Benchmark techniques vs. science ouverte : à qui revient la parole ?

Brigitte Bigi

Benchmark techniques vs. science ouverte : à qui revient la parole ?

L'exemple de SPPAS comme réponse concrète aux limites de l'évaluation numérique.

Ce document propose une prise de position sur les limites des benchmarks techniques dans le traitement automatique de la parole, en prenant le logiciel libre SPPAS comme étude de cas illustrant une approche conforme aux principes FAIR de la science ouverte.

Brigitte Bigi
LPL, CNRS, Aix Marseille Univ

15 juillet 2025

Mots-clés : SPPAS, segmentation automatique de la parole, benchmarks, science ouverte, FAIR, évaluation, reproductibilité, alignement forcé.

Résumé

Les systèmes de segmentation automatique de la parole reposent sur une série d'étapes fondamentales (normalisation, phonétisation, alignement) souvent dissimulées dans des pipelines opaques. Cette invisibilisation limite fortement le contrôle, la reproductibilité et l'adaptabilité des traitements, en particulier dans les contextes de recherche. Parallèlement, les benchmarks standards utilisés pour évaluer ces outils ne prennent en compte qu'un sous-ensemble des dimensions pertinentes, ignorant les phénomènes propres à la parole spontanée, la transparence du processus, ou encore la possibilité d'adaptation aux besoins d'un projet scientifique.

Dans ce contexte, SPPAS adopte une approche radicalement différente. Conçu comme un outil de recherche, il offre une architecture ouverte, modulaire et contrôlable, où chaque étape du traitement est explicite et chaque ressource modifiable. Ce positionnement, conforme aux principes FAIR (*Findable, Accessible, Interoperable, Reusable*), permet un usage rigoureux, reproductible et collaboratif. En refusant les évaluations réductrices, SPPAS met en lumière la nécessité de repenser les critères de qualité des outils scientifiques. Ce travail défend une évaluation fondée non sur la seule performance technique, mais sur l'utilisabilité réelle, l'ouverture des ressources, et le respect des principes de la science ouverte.

Au-delà de la critique, ce travail propose une réponse partielle mais concrète au problème de l'évaluation : une grille multi-critères combinant ouverture, adaptabilité et utilisabilité, offrant une base initiale pour des comparaisons plus pertinentes que celles fondées sur les seuls scores techniques.

1.

La segmentation automatique de la parole comme tâche invisibilisée

Vue rapide :

- ***Contexte :*** **la segmentation de la parole est une tâche ancienne, omniprésente, mais rarement mise en avant et encore mal outillée pour la recherche en SHS.**
- ***Problème :*** **la majorité des outils sont jugés sur des scores numériques issus de benchmarks techniques.**
- ***Thèse :*** **ces mesures ne reflètent pas les besoins réels des utilisateurs spécialisés car ces métriques techniques sont insuffisantes pour évaluer l'adéquation d'un outil à un usage scientifique.**
- ***Position*** **: SPPAS incarne une alternative conçue dans une logique de transparence, adaptabilité, reproductibilité et service à la communauté scientifique.**

La segmentation automatique de la parole est une tâche ancienne, omniprésente dans de nombreux travaux scientifiques, mais paradoxalement rarement mise au premier plan. Historiquement intégrée à la reconnaissance automatique de la parole (RAP), elle a longtemps été considérée comme une étape « intermédiaire », absorbée dans des chaînes de traitement plus larges. Pourtant, elle constitue un moment critique du traitement de la parole, en particulier pour les chercheurs en phonétique expérimentale, en sciences du langage, ou plus généralement en SHS.

Dès les années 1950, les premières expériences de reconnaissance phonémique ont intégré implicitement des mécanismes de segmentation. Cependant, ce n'est qu'avec la généralisation des modèles de Markov cachés (HMM), à partir des années 1970, que l'on a vu émerger des approches systématiques et modulaires d'alignement et de découpage temporel de la parole. Ces travaux ont posé les bases des systèmes d'alignement forcé, devenus essentiels dans des domaines aussi variés que la dictée vocale, la synthèse vocale ou l'analyse prosodique.

Sur cette base théorique, plusieurs générations d'outils logiciels ont vu le jour. Dans les années 1990, HTK - Hidden Markov Model Toolkit (Young et al., 2002) s'est imposé

comme une référence incontournable : puissant – bien que réservé à des utilisateurs avancés, il a profondément influencé la recherche. En parallèle, CMU Sphinx, développé dès les années 1980 à Carnegie Mellon (Seymore et al., 1998), a marqué les débuts de l'open source dans le domaine, en introduisant des systèmes de reconnaissance vocale à large vocabulaire, indépendants du locuteur. À partir de 1999, le projet Julius (Kawahara et al. 2000), développé par le Speech Research Center au Japon, a proposé une solution rapide, légère, et orientée vers les besoins pratiques. Son intégration dans des environnements Unix, sa compatibilité avec la plupart des distributions Linux et son installation directe via *Homebrew* sur macOS en font encore aujourd'hui l'un des outils les plus faciles à installer pour les utilisateurs non spécialistes. Au début des années 2010, c'est Kaldi (Povey et al., 2011) qui a redéfini les standards de la recherche moderne en reconnaissance vocale. Très modulaire, s'appuyant sur des scripts bien structurés et des techniques avancées (notamment les réseaux de neurones profonds), Kaldi est devenu la boîte à outils privilégiée des chercheurs, malgré sa complexité de prise en main.

Depuis 2020, une nouvelle génération de modèles fondés sur l'apprentissage auto-supervisé a profondément influencé les approches de traitement du signal vocal. Wav2vec 2.0 (Baevski et al., 2020) a été l'un des premiers à démontrer l'efficacité de représentations acoustiques apprises directement à partir du signal brut, sans supervision alignée. Il a été suivi de modèles comme HuBERT (Hsu et al., 2021) ou Whisper (Radford et al., 2022), qui poursuivent cette logique avec des architectures de type Transformer. Ces modèles sont désormais intégrés dans de nombreux frameworks open source, notamment Hugging Face Transformers (Wolf et al., 2020), et servent de base à des systèmes complets de transcription ou d'alignement. Des projets comme *WhisperX* ou certaines extensions de *pyannote.audio* utilisent ces représentations pré-entraînées pour proposer des segmentations automatiques fines, exploitables dans des contextes applicatifs. Cette évolution marque une rupture avec les chaînes traditionnelles basées sur HMM ou GMM décrites précédemment, en remplaçant l'architecture classique par des blocs neuronaux entraînés de bout en bout.

