TTS : Text To Speech

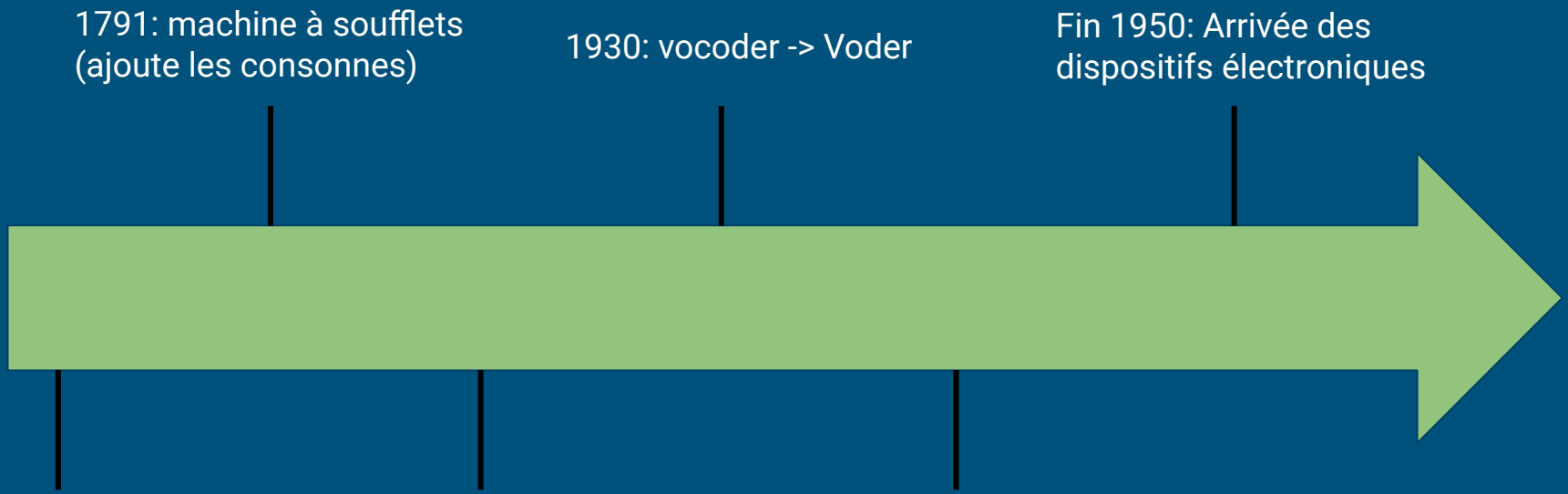
Granger Oscar - Cosotti Kévin

Sommaire

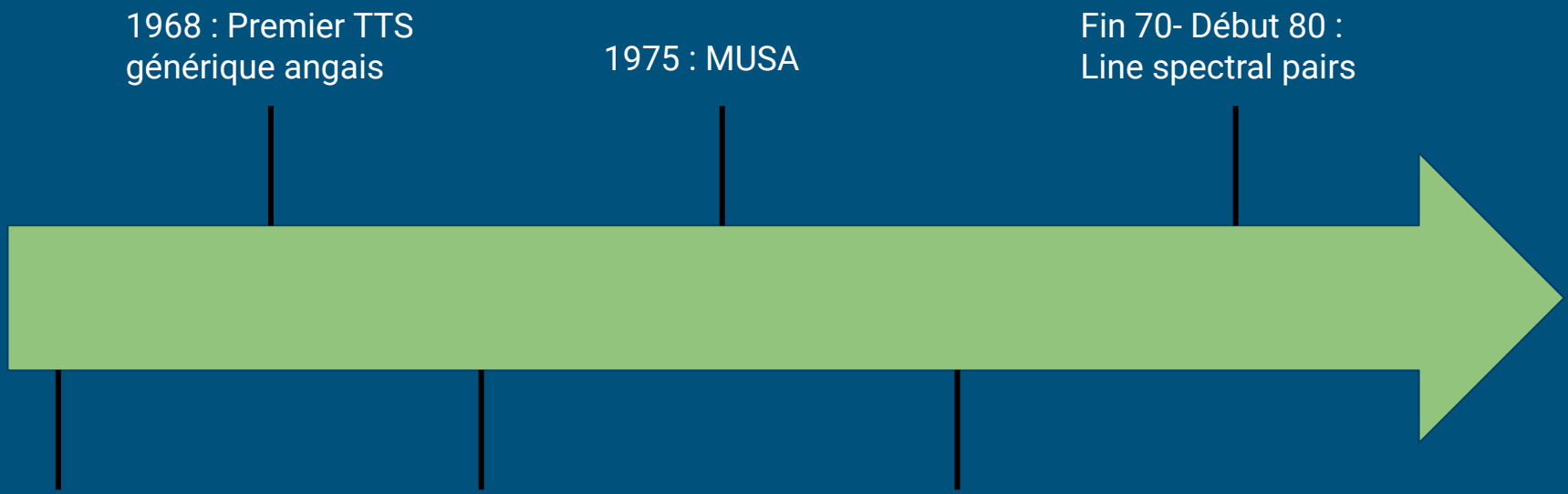
- Histoire
- Enjeux technologique
- Transcription phonétique
- Synthèse vocal
- Démo
- Conclusion

