Synthèse vocale

14 décembre 2020

1 Synthèse vocale

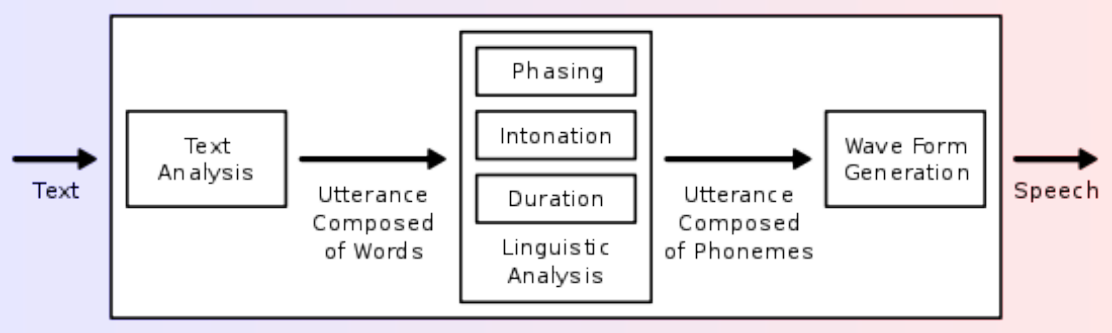

FIGURE 1 – Processus de synthèse vocale

1.1 Synthèse par formants

La synthèse vocale par formants consiste donc en la génération d'un spectre sonore grâce aux formants, qui représentent les caractéristiques uniques de chaque phonème, et à une suite de règles permettant d'établir les liens et l'évolution entre les formants.

Cette méthode répondait aux limitations techniques de son époque (années 60) : puisque la génération de la voix est entièrement mathématique, aucune base de donnée n'est nécessaire. L'espace mémoire demandée est donc très faible. Cet avantage est toujours d'actualité : bien que la synthèse par Deep Learning permet de synthétiser des voix quasi similaires à la voix humaines, ils demandent un espace mémoire (dure comme vive) important. A moins d'y accéder via le réseau, ces modèles ne peuvent pas être intégrés à tous les systèmes, contrairement aux modèles de synthèses par formants.

Il existe deux méthodes de synthèse par formants : la synthèse en série et la synthèse en parallèle.

Le modèle en série est celui qui correspond le plus au conduit vocal. Il est composé d'une série de résonateur, qui vont chacun synthétiser une fréquence. Ces dernières sont accumulées au fur et à mesure pour produire le son complexe qui correspond à la voyelle. Ce modèle est cependant beaucoup moins adapté à la production des consonnes. A l'inverse, le modèle en parallèle aligne les résonateurs en parallèle. Chaque résonateur est associé à un module qui contrôle l'amplitude de la fréquence. Ces dernières sont ensuite additionnées pour former l'onde complexe. Bien que

demandant plus de paramètres à gérer, ce modèle produit de meilleurs consonnes que le modèle en série.