Ces avancées techniques, bien que significatives, répondent en priorité aux besoins de la reconnaissance vocale à grande échelle — transcription, indexation, commande vocale — plutôt qu'à ceux de l'analyse linguistique fine. Or, dans des disciplines comme la phonétique, la linguistique ou l'étude du langage parlé, la segmentation demeure une tâche centrale et encore imparfaitement outillée. Elle ne se limite pas à aligner des phonèmes, mais exige de prendre en compte la diversité des conventions d'annotation, la variabilité des langues et dialectes, ou encore la richesse des phénomènes interactionnels (disfluences, amorces, rires, etc.). Bien que certains considèrent la segmentation comme une tâche « résolue » ou « dépassée », les réalités de terrain montrent au contraire sa persistance comme problème scientifique non trivial — en particulier dans les corpus spontanés, multi-locuteurs, ou issus de contextes minoritaires. Aucun des systèmes mentionnés ci-dessus, aussi puissant soit-il, ne répond aux besoins concrets de segmentation dans un cadre scientifique exploratoire : visualisation, correction, documentation, adaptation au domaine. Ils sont évalués par et

pour des développeurs, pas par les chercheurs en SHS qui utilisent les résultats pour produire de la connaissance. Finalement, les systèmes « s'améliorent », mais les utilisateurs ne sont pas mieux servis.

Il est donc nécessaire d'apporter un regard critique à l'évaluation de ces systèmes. Aujourd'hui encore, l'évaluation des outils de segmentation repose presque exclusivement sur des indicateurs chiffrés tels que l'UBPA - *Unit Boundary Positioning Accuracy*, qui mesure le pourcentage de frontières automatiques situées dans un intervalle (souvent 20 ms) autour des frontières manuelles. Ces métriques techniques sont devenues des standards implicites, valorisés dans les publications, les dépôts de code et les comparaisons entre systèmes. Des travaux ont montré que même entre annotateurs humains, l'accord plafonne à environ 93–94 % dans cet intervalle, ce qui souligne le caractère ambigu et graduel de nombreux phénomènes phonétiques (Hosom, 2009). Si ces chiffres fournissent une forme de repère, ils sont trop souvent interprétés de manière absolue, sans prise en compte du contexte d'usage. Pourtant, ils reflètent très imparfaitement les besoins réels des utilisateurs spécialisés. Ils mesurent une précision numérique, souvent calculée sur des corpus normés, mais ne disent rien de la capacité de l'outil à s'adapter à des usages concrets, à des corpus réels, à des contextes scientifiques variés. Par exemple, une amélioration de 1 % du taux d'accord (de 92 % à 93 %) peut sembler significative, mais son impact pratique dépend entièrement du type d'utilisation :

- Pour un phonéticien qui corrige manuellement les alignements dans un outil comme Praat (Boerasma, 2001), cette différence n'a aucun effet tangible : dans les deux cas, il doit vérifier l'ensemble des frontières et corriger manuellement les quelques cas fautifs.
- Pour une analyse automatique de grandes bases de données vocales (extraction de formants, mesures prosodiques), les erreurs d'alignement sont filtrées statistiquement en aval. Une légère différence de précision en amont n'affecte donc pas nécessairement la qualité finale des analyses, car les cas aberrants sont exclus.

L'interprétation brute des scores comme indicateur universel de qualité est donc trompeuse. Elle occulte la diversité des usages, des langues, des conventions, et des publics concernés. En focalisant l'attention sur des chiffres "objectifs", elle contribue à invisibiliser des dimensions pourtant essentielles : l'adaptabilité de l'outil, sa transparence, ou sa capacité à intégrer les contraintes spécifiques d'un corpus.

Dans ce contexte, le présent document défend la position suivante : un logiciel ne peut pas être jugé uniquement à l'aune de performances numériques.

> ⚠ **Performant ne veut pas dire utilisable**

Un outil peut produire des résultats techniquement excellents sans pour autant être réellement utilisable par celles et ceux à qui il est destiné. Dans le domaine de la segmentation de la parole, de nombreux systèmes — qu'ils soient historiques ou très récents — reposent sur des technologies de pointe, mais imposent encore des conditions d'usage restrictives : installation complexe, dépendance à un système d'exploitation particulier, recours exclusif à la ligne de commande, ou encore nécessité de configurer manuellement des paramètres sans documentation adaptée. Ces exigences, bien qu'acceptables pour des profils informatiques expérimentés, constituent un obstacle important pour une part significative des chercheurs en sciences du langage, en phonétique ou en linguistique interactionnelle, qui n'ont ni le temps ni la formation technique pour s'approprier de tels outils.

Cette réalité, souvent négligée dans les publications techniques ou les dépôts open source, souligne une vérité simple mais trop souvent oubliée : la performance algorithmique ne suffit pas. Concevoir un outil pour la recherche scientifique suppose également de prendre en compte l'ergonomie, la diversité des profils utilisateurs, la simplicité d'usage, et l'accessibilité.

SPPAS (Bigi 2015) s'inscrit dans une logique radicalement différente : celle d'une recherche logicielle ouverte, fondée sur la transparence des traitements, la modularité des composants, et l'adaptabilité aux besoins spécifiques des chercheurs en phonétique, linguistique, et plus généralement en sciences du langage. Son objectif n'est pas de produire de meilleurs scores sur des benchmarks standards, mais de mieux répondre aux usages concrets de la recherche scientifique de ces domaines. L'ambition initiale de SPPAS a donc été inverse à celle de nombreux outils d'alignement : concevoir un système complet, reproductible, contrôlable, mais suffisamment accessible pour être utilisé par des non-informaticiens. Cela s'est traduit par des choix structurels précis : la présence d'une interface graphique (dont l'utilisation est optionnelle), l'utilisation de fichiers de configuration lisibles et documentés, ou encore l'export direct vers des formats largement adoptés dans les communautés ciblées (comme les fichiers TextGrid compatibles Praat).

Et pourtant, malgré ces précautions, il reste des barrières : certains utilisateurs expriment déjà des réticences face à l'installation préalable de Python — preuve que même un outil pensé pour l'usage réel peut rencontrer des freins inattendus. Ce constat, trop souvent négligé dans les publications techniques rappelle une évidence essentielle qui est que **l'utilisabilité ne se mesure pas en millisecondes d'écart sur une frontière de phonème, mais en obstacles que l'on parvient — ou non — à lever pour permettre une appropriation effective.**

Cette posture scientifique — fondée sur la transparence, la reproductibilité et l'utilisabilité — est précisément ce que les benchmarks ne mesurent pas, et ce que ce document se propose de défendre.

2.

Une chaîne invisible mais toujours présente

Vue rapide :

- ***Contexte* : Tout outil de segmentation automatique repose sur une séquence d'étapes fondamentales (normalisation, phonétisation, alignement), qui sont systématiquement présentes dans tous les systèmes existants.**
- ***Problème* : Ces étapes sont souvent dissimulées ou fusionnées dans des boîtes noires logicielles, empêchant leur contrôle, leur compréhension ou leur adaptation par l'utilisateur final.**
- ***Thèse* : La séparation explicite et la capacité de modifications de ces étapes est essentielle pour garantir un usage scientifique fiable, reproductible et adapté à la diversité des contextes réels.**
- ***Position* : SPPAS rend visibles, configurables et partageables chacune des étapes du traitement, en accord avec les principes FAIR et dans une optique de service aux besoins concrets des chercheurs.**

Tout système de segmentation automatique, quelle que soit sa forme ou sa sophistication technique, repose inévitablement sur une même séquence d'opérations fondamentales :

1. Une normalisation du texte brut, qui transforme les chiffres, dates, abréviations ou ponctuations en formes compatibles avec une conversion en phonèmes ;
2. Une conversion en phonèmes, à l'aide de ressources linguistiques (dictionnaires, grammaires, réseaux de neurones entraînés) ;
3. Un alignement temporel qui projette la séquence de phonèmes sur le signal audio en exploitant un modèle (traditionnellement un modèle acoustique).

