Application BASTRI

Fiches Equipes

MULTISPEECH (SR0707BR)

La parole multimodale en interaction
MULTISPEECH (SR0656HR) MULTISPEECH

Statut: Décision signée

Responsable : Slim Ouni

Mots-clés de "A - Thèmes de recherche en Sciences du numérique - 2024" : *Aucun mot-clé.*

Mots-clés de "B - Autres sciences et domaines d'application - 2024" : *Aucun mot-clé.*

Domaine : Perception, Cognition, Interaction
Thème : Langue, parole et audio

Période : 01/07/2015 -> 30/09/2028
Dates d'évaluation : 02/10/2019 , 08/10/2015 ,

Etablissement(s) de rattachement : CNRS, U. DE LORRAINE
Laboratoire(s) partenaire(s) : LORIA (UMR7503)

CRI : Centre Inria de l'Université de Lorraine
Localisation : Centre Inria de l'Université de Lorraine
Code structure Inria : 051026-2

Numéro RNSR : 201421147E
N° de structure Inria: SR0707BR

Présentation

MULTISPEECH est une équipe-projet commune de l'Université de Lorraine, d'Inria et du CNRS. Elle fait partie du département D4 « Traitement automatique des langues et des connaissances » du LORIA.

Multispeech considère la parole comme un signal multimodal avec différentes facettes : acoustique, faciale, articulatoire, gestuelle, etc. L'objectif général de Multispeech est d'étudier l'analyse et la synthèse des différentes facettes de ce signal multimodal et leur coordination multimodale dans le contexte de l'interaction humain-humain ou humain-ordinateur.

Le programme de recherche est organisé selon les trois axes suivants :

- Apprentissage efficace en termes de données et préservation de la confidentialité.
- Extraction d'informations à partir de signaux vocaux.
- Parole multimodale : génération et interaction.

Axes de recherche

1 — Apprentissage efficace en termes de données et préservation de la confidentialité

Notre recherche porte principalement sur la conception de modèles d'apprentissage automatique pour les données de parole multimodales, en se concentrant sur des scénarios avec des données vocales limitées et des contraintes de confidentialité. Nous visons à intégrer les connaissances du domaine en combinant l'apprentissage profond avec l'expertise acoustique pour créer des modèles génératifs qui décrivent la distribution de probabilité des signaux vocaux et audio. Nous nous concentrons également sur l'apprentissage à partir de peu ou pas de données étiquetées, visant à apprendre des représentations qui désenchevêtrent les attributs de la parole en utilisant des méthodes d'apprentissage de représentation non supervisées, des branches supervisées et des branches adverses. Un aspect significatif de notre recherche est la préservation de la confidentialité en transformant la parole pour cacher l'identité des utilisateurs et d'autres attributs sensibles tout en maintenant les attributs nécessaires pour des tâches comme la reconnaissance automatique de la parole. Nous développons également des attaques fortes pour évaluer la confidentialité et travaillons sur la dissimulation des identifiants personnels et des attributs sensibles dans le contenu linguistique.

Contact

- **Responsable :** Slim Ouni
- **Tél :** 03.83.59.20.22
- **Secrétariat Tél :**

En savoir plus

- Site de l'équipe
- Site sur inria.fr
- Site du responsable
- Derniers Rapports d'Activité : 2015 , 2016 , 2017 , 2018 , 2019 , 2020 , 2021 , 2022 , 2023

Documents sur la structure

- Intranet
- Privés

Décisions

- 11067 (20/07/2015) : création
- 11683 (13/06/2016) : prolongation
- 14340 (30/07/2020) : prolongation
- 15592 (26/09/2022) : cessation du responsable
- 15593 (26/10/2022) : nomination responsable
- 16995 (26/04/2024) : prolongation
- 17293 (18/09/2024) : création

Localisation

- **Adresse postale :** Centre Inria de l'Université de Lorraine, 615 rue du Jardin Botanique, 54600 Villers-lès-Nancy France
- **Coordonnées GPS :** 48.666, 6.157

2 — Extraction d’informations à partir de signaux vocaux

Dans notre recherche, nous visons à extraire des informations pertinentes à partir de signaux vocaux dans des conditions réelles, en nous concentrant sur le contenu linguistique, l’identité et les états du locuteur, et l’information sur l’environnement vocal. Nous utilisons la reconnaissance vocale pour extraire des informations linguistiques et développons des méthodes pour améliorer sa robustesse dans des scénarios réels. Nous travaillons également sur l’analyse du contenu sémantique, la détection des discours de haine et l’extraction des informations phonétiques et prosodiques. L’identification du locuteur est cruciale pour la personnalisation de l’interaction humain-machine, et nous visons à reconnaître les états du locuteur comme l’émotion et le stress. Enfin, nous développons des méthodes de détection d’événements audio et explorons la modélisation des scènes sonores ambiantes, ainsi que l’inférence des propriétés acoustiques de l’environnement.

3 — Parole multimodale : génération et interaction

La parole est considérée comme une entité multimodale, et nous étudions la modélisation et l’analyse multimodales, la génération de la parole multimodale et l’interaction. Nous nous concentrons sur l’interaction, la fusion et la synchronisation des modalités pour un seul locuteur et entre les locuteurs dans une conversation. Nous visons à améliorer l’intelligibilité et la qualité de la parole grâce à l’amélioration de la parole audiovisuelle et aux méthodes d’apprentissage pour les données multimodales.

En termes de génération de parole multimodale, nous utilisons des techniques de synthèse articulatoire, acoustique et audiovisuelle. Nous nous appuyons sur la modélisation 2D et 3D de la dynamique du conduit vocal à partir de données d’IRM en temps réel, en considérant la génération du conduit vocal complet et la prédiction de l’ouverture de la glotte. Nous nous concentrons également sur la synthèse vocale audiovisuelle, y compris l’animation du visage liée à la parole et à l’expression faciale, et la modélisation de l’expressivité pour la synthèse acoustique seule et audiovisuelle.

Nous abordons également l’interaction où nous considérons les composants multimodaux utilisés pendant l’interaction par le locuteur et l’auditeur. Notre objectif est de générer simultanément la parole et les gestes par le locuteur et les gestes régulateurs pour l’auditeur. Nous considérons différents composants de dialogue : compréhension du langage parlé, gestion du dialogue et génération de langage naturel. Nous considérons le dialogue dans un contexte multimodal et rompons le schéma classique de gestion du dialogue pour tenir compte de l’évolution de l’interlocuteur pendant la réponse du locuteur.

Relations industrielles et internationales