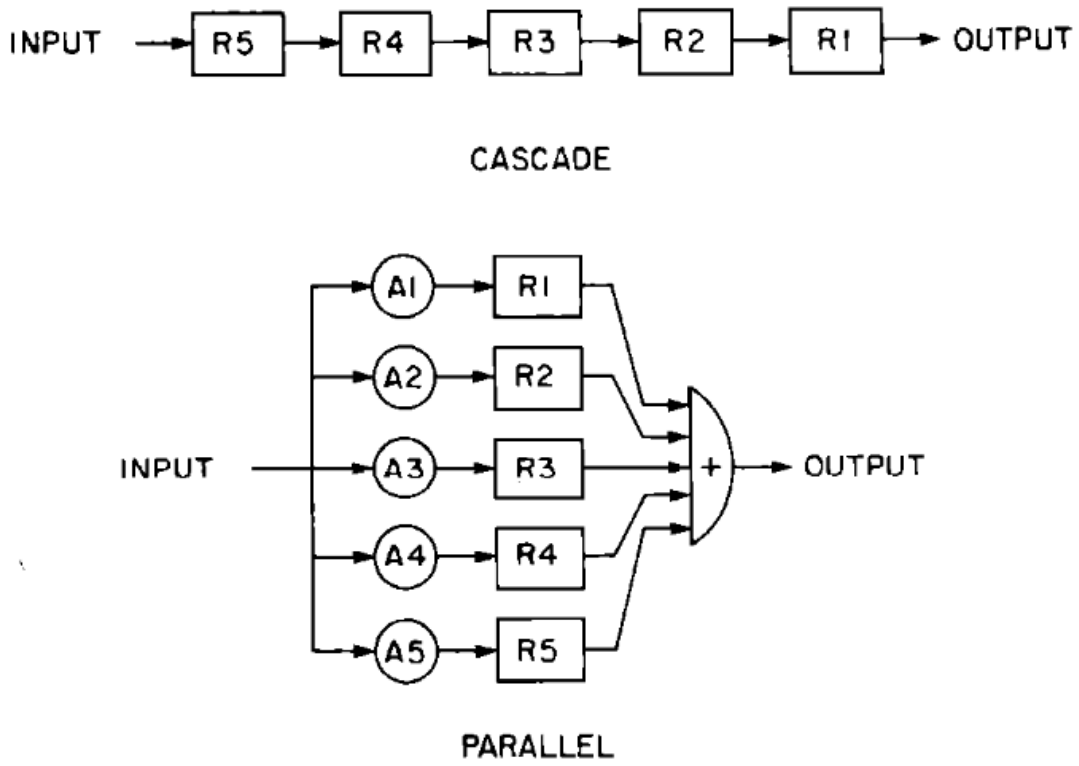

FIGURE 2 – Synthèse par formants en série (haut) et en parallèle (bas)

Le synthétiseur de Klatt vu au cours précédent combine les avantages de ces deux méthodes.

1.2 Synthèse par concaténation d'unités

Les progrès en termes de computation et d'espace mémoire ont permis de développer de nouveaux systèmes de synthèse vocale, qui reposent sur l'utilisation de fichiers audios. Ainsi, plutôt que de synthétiser mathématiquement les sons, ces modèles associent différents fichiers audios pour composer une phrase, qui sont contenus dans une base de données. Selon les méthodes, différentes unités peuvent être utilisées : des phonèmes, des paires de phonèmes (on parle de synthèse par **diphones**), des mots ou groupe de mots entiers. Ces deux derniers cas sont notamment utilisés lorsque cette synthèse est utilisée dans un **domaine spécifique** (ex : annonces de gare).

La base de données est créée à partir d'un corpus audio. Celui-ci est d'abord divisé selon la forme des unités choisies, généralement à l'aide d'un outil de reconnaissance vocale. Ces unités sont ensuite indexées selon leur paramètres audio (fréquence fondamentale, spectre, ...) mais aussi linguistique (position syllabique, phonèmes environnants).

Lors de la synthèse, le programme divise la phrase suivant le type d'unité choisies puis sélectionne les unités correspondantes dans la base de données. Un algorithme tel qu'un arbre de décision permet d'identifier les candidats les plus probables. Un système de règles vient ensuite

ajuster les paramètres tels que l'intensité ou l'intonation pour lisser l'association des unités.

A la suite de la synthèse par formants, la synthèse par concaténation produit une voix de très bonne qualité. Cette qualité est relative à la finesse de la base de données : plus les unités choisies seront petites et plus la qualité sera fine. Cependant, préparer une telle base de données demande un coup important, tant financier qu'en terme d'espace mémoire.

1.3 Synthèse articulatoire

La synthèse articulatoire est un type particulier de synthèse. Plutôt que de synthétiser la voix via les formants ou des fichiers audio, elle consiste à modéliser l'appareil phonatoire. L'utilisateur produit les sons en manipulant les différents élément de l'appareil. C'est donc un modèle beaucoup plus exploratoire et expérimental, qui est dédié à la recherche.

1.4 Synthèse basée sur les modèles de Markov cachés (MMC) et Deep Learning

Un modèle de Markov caché (HMM) est un modèle statistique permettant de générer des séquences d'unités discrètes. A partir d'un corpus audio d'entraînement, le modèle apprend à modéliser statistiquement la fréquence fondamentale, le spectre et la prosodie du discours. Une fois ces paramètres appris, le modèle est capable de synthétiser la voix, sans avoir recours à une base de données.