Ces trois étapes sont toujours présentes, même lorsqu'elles sont invisibles à l'utilisateur. Dans la majorité des outils récents, elles sont encapsulées dans un pipeline unique, parfois intégré dans un exécutable ou dans une API distante. Cette invisibilisation du pipeline est d'autant plus problématique que les benchmarks actuels n'évaluent en général que la dernière étape (l'alignement), et encore, selon un critère unique (la distance temporelle entre frontières de phonèmes). Toute erreur introduite plus tôt — mauvaise normalisation, phonétisation inadaptée — est alors invisible dans l'évaluation, bien qu'elle puisse fausser l'ensemble du résultat final.

Cette dissimulation des étapes n'est pas neutre puisqu'elle prive l'utilisateur de toute vue sur le traitement, l'empêche d'intervenir à un stade donné, et rend toute tentative de validation scientifique difficile, voire impossible. L'analyse linguistique ou phonétique exige pourtant transparence et traçabilité : savoir comment les données sont transformées, pouvoir modifier les ressources utilisées (par exemple le dictionnaire de prononciation), ou encore identifier quelle étape a introduit une erreur ou une approximation.

À ce stade, il est essentiel de distinguer deux niveaux de conception dans les systèmes de segmentation automatique :

- d'une part, les outils algorithmiques comme *HTK*, *Julius* ou *Kaldi*, qui réalisent l'alignement temporel en s'appuyant sur un modèle acoustique et une procédure d'optimisation (souvent via un algorithme de type Viterbi) ;
- d'autre part, les « *wrappers* » ou environnements utilisateurs comme *SPPAS*, *MFA* ou *WebMAUS*, qui encapsulent ces moteurs pour en faciliter l'usage. Ces wrappers prennent en charge la gestion des ressources linguistiques (dictionnaires, modèles), la configuration des appels, le format des données, l'organisation des sorties, voire l'intégration dans des pipelines d'analyse plus larges.

Cette distinction n'est pas seulement conceptuelle : elle est révélatrice d'un besoin structurel de médiation entre la puissance algorithmique d'un moteur et l'usage réel des utilisateurs. Le simple fait que des dizaines de wrappers aient été développés autour de HTK ou Kaldi témoigne de l'inadéquation persistante de ces moteurs bruts avec les besoins concrets de segmentation, d'annotation ou de correction.

Un inventaire publié en 2018 (https://github.com/pettarin/forced-alignment-tools) recensait déjà plus de 40 outils de ce type — chacun tentant, à sa manière, de combler le fossé entre les exigences techniques de l'alignement automatique et les réalités de l'usage scientifique. Et ce chiffre n'a cessé d'augmenter depuis. Cette prolifération illustre à quel point la question centrale n'est pas seulement "quel algorithme aligne le mieux", mais "*quel outil rend ce traitement réellement exploitable*".

Avec l'essor des modèles de reconnaissance vocale fondés sur des réseaux neuronaux profonds, comme *Wav2Vec 2.0*, *HuBERT* ou *Whisper*, un nouveau paradigme technique s'est imposé : la reconnaissance directe à partir du signal, sans explicitation des étapes intermédiaires. Ces modèles, souvent encapsulés dans des bibliothèques comme Hugging Face ou PyTorch, permettent d'obtenir rapidement des transcriptions à partir d'un simple fichier audio. Ils sont ainsi devenus des composants-clés de nouveaux wrappers comme *WhisperX* ou *pyaudio-align.* Pour les informaticiens, cette évolution représente un gain d'utilisabilité considérable par rapport aux moteurs précédents (HTK, Kaldi), souvent lourds à compiler et à configurer. Mais pour les chercheurs non techniciens, la barrière reste intacte : installation en ligne de commande, dépendances multiples, absence d'interface, sorties peu documentées. La promesse de simplicité reste, dans les faits, largement illusoire. Mais surtout, cette nouvelle génération d'outils hérite d'une limite conceptuelle profonde : elle ne voit pas ce qui n'est pas lexical.

Les modèles comme *Whisper* ou *Wav2Vec* sont entraînés pour transcrire efficacement de la parole fluide, standardisée, exempte de « perturbations ». Ces modèles ont démontré d'excellentes performances pour la transcription automatique de la parole — mais cette transcription est conçue dans une perspective lexicale, fluide, épurée. Les représentations auto-supervisées qu'ils exploitent sont entraînées sur des corpus normalisés, dépouillés de tout ce qui les gène. Dans leur logique, tout ce qui relève des rires, soupirs, chevauchements, hésitations, amorces, disfluences — est traité comme du bruit. Ces événements, pourtant au cœur des analyses en linguistique interactionnelle ou en phonétique conversationnelle (par exemple (Bigi & Meunier, 2018)), sont soit ignorés, soit activement filtrés. Le pipeline de ces modèles ne prévoit aucune place pour les formes non lexicales, et n'offre aucun mécanisme natif pour les détecter, les annoter ou même simplement les laisser visibles.

Ce silence algorithmique est problématique. Il ne s'agit pas d'un simple "manque de performance" : c'est un biais de conception. Ces outils ne ratent pas les disfluences et autres événements — ils ne les cherchent même pas. Et dans des contextes de recherche où ces phénomènes sont centraux, leur usage conduit à une normalisation forcée de la parole, en rupture avec les objectifs mêmes de l'analyse. Car tous ces phénomènes sont au cœur de l'analyse du langage parlé réel :

- en phonétique (ex. durée, prosodie, pauses)
- en interaction (ex. gestion du tour de parole)
- en sociolinguistique (ex. marques de style ou d'émotion)

Ce que ces modèles offrent en précision temporelle, ils le perdent en fidélité phénoménologique. Ils sont conçus pour "entendre des mots" — pas pour écouter la parole humaine dans toute sa richesse. **Un outil qui aligne parfaitement des mots, mais ignore ce qui fait la spécificité du parler humain, ne peut pas répondre aux attentes des chercheurs en sciences du langage.**

Chaque phonéticien travaille avec des objectifs, des corpus et des conventions spécifiques. Il n'existe pas de "standard universel" de l'annotation phonétique : certains projets privilégient une segmentation fine de certains phonèmes, d'autres se satisfont de quelques imprécisions au profit d'une validité globalement correcte. Les conventions, comme les objectifs, varient selon les langues, les domaines, ou les finalités de recherche. Cette diversité n'est pas une faiblesse mais une richesse du champ — à condition qu'elle soit prise en compte par les outils.