DANS LE PASSÉ

L’ère mécanique



L’ère électronique



Enjeux de la synthèse vocal

La prosodie

Ce qui rend “naturel” la parole.

L’accentuation

Le Ton

Le débit

Les pauses

Le rythme

Transcription phonétique

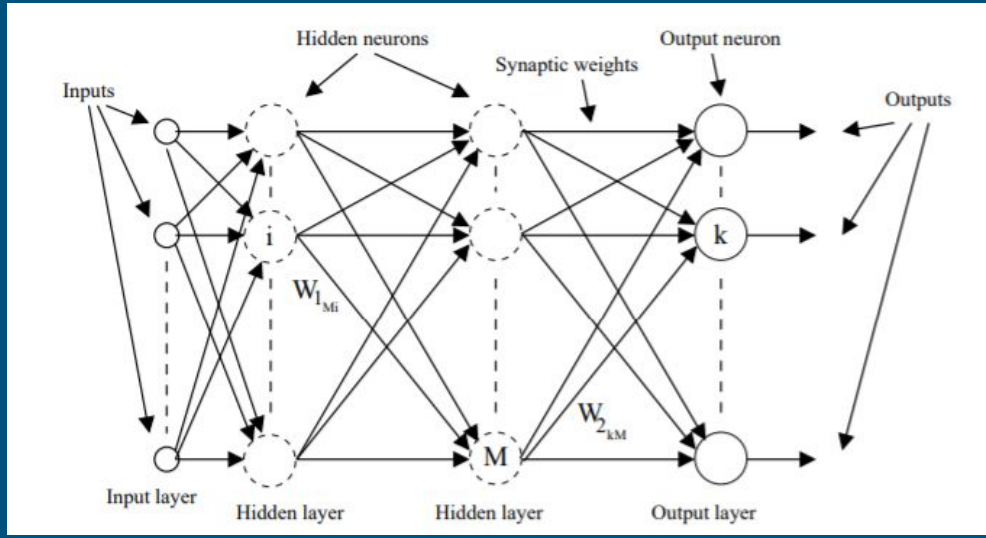
- Aussi appelé Text-to-phoneme (TTP) ou Grapheme-to-phoneme (GTP)
- Convertit un mot ou ensemble de mots en leur équivalent phonétique

  ex : Hello world -> hɛˈləʊ wɜːld

- Souvent utilisé en complément d’un dictionnaire
- Requiert l’usage de fonction non-linéaire -> réseau de neurones