Les modèles cachés de Markov ont été utilisés dès les années 70 dans de nombreux domaines demandant des génération de séquences, tels que la reconnaissance de caractères, l'analyse en parties du discours, traduction automatique.

Les performances obtenues par les HMM sont restés inégalées jusque dans les années 2010, avec l'arrivée du Deep Learning et des modèles séquentiels tels que LSTM. En 2016, DeepMind sort WaveNet, un des premiers modèles de synthèse vocale en DeepLearning. Celui-ci est rapidement suivi par Tacotron (Google) et VoiceLoops (Facebook). En réalité, cette méthode est constituée de deux modèles : le premier modèle apprend à faire correspondre une séquence de caractères (mots ou phonèmes) à un spectrogramme. Le second, appelé vocoder, apprend à transformer le spectrogramme en audio.

2 Reconnaissance de la parole

Malgré les différentes approches, le principe reste identique : une première étape consiste à traiter le signal de telle sorte à le diviser en petites fenêtres (de l'ordre de 10Ms à 30ms). Deuxièmement, un modèle (généralement HMM ou Deep Learning) identifie les termes du lexique les plus probables pour ce signal. Enfin, un dernier modèle vient concaténer chaque unité, en filtrant les résultats du modèle précédent en fonction du contexte.

Les premiers systèmes développés entre les années 50 et 70 permettaient au mieux de reconnaître les 10 premiers chiffres ou bien une centaine de mots au maximum. Ceux-ci reposaient principalement sur l'algorithme de **Déformation temporelle dynamique** (*Dynamic Time Warping*), qui calcule la similarité entre deux séquences qui peuvent varier au cours du temps. Il permet par exemple d'identifier deux énoncés comme identiques, même si l'un est prononcé plus rapidement que l'autre.

Les premiers vrais modèles capable de reconnaître plusieurs milliers de mots reposaient également sur les modèles cachés de Markov, comme pour la synthèse vocale. Ceci a permis la création de logiciels tels que Dragon pour la dictée vocale, ainsi que des services téléphoniques de commandes vocales. Comme pour le Deep Learning, les performances obtenues par les HMM

sont restés inégalées jusque dans les années 2010, avec l'arrivée du Deep Learning et des modèles séquentiels tels que LSTM. Les services tels que Siri, Cortana, Google Assistant, reposent ainsi sur du Deep Learning et non des HMM. Cependant, comme pour la synthèse vocale, les modèles de reconnaissance de la parole en Deep Learning sont les plus performants, mais aussi les plus massifs. Ils sont donc difficiles à intégrer à des systèmes. Pour cette raison, les modèles HMM sont préférables lorsque l'on ne peut avoir accès à Internet.

3 Logiciels

eSpeak est un logiciel libre (pour Linux et Windows) de synthèse vocale par formants. Il permet la synthèse de plus de cent langues (dont des langues inventées). Il permet également certains traitement linguistiques du textes, dont la transcription en phonétique.

MBROLA est un projet international qui produit et fournit des bases de données de diphones dans plusieurs langues. MBROLA n'est pas un système de synthèse vocale en soit : ces fichiers doivent être utilisés avec d'autres systèmes, tels que eSpeak. A l'origine non-libre de droit pour des usages commerciaux, le projet a été ouvert en 2019 en donnant accès à l'outil de création de voix.

CMU Sphinx est un ensemble d'outils de reconnaissance de la parole développés à l'université de Carnegie Mellon. Ces outils reposent sur des modèles HMM, et sont facilement intégrables à toute sorte de projets. Ces outils permettent également d'entraîner soi-même ses modèles.

Mozilla Common Voice est un projet visant à recueillir des voix et à permettre le développement d'outils de reconnaissance et de synthèse vocale non-propriétaires. Les projets DeepSpeech (reconnaissance vocale) et Mozilla TTS (synthèse vocale) font partie de ce projet. Ils reposent tous deux sur l'entraînement de modèles Deep Learning

— eSpeak : http ://espeak.sourceforge.net/
— MBROLA : `https://github.com/numediart/MBROLA`
— CMU Sphinx : https ://cmusphinx.github.io/
— DeepSpeech : https ://github.com/mozilla/DeepSpeech
— Mozilla Tacotron : https ://github.com/mozilla/TTS