C'est précisément en raison de cette exigence de contrôle, de précision, et d'adaptation que les annotations ont longtemps été réalisées à la main. Le chercheur devait pouvoir décider des unités pertinentes, des critères de segmentation, ou des cas limites à trancher. Or les systèmes de segmentation automatique actuels, lorsqu'ils imposent des chaînes de traitement opaques ou des conventions fixes, reproduisent l'illusion d'un alignement objectif, au détriment de l'ajustement raisonné au terrain. Un bon outil ne doit donc pas imposer un modèle unique. Il doit, au contraire, fournir à l'utilisateur les moyens d'adapter les ressources, d'explorer les variantes, de documenter ses choix.

Cette adaptabilité raisonnée n'est pas un luxe, mais une exigence minimale pour garantir la validité scientifique des annotations produites.

> 📌 **Idée principale** : La segmentation est toujours un processus en plusieurs étapes. Pour qu'un outil soit réellement scientifique, ces étapes doivent être visibles, compréhensibles et ajustables par l'utilisateur.

Face à cette opacité généralisée des systèmes de segmentation, qu'elle soit technique ou méthodologique**,** SPPAS adopte une posture radicalement différente**,** fondée sur les principes **FAIR** (*Findable, Accessible, Interoperable, Reusable*), aujourd'hui reconnus comme un socle des bonnes pratiques en **science ouverte**. Cette exigence de transparence, bien qu'inscrite depuis 2016 dans le droit français via l'article 30 de la loi pour une République numérique - https://www.legifrance.gouv.fr/jorf/article_jo/JORFARTI000033202841 (accès ouvert par défaut aux résultats de la recherche publique), faisait déjà partie intégrante de la conception de SPPAS dès sa première publication en 2012. L'ouverture, la possibilité de modifier les ressources, et de partager les traitements sont au cœur même de sa conception.

Dans SPPAS, chaque étape du processus est clairement identifiée : la normalisation du texte, la conversion graphème-phonème et l'alignement temporel sont séparées, explicites et contrôlables. Les ressources utilisées sont toutes fournies, documentées et modifiables — qu'il s'agisse des lexiques, des dictionnaires de prononciation, des modèles acoustiques ou des règles de traitement. Enfin, chaque traitement est traçable et reproductible : aucun paramètre n'est dissimulé, aucune transformation n'est appliquée sans que l'utilisateur puisse la comprendre, l'ajuster ou la corriger. L'ensemble du code source est publié sous licence libre (AGPL), et les ressources sont elles aussi distribuées sous des licences ouvertes compatibles avec la recherche et la réutilisation. **Ce n'est pas un simple choix de conception — c'est une exigence scientifique**. Dans des disciplines où la parole est objet d'analyse fine, où les conventions varient selon les langues, les corpus ou les objectifs de recherche, un outil "boîte noire" est une contradiction. On ne peut pas étudier ce que l'on ne voit pas.

> 📦 SPPAS rend visible ce que les autres cachent

Contrairement à la majorité des wrappers, SPPAS a été pensé dès l'origine comme un outil de recherche, au service des usages concrets. Il ne se contente pas de fournir un alignement, mais offre à l'utilisateur un contrôle complet sur le processus : le dictionnaire phonétique peut être enrichi ou modifié, les variantes inappropriées supprimées, les ressources partagées ou adaptées à des corpus spécifiques (langage

enfantin, locuteurs dialectaux, parole pathologique...). La transparence est structurelle : les fichiers intermédiaires sont lisibles, les modules configurables, les étapes séparées et explicitées. L'interface graphique permet une prise en main accessible, mais n'enferme pas l'utilisateur dans un parcours contraint.

Ce positionnement reflète une conviction forte : un outil scientifique ne doit pas seulement "donner un résultat", mais s'inscrire dans un dialogue avec l'utilisateur, en respectant les exigences de traçabilité, d'adaptabilité, et de reproductibilité, propres à la recherche.

La pertinence de SPPAS tient dans sa capacité à s'ajuster aux pratiques de la recherche. Or, **c'est précisément cette dimension — ce qu'un outil rend possible, et non uniquement ce qu'il produit — que les benchmarks actuels laissent dans l'ombre.**

💡 **En résumé,** SPPAS, c'est « FAIR » de la science !

Tous les outils de segmentation réalisent les mêmes étapes (normalisation, phonétisation, alignement), mais trop souvent de manière opaque. SPPAS les rend visibles, modifiables et traçables — conformément aux principes FAIR — ce qui permet aux chercheurs d'ajuster le traitement à leurs besoins réels. Cette adaptabilité, pourtant essentielle, est ignorée par les benchmarks techniques.

3.

Les limites fondamentales des benchmarks actuels

Vue rapide :

- ***Contexte*** **: Les benchmarks techniques (UBPA, PER, WER, etc.) sont devenus la référence quasi exclusive pour évaluer les outils de segmentation de la parole.**
- ***Problème*** **: Ces benchmarks se focalisent sur des écarts locaux par rapport à des références manuelles, sans jamais prendre en compte la transparence, la reproductibilité ni l'adéquation à l'usage réel.**
- ***Thèse*** **: Un outil peut obtenir un excellent score tout en restant inutilisable pour les chercheurs en linguistique, phonétique ou SHS.**
- ***Position*** **: SPPAS propose une approche radicalement différente, fondée non sur l'optimisation d'un score unique, mais sur la lisibilité et la pertinence de chaque étape pour les usages scientifiques réels.**

Malgré leur technicité apparente, la plupart des systèmes de segmentation automatique reposent sur les mêmes fondements : une chaîne de traitement implicite rarement exposée et jamais évaluée dans sa globalité. Cette invisibilité méthodologique est renforcée par la place dominante qu'occupent aujourd'hui **les benchmarks techniques**, lesquels condensent la qualité d'un outil en quelques **scores numériques standardisés**. Pourtant, ces évaluations masquent des dimensions essentielles : la transparence des traitements, leur traçabilité, leur reproductibilité, ou encore leur adéquation aux données réelles. Cette section examine les limites structurelles des benchmarks actuels, en analysant successivement les trois étapes fondamentales d'un système de segmentation : la normalisation, la phonétisation et l'alignement forcé. Pour chacune d'elles, il s'agit de préciser ce que les benchmarks mesurent — et surtout ce qu'ils ne mesurent pas.

3.1 Normaliser pour phonétiser : une étape (trop) invisible

Avant toute conversion d'un texte en phonèmes, une étape préalable et indispensable s'impose : la normalisation du texte. Il s'agit d'un ensemble d'opérations visant à transformer un texte brut en une séquence linguistiquement et phonétiquement exploitable. Cette normalisation peut inclure la segmentation en unités, la suppression ou la transformation de signes de ponctuation, le traitement des chiffres, sigles, dates, abréviations, formes contractées ou régionales. Certains outils récents, notamment dans

les chaînes Python, appliquent même des transformations particulièrement agressives, comme la romanisation systématique des caractères, ce qui revient à supprimer les accents ou distinctions diacritiques importantes : “côte”, “côté” deviennent indistinctement “cote”, tout comme “élevé” et “élève” deviennent “eleve” — avec des conséquences évidentes sur la phonétisation.

