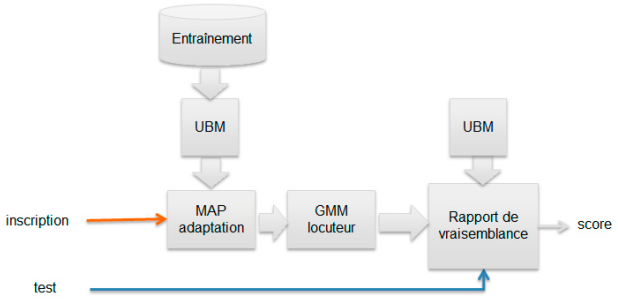
*Survol des principaux jalons ayant marqué la recherche en biométrie vocale*

## 1. Gaussian Mixture Model - Universal Background Model (GMM UBM)

La première méthode d'identification vocale avec un taux de succès intéressant se nomme la technique GMM-UBM. Il s'agit du premier système de biométrie vocale où le logiciel n'essaie pas simplement de comparer l'enregistrement d'inscription d'un usager et l'enregistrement d'authentification. En effet, le système GMM UBM s'entraîne selon une méthode bayesienne, en comparant une grande quantité de données audio entre elles, afin que le système puisse apprendre de lui-même l'étendue des variations de la voix humaine, selon la personne et la langue parlée, entre autres. Cela veut dire que les données qui servent à entraîner la machine ne sont pas les mêmes enregistrements audios qu'on lui demandera d'identifier, mais plutôt des échantillons de voix les plus divers possible, lui permettant ainsi de percevoir tout l'éventail des variations possibles de la voix humaine et d'en faire un modèle universel (Universal Background Model).

Ensuite, lorsqu'on on obtient un enregistrement d'inscription pour un individu, on se sert du modèle universel pour créer un modèle de voix pour cet individu qui prend en compte toutes les variations de cette voix spécifique (speaker-specific model).

Lorsque l'individu veut s'identifier, le système compare l'échantillon d'authentification avec le modèle de voix universel et avec le modèle spécifique de chaque utilisateur, afin d'émettre un coefficient de vraisemblance (likelihood ratio), c'est-à-dire la probabilité que l'échantillon d'authentification provienne de la même source que la voix modèle individuelle. C'est la comparaison entre la voix unique d'un individu et la voix «universelle» générée par l'analyse et la combinaison de grandes quantités de données qui détermine le score. Il revient ensuite aux opérateurs humains de déterminer à partir de quel seuil la machine peut confirmer l'identification.




Un des grands avantages de ce modèle statistique est qu'on peut sans cesse le mettre à jour et y intégrer une plus grande quantité de données, ce qui aide le système à devenir plus performant. La grande percée du GMM UBM fut l'utilisation des connaissances générales sur le son et ses variations afin d'entraîner la machine, au lieu de comparer directement deux échantillons de voix l'un avec l'autre, comme les systèmes précédents faisaient.

Avec le GMM UBM, l'identification vocale commence à être possible : par contre, avec un taux d'erreur de 15 %, il est presque impossible de lui trouver une application concrète dans notre monde.

## 2. Analyse factorielle jointe (JFA)

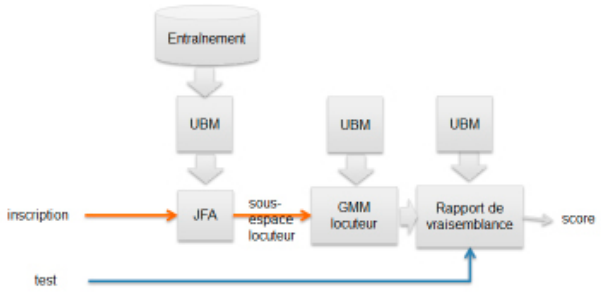
En 2004, Patrick Kenny, chercheur au CRIM, introduit une innovation : la Joint Factor Analysis (JFA), ou analyse factorielle jointe. Cette méthode fait chuter drastiquement les taux d'erreur : c'est une révolution dans le domaine.

La JFA est le résultat de deux réflexions principales. La première découle du fait que lorsqu'on développe un modèle de voix spécifique à un usager grâce au GMM UBM, le système se sert de l'enregistrement de la voix de la personne afin de modifier certaines parties du modèle universel. Par ailleurs, le système peut seulement modifier les parties du modèle pour lesquelles l'enregistrement individuel fournit de l'information directement.

Par exemple, dans ce modèle, il est impossible de présupposer comment un individu prononcerait un son ou une syllabe, s'il ne le fait pas durant l'enregistrement. L'intuition novatrice de Patrick Kenny fut de se demander si toutes les caractéristiques de la voix humaine sont réellement indépendantes l'une de l'autre, ou s'il n'existerait pas plutôt des facteurs sous-jacents, plus profonds, qui nous permettraient d'observer les liens qui existent entre certaines d'entre elles. Serait-il alors possible de déduire, ou à tout le moins d'estimer la présence de certaines caractéristiques absentes de l'enregistrement individuel, mais qui permettraient de raffiner grandement le modèle de voix spécifique? Cette méthode permettrait de modifier le modèle universel entièrement pour l'adapter à une voix individuelle, pas seulement d'en changer les caractéristiques présentes dans l'enregistrement. On pourrait estimer comment chaque individu prononce chaque lettre et phonème, même celles qu'il ne prononce jamais durant l'enregistrement servant à créer le modèle. Évidemment, si ce modèle fonctionnait, le taux de fiabilité serait probablement grandement amélioré.