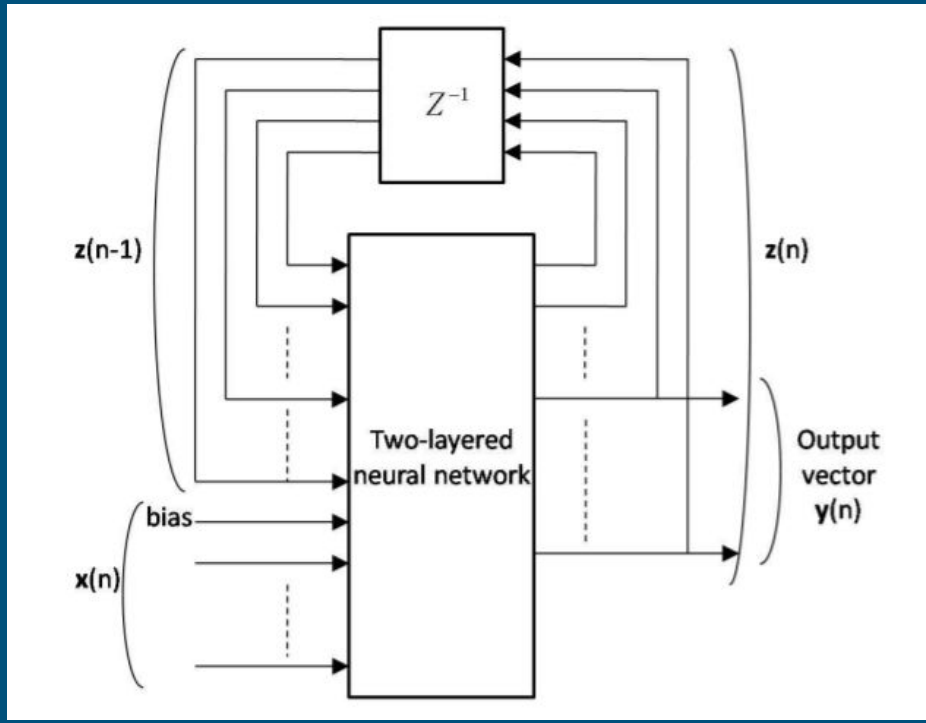
Multilayer Perceptron (MLP)

- Error back-propagation with momentum
- Taux d’apprentissage pour contrôler la vitesse de convergence et la stabilité du modèle



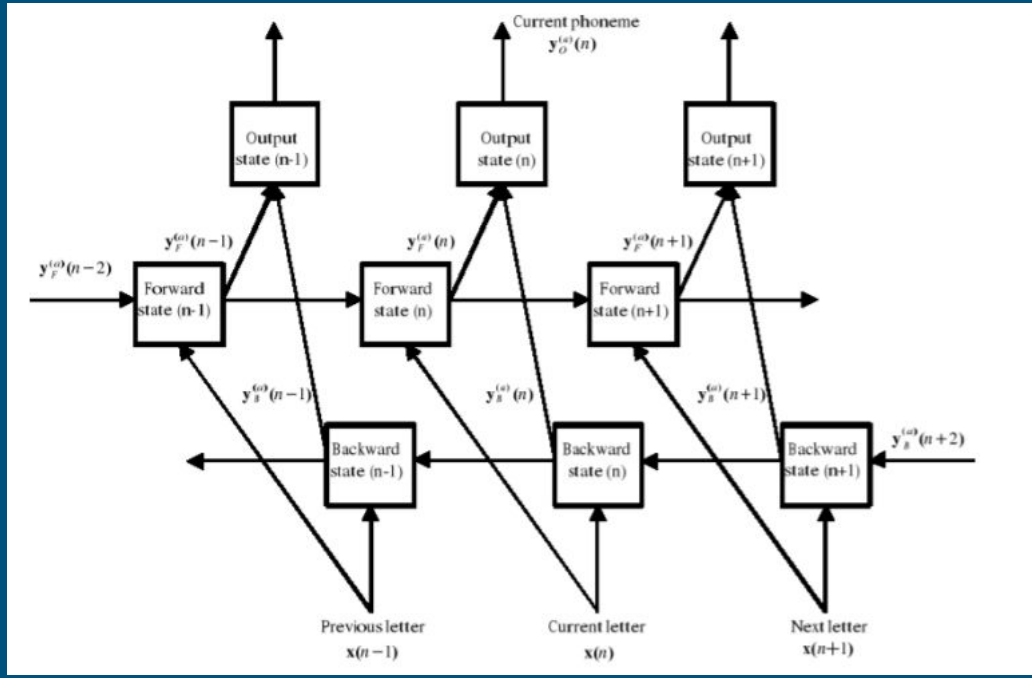
Recurrent Neural Network (RNN)