Dans les langues où les conventions orthographiques sont stables — comme l’anglais écrit standard — cette étape est souvent intégrée discrètement dans le pipeline, voire passée sous silence. Mais dans la plupart des langues, cette normalisation est loin d’être triviale : en français, les contractions, élisions et variations typographiques dépendent du domaine, du registre ou du style ; dans les langues non segmentées (comme le thaï, le khmer, le chinois traditionnel ou simplifié, …), l’absence d’espaces entre les mots exige des algorithmes spécifiques de segmentation morphosyntaxique. Dans tous ces cas, la qualité de la normalisation conditionne directement la fiabilité des étapes suivantes.

Pourtant, cette étape est rarement évaluée en tant que telle. Elle est soit implicitement incluse dans l’évaluation des modules G2P (grapheme-to-phoneme), soit absente des benchmarks. Les rares études existantes se concentrent souvent sur des cas limités (comme la conversion de nombres en lettres en synthèse vocale), et ne tiennent pas compte des phénomènes présents dans les corpus oraux : abréviations, hésitations, mots inventés, formes incomplètes, etc. De plus, les outils actuels n’offrent que peu ou pas de transparence sur ce qu’ils transforment, ni comment ils le font.

SPPAS, au contraire, fait de la normalisation une étape autonome, documentée, publiée (Bigi, 2014) et modulaire du pipeline. Il propose une solution basée sur des modules génériques, valables pour de nombreuses langues ; et des modules spécialisés par langue, remplaçables ou ajustables selon les besoins.

Ce choix technique et scientifique permet de trouver un équilibre entre deux extrêmes : d’un côté, une approche “universelle” simpliste qui ignore les spécificités linguistiques ; de l’autre, une approche par langue totalement indépendante, difficile à maintenir et mutualiser. En segmentant clairement cette étape, en en fournissant les composants, et en laissant la possibilité de les modifier ou les remplacer, SPPAS garantit la reproductibilité, l’adaptabilité et la traçabilité du traitement, en accord avec les principes FAIR.

> 👉 Aucun benchmark actuel ne permet de mesurer la qualité, la modularité ou la pertinence de cette normalisation.

3.2 Phonétisation : ce que mesure (réellement) PER

La phonétisation (ou conversion grapheme-to-phoneme, G2P) est l’étape qui consiste à transformer une séquence de mots écrits en une séquence de phonèmes, en vue de l’aligner avec un signal audio ou de le synthétiser. Cette étape repose historiquement :

- soit sur des dictionnaires phonétiques manuels,
- soit sur des règles de prononciation explicites,
- soit sur des modèles statistiques ou neuronaux apprenant ces correspondances à partir de données annotées.

Si cette tâche paraît bien définie sur des données écrites, lexicales, et standardisées, elle devient beaucoup plus délicate dès que l'on aborde la parole réelle, marquée par des formes inédites, des mots hors lexique, des réductions phonétiques, ou des effets dialectaux.

Pour la segmentation automatique de la parole, la phonétisation s'intercale entre la normalisation du texte et l'alignement avec le signal audio, et constitue un maillon indispensable de tout système de traitement de la parole. Pourtant, dans de nombreux outils, cette étape est intégrée de manière invisible : l'utilisateur n'a aucun moyen de savoir comment une transcription phonétique a été produite, ni pourquoi, ni comment la corriger ou l'adapter.

Les évaluations classiques de G2P reposent sur des métriques comme le PER - *Phoneme Error Rate*, calculé entre une référence et la sortie du système sur des corpus alignés et normés comme CMUdict ou Librispeech (Panayotov et al., 2015). Ces benchmarks mesurent essentiellement la capacité à reproduire une prononciation canonique pour des mots déjà connus. Dans ce cadre, les performances peuvent être impressionnantes. Certains modèles atteignent une PER inférieure à 0,5 % pour des langues comme le coréen, et les architectures de type Transformer surpassent désormais les LSTM sur la plupart des jeux de données standards. Les LSTM - *Long Short-Term Memory*, sont des réseaux de neurones récurrents conçus pour modéliser des séquences, en conservant une mémoire des éléments passés. Ils ont longtemps été l'architecture dominante pour les tâches séquentielles en Taitement Automatique des Langues et en Reconnaissance Automatique de la Parole. Les Transformers (Vaswani et al., 2017) reposent entièrement sur des mécanismes d'attention, sans récursivité. Ils permettent un parallélisme plus efficace et capturent les dépendances longues plus facilement. Depuis 2018, ils ont largement remplacé les LSTM dans la plupart des tâches et commencent à dominer aussi dans la reconnaissance automatique de la parole, la traduction et le G2P.

Mais ces scores élevés masquent une réalité : **les benchmarks actuels ne testent que des cas idéaux**. Ils ne disent rien sur la capacité à gérer des mots inconnus, des réductions orales, des amorces, des hésitations, ou des variantes dialectales. En bref, ils ne disent rien sur la parole telle qu'elle est effectivement produite et étudiée dans les recherches en linguistique. Or ces formes non canoniques sont particulièrement riches en informations phonétiques, sociolinguistiques ou interactionnelles.

Face à cela, SPPAS adopte une approche radicalement différente. Tout d'abord, la phonétisation repose sur un dictionnaire ouvert et modifiable. Pour chaque mot connu, toutes les variantes sont conservées. Si un mot est inconnu, un mécanisme de *longest matching* est utilisé pour le découper en segments phonétisables, à partir d'unités présentes dans le dictionnaire (lettres, syllabes, morphèmes). Cela permet à la fois de

générer automatiquement des transcriptions plausibles dans les langues peu dotées, et d'atteindre une large couverture lexicale dans les langues mieux pourvues, sans créer de "trou" dans les données.

Crucialement, SPPAS ne choisit pas à la place de l'utilisateur. Il n'impose pas une prononciation unique, ni ne cache les variantes. Le choix de la séquence phonétique la plus compatible est délégué à l'algorithme d'alignement, qui exploite le signal pour résoudre l'ambiguïté. Ce découplage entre génération et sélection garantit non seulement une transparence maximale du traitement, mais permet aussi d'obtenir une phonétisation aussi proche que possible de la réalité acoustique observée — sans imposer de standardisation a priori. Là où les benchmarks mesurent l'adhésion à un standard illusoire, SPPAS privilégie l'adaptation au corpus, la liberté de modifier les ressources, et la reproductibilité des résultats. Ce sont ces propriétés, pourtant essentielles pour la recherche, que les benchmarks échouent à mesurer. Enfin, comme pour la normalisation, SPPAS fait de la phonétisation une étape autonome, documentée, publiée (Bigi, 2016) et modulaire du pipeline.

