COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 641
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par

THIBAULT BAÑERAS-ROUX

Analyse et compréhension de l'évaluation des systèmes de reconnaissance automatique de la parole : vers des métriques intégrant la perception humaine

Thèse présentée et soutenue à « Lieu », le « date »
Unité de recherche : Laboratoire des sciences du numérique à Nantes (LS2N)
Thèse N° : « si pertinent »

Rapporteurs avant soutenance :

Irina ILLINA — Maîtresse de Conférences, Université de Lorraine (LORIA/INRIA)
Cyril GROUIN — Maître de Conférences, Université Paris-Saclay (LISN)

Composition du Jury :

Président :	Prénom NOM	Fonction et établissement d'exercice *(à préciser après la soutenance)*
Examinateurs :	Prénom NOM	Fonction et établissement d'exercice
	Béatrice DAILLE	Professeure des Universités, Nantes Université (LS2N)
	Martine ADDA-DECKER	Directrice de Recherche CNRS, Sorbonne Nouvelle (LPP)
	Benjamin LECOUTEUX	Professeur des Universités, Université Grenoble Alpes (LIG)
Dir. de thèse :	Richard DUFOUR	Professeur des Universités, Nantes Université (LS2N)
Co-enc. de thèse :	Jane WOTTAWA	Maîtresse de Conférences, Le Mans Université (LIUM)
Co-enc. de thèse :	Mickael ROUVIER	Maître de Conférences, Avignon Université (LIA)

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

Titre : Analyse et compréhension de l'évaluation des systèmes de reconnaissance automatique de la parole : vers des métriques intégrant la perception humaine

Mot clés : reconnaissance automatique de la parole, métriques d'évaluation, perception humaine, sémantique

Résumé : De nos jours, le taux d'erreur mot reste la métrique la plus utilisée pour évaluer les systèmes de reconnaissance automatique de la parole (RAP). Toutefois, cette métrique présente des limites en matière de corrélation avec la perception humaine et ne se concentre que sur la préservation de l'orthographe. Dans cette thèse, nous proposons des métriques alternatives qui peuvent évaluer l'orthographe, mais aussi la grammaire, la sémantique ou la phonétique.

Pour analyser la capacité de ces métriques à refléter la qualité des transcriptions du point de vue des utilisateurs, nous avons constitué un jeu de données nommé HATS, annoté par 143 sujets francophones. Chaque annotateur a examiné 50 triplets, composés d'une transcription de référence manuelle et de deux hypothèses issues de différents systèmes de RAP, afin de déterminer quelle hypothèse était, selon eux, la plus fidèle.

En calculant le nombre de fois où une métrique est d'accord avec les choix des annotateurs, on obtient une mesure de sa corrélation avec la perception humaine. Ce corpus permet ainsi de hiérarchiser les différentes métriques selon le jugement d'un lecteur humain. Nos résultats montrent que SemDist, une métrique basée sur les représentations sémantiques de BERT pour comparer deux phrases, s'avère la plus pertinente pour évaluer les transcriptions du point de vue perceptif. À l'inverse, le taux d'erreur mot figure parmi les moins performants, ce qui interroge sur son utilisation systématique comme unique métrique, alors que d'autres alternatives prometteuses sont largement négligées.

Nous avons également mené une étude sur l'impact de plusieurs hyperparamètres des systèmes de RAP, tels que le réordonnancement des hypothèses avec des modèles de langage, la tokenisation et l'utilisation de modules SSL. En plus de l'analyse qualitative de ces paramètres, nos recherches révèlent que chaque métrique évalue des aspects différents des systèmes et que les métriques ne convergent pas toujours dans leur classement des systèmes. Cette disparité, combinée aux limites du taux d'erreur mot, justifie l'utilisation de plusieurs métriques pour une évaluation plus fine.

Enfin, nous proposons une approche innovante pour rendre les métriques sémantiques plus interprétables. Ces métriques se contentent actuellement de fournir des scores bruts basés sur des similarités cosinus entre représentations sémantiques, rendant difficile l'interprétation des erreurs. Afin de rendre ces mesures plus accessibles, nous avons développé une méthode appelée minED, qui vise à améliorer la compréhension et la transparence de l'évaluation des systèmes de RAP, en prenant en compte les aspects sémantiques ainsi que la perception humaine. De plus, une variante de cette méthode permet d'évaluer la gravité de chaque erreur pour la compréhension globale d'une phrase, offrant ainsi des informations précieuses non seulement sur les erreurs des systèmes, mais aussi sur le fonctionnement des métriques elles-mêmes.