Technologies vocales

LETEILLIER Marion
Numéro d'étudiant : 12400026

UFR LLASIC
Département d'Informatique intégrée en Langues, Lettres et Langage

Master 1 Mention Sciences du Langage Parcours Industries de la Langue

Enseignante responsable : Véronique Aubergé

Année universitaire
2024 – 2025

Facultatif : question bonus + 2 points

Quels sont les nouveaux défis de la synthèse vocale dans l'avenir proche ? Quelle(s) type(s) de techniques algorithmiques risque(nt) d'être le support pour répondre à ces nouveaux défis ?

Les nouveaux défis de la synthèse vocale dans l'avenir proche sont multiples. Tout d'abord, les systèmes d'aujourd'hui sont souvent perçus comme assez monotones. La voix utilisée dans l'assistant vocal Alexa (Amazon) ne fait pas varier sa tonalité, ne laisse pas transparaître d'émotions. Une sorte de barrière se créée avec nous, utilisateurs, car nous ne pouvons pas établir de lien profond. Pour améliorer cette relation, il faudrait que cette voix adapte son ton en fonction du contexte et de l'utilisateur. Le contexte de l'échange est aussi important selon le contenu. De plus, les accents régionaux tout comme les variations phonétiques sont des défis majeurs. Lorsque, par exemple, un utilisateur avec un fort accent marseillais dicte un mot, le système en question ne reconnaît pas forcément. Il faudrait qu'il soit capable de comprendre n'importe quel accent. Nous pouvons aussi rajouter que les deepfakes, plus particulièrement vocaux, sont un défi éthique à prendre très au sérieux. Récemment, la voix imitant celle d'un PDG utilisée pour des transferts d'argent en est le parfait exemple. Une régulation est nécessaire pour des mesures de sécurité.

Les techniques algorithmiques risquant d'être le support pour répondre à ces nouveaux défis sont aussi nombreux. Les modèles avancés basés sur le deep learning (WaveNet, Tacotron 2) ont été conçus pour produire des voix très naturelles, grâce à des milliers d'échantillons vocaux. Les chuchotements (pas de vibration des cordes vocales) et les émotions sont encore quelque chose de difficile à cerner. La combinaison de règles linguistiques explicites et de modèles neuronaux pourrait être aussi un support pour distinguer les homophones (règles contextuelles). Il y a aussi le front-end qui pourrait être amélioré pour éviter les désagréments suive à un mot mal « phonétiser ». Google utilise le transfert d'apprentissage pour entraîner des modèles vocaux pour des langues où nous disposons très peu de ressources, et cela pourrait s'étendre pour des langues régionales.

Partie 1 (sur 13 points)

Q1 : (3 points)

• Quelles sont les différentes briques de technologies vocales, dites pour chacune ce qu'elle prend en entrée et ce qu'elle livre en sortie.

Il y a différentes briques de technologies vocales. Nous avons la synthèse vocale (Text to speech) qui prends en entrée un texte, qu'il soit interprété ou généré avec en sortie le signal audio de ce texte qui simule une voix humaine. La phonétisation est aussi importante, avec un mot ou texte en entrée pour un convertissement en symboles phonétiques en sortie. La reconnaissance vocale automatique (automatic speech recognition) prend en entrée un signal audio de paroles humaines avec une transcription textuelle en sortie. Il y aussi le vocodeur, qui est une représentation numérique acoustique du signal vocal avec en sortie ce signal audio (Hifi gan).

Q2 : (1 point)

• Qu'est-ce que la synthèse à trous ?

La synthèse à trous ne génère pas l'entièreté du signal vocal, mais complète ce signal par des segments manquants. Le contexte est pris en compte pour plus de cohérence et ainsi générer les segments manquants. Cela permet d'améliorer la qualité audio par exemple.

Q3 : (1 point)

• Dans un système de synthèse vocale à partir du texte (TTS) que prend en entrée et délivre en sortie le Front End ? Idem le Back End ?

Dans un système de synthèse vocale à partir du texte (TTS), le Front End prend en entrée un texte écrit fourni par l'utilisateur et en sortie une représentation phonétique avec de la prosodie. (paramètres vocaux). Le Back End prend en entrée cette représentation du Front End avec en sortie le signal vocal sous forme de parole. Ils forment un duo essentiel.

Q4 : (3 points)

• Quelles sont les deux grandes approches techniques dans les systèmes TTS commercialisés actuellement ? Quels sont les avantages et les inconvénients de ces deux approches ?