> 👉 Les benchmarks de G2P évaluent une capacité à produire des séquences de phonèmes corrects sur des mots connus, dans une langue canonique. Ils ignorent tout ce qui caractérise la parole réelle : spontanée, désordonnée, idiosyncratique, dialectale, émotionnelle, inachevée.

3.3 Alignement forcé : ce que mesure (réellement) l'UBPA

L'alignement forcé est la tâche qui consiste à associer chaque segment d'un signal audio à une unité linguistique (souvent des phonèmes), en déterminant les bornes temporelles précises de ces unités. Ce processus intervient après la génération d'une ou plusieurs séquences phonétiques issues de la transcription textuelle, par un module de phonétisation. Dans la plupart des systèmes, l'alignement ne sélectionne pas une séquence, mais prend en entrée une grammaire, c'est-à-dire un ensemble de variantes phonétiques possibles, et détermine la séquence la plus probable en fonction du signal.

Dans ce contexte, l'UBPA - *Unit Boundary Positioning Accuracy*, est une métrique souvent utilisée pour évaluer la précision temporelle de l'alignement. Elle calcule la proportion de frontières détectées automatiquement qui tombent dans une certaine tolérance (souvent ±20 ms) autour des frontières définies manuellement. En apparence simple et objective, cette mesure repose sur plusieurs hypothèses implicites qui en limitent fortement la portée. D'abord, pour que l'UBPA soit calculable, il faut fournir une séquence unique de phonèmes correspondant exactement à l'annotation manuelle. Cela neutralise toute variabilité liée au G2P (qui pourrait proposer plusieurs séquences possibles) et réduit artificiellement la tâche à un problème d'alignement pur, dans des conditions rarement réalistes. L'UBPA ne mesure donc ni la capacité du système à choisir la bonne séquence de phonèmes, ni l'adaptabilité à des prononciations non canoniques,

ni la gestion des phénomènes oraux typiques (amorphes, chevauchements, interjections…). Ensuite, l'UBPA évalue uniquement le comportement du moteur acoustique et de son modèle statistique, historiquement basé sur des HMM et une recherche Viterbi. Dans les systèmes classiques, cela signifie que le score reflète la qualité du couple *moteur + modèle*, et non celle de l'outil complet. Lorsque les wrapers tels que SPPAS ou WebMAUS encapsulent des moteurs comme HTK ou Julius, c'est ce couple qui est évalué. Le reste de l'environnement — traitements en amont, ergonomie, documentation, interfaces de correction — reste entièrement hors du champ de la métrique. Par exemple, si l'on pouvait extraire un modèle acoustique depuis WebMAUS et qu'on l'utilise dans SPPAS avec le même moteur (HTK en l'occurrence), on obtiendrait des résultats quasi-identiques. L'UBPA ne mesure que les frontières obtenues des phonèmes et ces dernières ne dépendent que du couple *moteur+modèle*.

Au-delà de ses limites méthodologiques internes, l'UBPA reste aveugle à de nombreux critères qui conditionnent pourtant l'usage scientifique réel d'un outil de segmentation. Elle ne dit rien de la gestion des phénomènes linguistiques typiques de la parole spontanée — disfluences, amorces, chevauchements, pauses, interjections ou rires — qui sont souvent absents des benchmarks, mais omniprésents dans les corpus de recherche. Elle ignore également la variabilité dialectale, les réductions phonétiques ou les variations stylistiques, autant d'éléments pourtant cruciaux en phonétique, sociolinguistique ou analyse interactionnelle. Surtout, l'UBPA ne prend en compte ni l'adaptabilité d'un outil à des corpus spécifiques, ni la transparence du traitement, ni la reproductibilité des résultats. Elle ne renseigne ni sur la documentation fournie, ni sur la facilité à éditer/adapter les ressources linguistiques, ni sur la capacité à corriger ou visualiser les résultats. Ces dimensions — interface utilisateur, contrôle des paramètres, édition des données — sont pourtant fondamentales dans une démarche scientifique rigoureuse.

Pire encore, en valorisant exclusivement la précision temporelle d'un moteur encapsulé, l'UBPA peut involontairement favoriser des systèmes opaques ou rigides, difficiles à adapter, mais optimisés pour un score de benchmark donné. Autrement dit, les outils les mieux classés selon cette métrique sont parfois les moins exploitables pour un chercheur qui souhaite analyser, corriger ou réutiliser les sorties. Surtout, il faut le rappeler avec fermeté : l'UBPA ne mesure en aucun cas la qualité d'un wrapper. Elle évalue uniquement le comportement du couple moteur d'alignement + modèle acoustique. Ainsi, toute étude prétendant comparer des wrappers sur la base de cette seule métrique repose sur une erreur méthodologique : elle confond la performance d'un composant interne avec la valeur scientifique d'un environnement logiciel dans son ensemble.

> 👉 L'UBPA permet d'évaluer un algorithme d'alignement forcé sur corpus standardisé. Elle ne dit rien sur la transparence du traitement, la modularité des ressources, ou l'adaptabilité aux corpus réels.

4.

Conclusion

Vue rapide :

Position : **Le respect de la science ouverte et des principes FAIR implique de rendre chaque étape visible, contrôlable, reproductible. C'est cette transparence que les benchmarks ignorent. Dans la recherche scientifique, un bon score ne suffit pas si on ne sait pas comment on l'a obtenu.**

Tout système de segmentation automatique, quelle que soit sa forme ou sa sophistication, repose inévitablement sur une même séquence d'opérations fondamentales : d'abord une normalisation du texte brut, ensuite une conversion en phonèmes, puis un alignement avec le signal audio. Ces étapes sont toujours présentes, même si elles sont souvent dissimulées dans des interfaces simplifiées ou des pipelines opaques. Cette dissimulation n'est pas anodine : elle prive l'utilisateur de toute capacité de contrôle, d'adaptation ou de vérification. Elle transforme une tâche scientifique en une opération algorithmique non reproductible.

Cette opacité des systèmes de traitement de la parole est confortée par des benchmarks techniques — qui ne prennent en compte ni l'ensemble des phénomènes de la parole, ni la transparence, la reproductibilité ou la traçabilité du processus de segmentation. Un outil peut obtenir un "bon score" tout en restant totalement impropre à toute forme d'analyse linguistique ou phonétique. Le véritable choix pour les utilisateurs n'est donc pas entre des outils « plus ou moins performants » au sens des benchmarks, mais entre des systèmes ouverts, adaptables et reproductibles, et des systèmes fermés, rigides, difficilement exploitables en contexte réel.