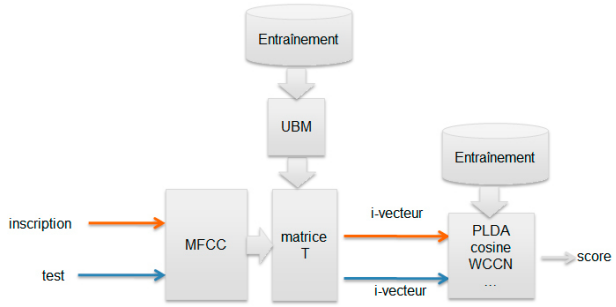
Il se peut bien que certaines variations de la voix soient liées les unes avec les autres. Par exemple, si on découvre qu'entre un homme et une femme, les variations entre la façon de prononcer « a » ou « i » varient toujours dans la même direction, alors il devient possible de raffiner notre modèle simplement en précisant le genre du sujet. Donc, à partir d'un modèle qui contient un plus petit nombre de facteurs sous-jacents, on peut développer un modèle complet qui serait à même d'offrir de meilleures prédictions.

Ensuite, Patrick Kenny découvre que les variations sonores reliées à la voix humaine et celles liées au canal et à l'environnement (médium de communication, téléphone, type de micro, bruit de fond) sont en fait deux groupes de données orthogonales. C'est-à-dire qu'elles sont situées sur 2 plans différents lorsqu'on visualise les données. Donc, si on trouve la bonne rotation, le bon angle pour « visualiser » les données, on peut séparer ce qui appartient aux caractéristiques de la voix elle-même et ce qui y est extérieur. Cette technique se révèle rapidement essentielle pour nettoyer les données et raffiner les modèles! On peut maintenant séparer approximativement la voix et ses variations des facteurs externes.



## 3. -Vecteurs

Le chercheur Najim Dehak a poussé les réflexions de Patrick Kenny plus loin. Comme les facteurs sous-jacents étaient à l'époque utilisés afin de bâtir et de raffiner les modèles (GMM UBM), il s'est demandé s'il ne serait pas plus efficace d'utiliser directement ces facteurs comme base de comparaison, sans passer par des modèles complets. Sa solution : extraire les facteurs de l'enregistrement d'inscription ainsi que ceux de l'enregistrement d'authentification (i-vecteurs) et les comparer. Cette méthode offre des avantages énormes, particulièrement au niveau du temps requis et de la simplification des calculs. La méthode des i-vecteurs est celle qui a le meilleur taux de succès aujourd'hui.




## 4. Réseaux de neurones

Les réseaux de neurones ont plusieurs applications efficaces dans de nombreux domaines d'innovation, mais pour la biométrie vocale ils en sont encore à l'étape du développement. Pour l'instant, les réseaux de neurones sont surtout utiles afin de faciliter le classement des données audio et des enregistrements. Ils sont encore peu utilisés comme technique d'identification vocale. Par ailleurs, en 2016, les premiers systèmes basés sur le *deep learning* ont finalement commencé à donner des résultats satisfaisants en biométrie vocale, après de nombreux succès dans le domaine de la reconnaissance de la parole. Cette avenue de recherche devrait se développer énormément durant les années à venir.

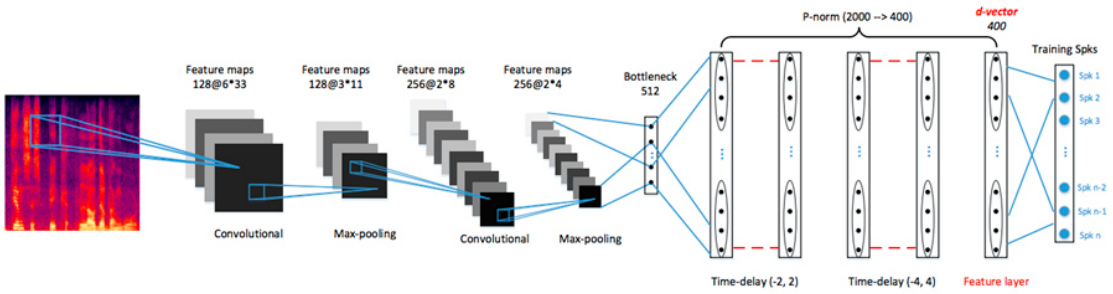
## Comment évaluer l'état de l'art en biométrie vocale?

Pour déterminer les techniques de biométrie vocale les plus efficaces ainsi que les avenues de recherche prometteuses, la plupart des chercheurs se fient aux résultats des évaluations du **NIST SRE (National Institute of Standards and Technology – Speaker Recognition Evaluation)**. Ces compétitions internationales permettent à plusieurs équipes de comparer leurs méthodes et leurs résultats en utilisant les mêmes données. Le CRIM participe à ces évaluations technologiques depuis 2005, et ses experts se sont toujours très bien classés dans ces compétitions internationales.

➔ **Pour en savoir plus, consultez la fiche**

## À SURVEILLER

Certains chercheurs ont proposé un type de système qui semble prometteur pour l'avenir. Il s'agit d'un modèle avec deux réseaux de neurones : le premier classifie les phonèmes (pour déterminer les sons, donc les mots, qui sont dits), et l'autre les caractéristiques de la voix (pour identifier le locuteur). L'innovation majeure, c'est que chacun des deux systèmes fournit ses résultats à l'autre, et vice versa. Les deux systèmes s'entraident afin d'augmenter le taux de réussite. L'idée générale est que si on sait ce qui a été dit, il est peut-être plus facile de savoir qui l'a dit, et que lorsqu'on sait qui parle et les variations possibles de la voix, il sera plus facile de décrypter le contenu du son (ce qui est dit). Qui sait, cette innovation aura peut-être bientôt le potentiel de remplacer les i-vecteurs comme technique principale d'identification vocale !




*Tiré de Wang 2017*