Les deux grandes approches techniques dans les systèmes TTS commercialisées actuellement sont la synthèse par concaténation et la synthèse basée sur le machine learning. Pour la synthèse par concaténation, on utilise des échantillons de voix assemblés pour produire un signal vocal. Les avantages sont que les voix sont naturelles, engendrant une expérience plutôt réaliste, et réutilisable à l'infini. Les inconvénients sont la difficulté de transitions entre les segments, avec aussi des intonations complexes à ressortir. Acapella est une entreprise qui confirmait que cette approche était assez limitée pour des textes nouveaux. La synthèse basée sur le machine learning avec des DNN génère des voix à partir de textes. Les avantages sont que l'on peut apprendre à synthétiser des styles variés (formel, informel) avec une reproduction de la prosodie, des émotions. La parole est assez fluide. Mais les inconvénients sont que cela engendre une grande puissance de calcul et une dépendance aux données. Wavenet est un exemple de cette synthèse.

Q5 : (1 point)

Si on enchaîne un système de génération de texte à partir de concepts et un système TTS, qu'obtient-on ? A quels types d'applications est-ce, ou pas, adapté ?

Si on enchaîne un système de génération de texte à partir de concepts et un système TTS, nous obtenons une voix synthétique. Elle va pouvoir s'adapter au contexte à partir des informations qu'elles reçoie avec une certaine cohérence. Pour la génération de texte, Alexa ou Google Assistant en sont le parfait exemple pour générer des réponses de chatbots. Puis avec un système TTS, cela se converti en voix artificielle. C'est adapté pour répondre aux questions des utilisateurs par le biais d'assistants vocaux comme Alexa, ou pour de la création de contenu (générer des histoires pour les enfants). Mais inadapté pour répondre à un utilisateur en souffrance émotionnelle où un ton non adapté peut entraîner de la frustration.

Q6 : (4 points)

• Pour un système de dialogue en condition réellement spontanée, en quoi et par quoi les systèmes actuels sont-ils encore limités ? Selon vous les solutions seront elles dans les techniques, les data, ou les connaissances linguistiques ?

Pour un système de dialogue en condition réellement spontanée, les systèmes actuels sont encore limités. Les voix artificielles manquent de flexibilité pour répondre aux différents états émotionnels des utilisateur, créant une certaine distance interactionnelle. Si un utilisateur exprime quelque chose et que le système répond de manière monotone, cela peut engendre une tension au lieu de l'apaiser. Les

assistants vocaux ne relient pas aussi les différentes requêtes des utilisateurs et limite ainsi la conversation. Si un utlisateur plusieurs questions à la suite, le système ne peut pas forcément faire le lien avec la 1ère requête s'il la repose de manière non explicite plus tard (c'est quoi **sa** capitale ?). Les langues peu dotées manquent pour entraîner ces modèles de manière efficace, mettant une barrière à la diversité linguistique. Un dialecte en particulier peut être mal compris par manque de données.

Les solutions seront dans les trois. Dans les techniques pour améliorer les modèles de deep learning avec par exemple Hifi gan qui améliorent la fluidité, précision des réponses avec une meilleure anticipation des besoins utilisateurs. Dans les données, par les corpus avec de meilleures qualités (variations linguistiques). Et dans les connaissances linguistiques avec la théorie de l'esprit artificielle pour améliorer la compréhension des intentions de l'utilisateur.

Partie 2 (sur 7 points)

Q1 : (1 point)

• Quels sont les 2 ordres qui régissent l'application des règles dans le langage de transcription orthographe phonétique vu en cours ?

On lit la grammaire de haut en bas pour les règles. Entre les 2 « + », c'est ce qu'on doit prendre en entrée et réécrire en sortie suivant le contexte. Dans notre algorithme de transcription, quand il va appliquer les règles, il va d'abord prendre la première. S'il l'applique, il ne passe pas à la suivante. Il continu à l'endroit où on sera à transcrire dans le texte. Si la première règle ne s'applique pas, il passe à la suivante, et ainsi de suite (car elle est dépendante de ce que j'ai fait avant). Pour la façon dont les règles seront appliquées dans le texte, c'est de gauche à droite.

Q2 : (2 points)

Sur la partie de grammaire ci-dessous :
Comment sont transcrits les *p* dans *appel*

La transcription correcte de appel est [apEl], si cette grammaire ne produit pas cette transcription correcte, corriger la grammaire.

+pp+=[p] pour la transcription des p

Q3 : (4 points)

Y a-t-il une ou des règles inacessibles et/ou redondantes ? Désigner la/les et proposer une solution sans règle inaccessible ni redondante

Si nous prenons les règles du « p » avec (p)+p+=[] et +p+=[p], nous remarquons qu'elles sont redondantes. **+p+(p)=[]** est en plus inaccessible.

(p)+p+=[] **redondantes**

+p+(p)=[] **inaccessible**

+p+=[p]

Pour une solution sans règle inaccessible ni redondante on a +pp+=[p]