Plusieurs cas pratiques illustrent cette inadéquation. Lors d'une correction manuelle, par exemple, une différence de 6 % ou 7 % sur un score de précision n'a pratiquement aucun effet réel. Dans les deux cas, l'utilisateur est contraint de relire l'intégralité du fichier pour s'assurer de la qualité des frontières. Le gain mesuré ne se traduit donc ni en gain de temps, ni en réduction de l'effort cognitif requis. De même, dans les analyses de grande ampleur, l'extraction des informations comme les valeurs de formants des voyelles est automatisée, puis les valeurs aberrantes sont repérées à l'aide de mesures statistiques afin d'être éliminées. Ainsi, une frontière mal placée lors de l'alignement entraîne l'élimination d'un segment : que l'UBPA soit à 93 % ou à 97 % n'induit aucun changement dans l'analyse qui s'ensuit. Enfin, les corpus de parole spontanée, dialoguée

ou pathologique, par exemple, sont traversés de phénomènes que ces outils ignorent systématiquement : réductions, amorces, hésitations, mots inventés, néologismes, expressions dialectales. Or ce sont précisément ces formes atypiques qui intéressent la recherche linguistique, sociolinguistique ou interactionnelle, et que les pipelines opaques tendent à supprimer ou à normaliser.

Dans ce contexte, SPPAS propose une alternative radicalement différente. Conçu comme un outil de recherche plutôt qu'un produit logiciel, SPPAS repose sur des principes clairs, rendant chaque étape du processus visible, chaque ressource modifiable, chaque sortie traçable. Ce positionnement, aligné avec les principes **FAIR**, fait de SPPAS un outil adapté à une utilisation scientifique rigoureuse. Cette approche dépasse largement les critères habituels des benchmarks.

> ***Ce que SPPAS fait, les autres ne le font pas*** : SPPAS rend visible, modifiable et partageable chaque étape du processus de segmentation.

SPPAS a été conçu comme une infrastructure logicielle ouverte, modulaire et contrôlable, en rupture avec les approches opaques des pipelines standards. Il repose sur une séparation explicite entre le moteur d'alignement, interchangeable, et les ressources linguistiques, entièrement accessibles à l'utilisateur. Chaque étape du traitement (normalisation, phonétisation, alignement) a fait l'objet de validations scientifiques documentées et publiées, attestant de la rigueur méthodologique du système. Les ressources utilisées, telles que les dictionnaires phonétiques ou les modèles acoustiques, sont non seulement visibles, mais aussi éditables, remplaçables et partageables. Cette architecture explicite permet une maîtrise fine de l'ensemble du processus, à rebours des systèmes fermés où les choix linguistiques sont dissimulés ou imposés.

Ce positionnement a été reconnu au niveau institutionnel. En 2022, SPPAS a reçu un accessit au prix spécial du jury dans le cadre du **Prix Science Ouverte des Logiciels Libres de la recherche**, décerné par le Ministère chargé de l'Enseignement Supérieur et de la Recherche. Car le logiciel joue un rôle clé dans la recherche scientifique, dont il est à la fois un outil, un résultat et un objet d'étude. Comme le rappelle le deuxième Plan national pour la science ouverte, « *la mise à disposition des codes sources des logiciels, avec la possibilité de les modifier, les réutiliser et les diffuser, est un enjeu majeur pour permettre la reproductibilité des résultats scientifiques et soutenir le partage et la création de connaissances, dans une logique de science ouverte*. »

SPPAS échappe volontairement aux logiques classiques d'évaluation standardisée. Son architecture modulaire, ouverte et configurable rend impossible toute évaluation unidimensionnelle, fondée sur un unique score ou une métrique isolée. Sa richesse fonctionnelle, pensée pour des usages scientifiques réels, déborde largement les cas d'usage figés ou artificiels des benchmarks habituels. Et c'est précisément ce que ces

benchmarks échouent à mesurer — la transparence, l'adaptabilité, la traçabilité — qui constitue le cœur de son utilité pour les chercheurs.

Il est urgent de dépasser l'hégémonie des benchmarks techniques comme seuls instruments de validation des outils scientifiques. Bien qu'utiles dans certains contextes normés, ces évaluations unidimensionnelles échouent à rendre compte des critères fondamentaux de la pratique de recherche : compréhension des processus, adaptabilité aux données réelles, reproductibilité des résultats. Ce constat appelle à un changement de paradigme dans l'évaluation des outils logiciels de et pour la recherche. Les critères de qualité doivent inclure, aux côtés des performances techniques, des dimensions essentielles telles que :

- l'ouverture du code et des ressources, condition première d'un usage scientifique transparent ;
- l'explicabilité des traitements, qui permet à l'utilisateur de comprendre, corriger ou ajuster chaque étape du processus ;
- la reproductibilité des analyses, rendue possible par la traçabilité complète des données, des paramètres et des algorithmes.

Ce sont ces qualités, et non les seuls chiffres de performance sur corpus standardisés, qui garantissent l'adéquation d'un outil aux exigences de la recherche. Valoriser des systèmes capables de s'adapter à la diversité des terrains, de rendre visibles les choix méthodologiques et de respecter les principes de la science ouverte est une nécessité. **Repenser les modalités d'évaluation, c'est aussi affirmer que l'enjeu n'est pas seulement technique : il est épistémologique et éthique.**

5.

Perspectives : vers une approche qualitative de l'évaluation

Ce document ne se contente pas de critiquer les limites des évaluations actuelles : il propose une voie concrète pour les dépasser. Il ne s'agit pas seulement de dénoncer l'inadéquation des benchmarks dans certains contextes, mais d'ouvrir un espace de réflexion sur ce que pourrait (et devrait) être une évaluation utile à la recherche.

C'est dans cet esprit que nous proposons, *à titre indicatif*, un tableau synthétique comparant plusieurs wrappers de segmentation automatique sur des critères élargis. Ces critères ne prétendent pas à l'exhaustivité, ni à l'universalité, mais reflètent des dimensions concrètes, fréquemment rencontrées dans les pratiques scientifiques : transparence, ouverture, contrôlabilité, documentation, respect des licences, possibilités de personnalisation, etc. L'ensemble des critères ont été réparties en 3 dimensions distinctes : 1/ les licences et la conformité légale afin de valider la pertinence avec les critères FAIR, 2/ les fonctionnalités scientifiques afin d'évaluer la pertinence et l'adaptabilité à la tâche, et 3/ l'intégration et la prise en main pour les utilisateurs.

Légende des tableaux :

- ✔ : critère rempli
- ⚠ : critère partiellement rempli ou douteux
- ❌ : critère non rempli
- 🚫 : critère hautement problématique
- ❓ : non connu/non défini
- 🏅 : prix/distinction
- ⚙ : infrastructure

5.1 Licences et conformité

Avant toute considération technique, l'évaluation d'un outil doit porter sur sa légalité d'usage et sa conformité aux principes de la science ouverte. Licence du code, réutilisabilité des ressources linguistiques, respect des licences tierces : ces éléments conditionnent la capacité à diffuser, modifier ou publier des travaux reposant sur l'outil. Ce premier tableau synthétise ces aspects essentiels, souvent absents des publications techniques.

	SPPAS	EasyAlign	WebMAUS	MFA	TorchAudio	Gentle
Licence du code	AGPL v3	❓	❓	MIT	BSD-2 -Clause	MIT
Licence des ressources linguistiques	GPLv3/ CC	❓	❓	MIT	MIT/ varié	MIT
Respect des licences tierces	✔	🚫	✔	✔	✔	✔
Code source : **F – Findable** **A – Accessible** **I – Interoperable** **R - Reusable**	 ✔ ✔ ✔ ✔	 ❌ ⚠ ❌ ❌	 ❌ ❌ ❌ ❌	 ✔ ✔ ❌ ✔	 ✔ ✔ ✔ ✔	 ✔ ✔ ⚠ ⚠
Ressources : **F – Findable** **A – Accessible** **I – Interoperable** **R - Reusable**	 ✔ ✔ ✔ ✔	 ⚠ ✔ ❌ ❌ 	 ⚠ ⚠ ⚠ ❌	 ✔ ✔ ❌ ⚠	 ⚠ ✔ ✔ ⚠	 ❌ ❌ ❌ ❌
Maintenu	✔	❌	✔	✔	✔	✔
Reconnaissance institutionnelle	🏅 MESR	❌	⚙ CLARIN	❌	❌	❌

Tableau 1 : Statut légal et conformité éthique des logiciels et des ressources

5.2 Fonctionnalités scientifiques

Un bon score à un benchmark ne garantit en rien l'utilité d'un outil pour la recherche. Ce tableau se concentre sur les fonctionnalités linguistiques réellement nécessaires en analyse de la parole : séparation explicite des étapes, gestion des formes atypiques, correction possible, ... C'est dans ces dimensions que se jouent l'adéquation au terrain, l'adaptabilité, et la transparence du traitement.

	SPPAS	EasyAlign	WebMAUS	MFA	TorchAudio	Gentle
Séparation claire des étapes	✔	✔	✘	✔	✘	✘
Segmentation semi-automatique possible	✔	✘	✘	✘	✘	✘
Phonétisation des mots inconnus	✔	✘	✘	✔	✘	✘
Gestion des variantes de prononciation	✔	✘	✘	✔	✘	✘
Alignement des rires	✔	✘	✘	✘	✘	✘
Gestion spécifique des hésitations	✔	✘	✘	✘	✘	✘

Tableau 2 : Capacités linguistiques et traitement des cas réels

5.3 Intégration et prise en main

Un outil de traitement de la parole destiné à la recherche ne peut se limiter à un moteur performant. Il doit offrir des moyens concrets d'interaction adaptés à divers profils d'utilisateurs : chercheurs sans compétences en programmation, développeurs, ingénieurs. Ce tableau examine les modalités d'usage.

	SPPAS	EasyAlign	WebMAUS	MFA	TorchAudio	Gentle
Interface graphique (web ou locale)	✔	✔	✔	✘	✘	✔
Utilisation via API	✔	✘	✘	✔	✔	✘
Utilisation via CLI	✔	✘	✘	✔	✔	✔
Formats de sortie lisibles	✔	✔	✔	✔	✔	✔

	SPPAS	EasyAlign	WebMAUS	MFA	TorchAudio	Gentle
Personnalisation facile (édition, remplacement)	✔	✔	✘	✔	✔	✘
Documentation développeur	✔	✘	✘	✔	✔	✘
Documentation utilisateur	✔	✔	✔	✔	✘	✘
Tutoriels didactiques	✔	✘	✘	✔	✘	✘
Compatibilité multi-plateforme	✔	✘	✔	✔	✔	✔

Tableau 3 : Souplesse d'usage et modalités concrètes d'interaction

⚠️ **Avertissement :** malgré tous les soins apportés à la collecte et à la vérification des informations, ces tableaux peuvent comporter des erreurs ou des approximations. Ils sont fournis à titre purement indicatif, dans une perspective comparative et non normative.

L'état des outils ayant pu évoluer, ces données reflètent la situation observée en **juillet 2025** et ne sauraient être tenues pour définitives.

Références bibliographiques

Baevski, A., Zhou, H., Mohamed, A., Auli, M. (2020). wav2vec 2.0: a framework for self-supervised learning of speech representations. Proceedings of the 34th International Conference on Neural Information Processing Systems, Article No.: 1044, pp 12449—12460.

Bigi, B. (2014). A multilingual text normalization approach. *Human Language Technology Challenges for Computer Science and Linguistics, LNAI 8387,* pp. 515–526.

Bigi, B. (2015). SPPAS-multi-lingual approaches to the automatic annotation of speech. *The Phonetician. Journal of the International Society of Phonetic Sciences*, *111*(ISSN: 0741-6164), 54-69.

Bigi, B. (2016). A phonetization approach for the forced-alignment task in SPPAS. *Human Language Technology. Challenges for Computer Science and Linguistics, LNAI 9561,* pp. 515–526.

Bigi, B., & Meunier, C. (2018). Automatic segmentation of spontaneous speech. *Revista de Estudos da Linguagem*, *26*(4).

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot. Int.*, *5*(9), 341-345.

Hosom, J. P. (2009). Speaker-independent phoneme alignment using transition-dependent states. *Speech communication*, *51*(4), 352-368.

Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, *29*, 3451-3460.

Kawahara, T., Lee, A., Kobayashi, T., Takeda, K., Minematsu, N., Sagayama, S., ... & Shikano, K. (2000). Free software toolkit for Japanese large vocabulary continuous speech recognition. In Proc. Int'l Conf. on Spoken Language Processing (ICSLP), Vol. 4, pp. 476--479.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). *Librispeech: An ASR corpus based on public domain audio books*. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5206–5210). IEEE. https://doi.org/10.1109/ICASSP.2015.7178964

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding,* vol. 1.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision*. OpenAI. https://openai.com/research/whisper

Seymore, K., Rosenfeld, R., Chen, S., Eskenazi, M., Gouvea, E., Reddy, R., ... & Thayer, E. (1998). The 1997 CMU Sphinx-3 English broadcast news transcription system.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Rush, A. M. (2020). *Transformers: State-of-the-art natural language processing*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., ... & Woodland, P. (2002). The HTK book. *Cambridge university engineering department*, *3*(175), 12.

Liens utiles (juillet 2025) :

CMUdict - Carnegie Mellon University. *The CMU Pronouncing Dictionary*. Version 0.7b : http://www.speech.cs.cmu.edu/cgi-bin/cmudict

CMUSphinx : https://cmusphinx.github.io/

Julius CSR Engine : https://github.com/julius-speech/Julius

Gentle : https://github.com/strob/gentle

HTK – Hidden Markov Toolkit : https://htk.eng.cam.ac.uk/

MFA - Montreal Forced Aligner : https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

Praat : https://praat.org/

pyannote.audio : Neural building blocks for speaker diarization : https://github.com/pyannote/pyannote-audio

PyTorch – TorchAudio – Alignement forcé : https://docs.pytorch.org/audio/main/generated/torchaudio.functional.forced_align.html

https://github.com/pytorch/audio/blob/main/examples/tutorials/forced_alignment_tutorial.py

SPPAS : https://sppas.org/

Wave2vec : https://ai.meta.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/

WebMAUS :

https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic

WhisperX - Time-Accurate Speech Transcription with Word-Level Timestamps : https://github.com/m-bain/whisperx