- Boucles de feedback (global ou local)
- Crée une dépendance temporelle



Bidirectional Recurrent Neural Network (BRNN)

- Sépare l’état des neurones (état avant/après)
- Permet d’avoir le contexte avant et après l’état actuel



Synthèse Vocal

3 Grands types de Synthèse

Synthèse par concaténation :

- plutôt “naturel”, grosse base de donnée

Synthèse par formant

- Très modulable, très “robotique”

Deep Learning

- Très naturel, coûteux en calcul

Synthèse par concaténation

Concaténation de de son pré-enregistré

- Par brique élémentaire :

  très modulable, “glitch auditif” fréquent

- Par mot ou groupe de mots :

  peu modulable sans faire exploser la base de donnée

Synthèse par concaténation

- par diphone :

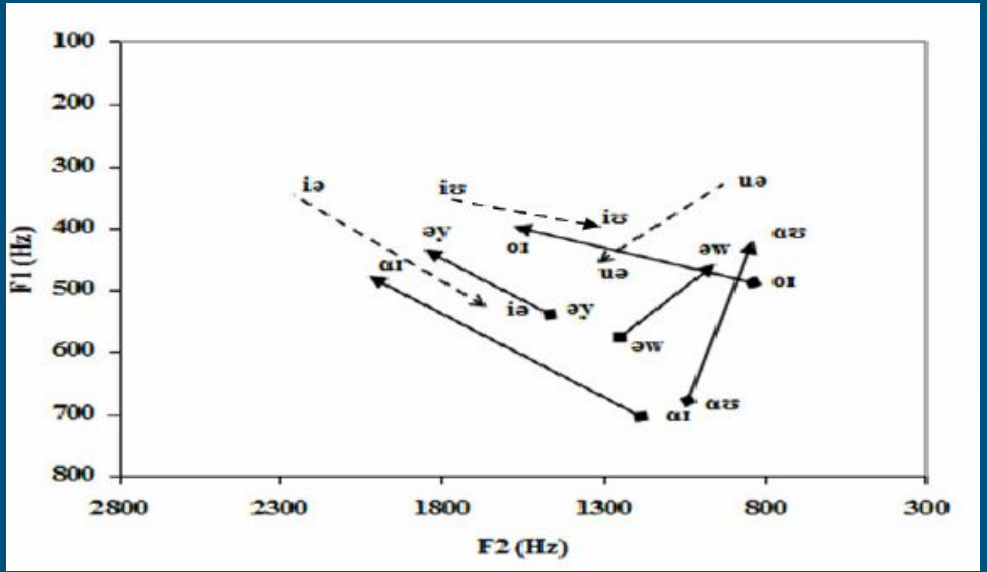
  Mix de concaténation et de formant, peu utilisé

Synthèse par Formant

Formant : Fréquence sonore fondamentale spécifique à un son du langage parlé.

Son généré directement par ordinateur

utilisable en embarqué



Deep learning

Pas besoins de passer par des phonèmes

Très performant

imite la voix de personne réel

aucunement modulable

ton en décalage possible avec le propos

TRANSITION

Démo

Mbrola :

Synthétisation par diphone

uniquement “phoneme to speech”

gratuit pour les utilisations non-lucrative

Fait par des francophones

Démo - Exemple Bonus

“Fukkireta” - Utauloid

Obama Deep Fake -Tacotron 2

Conclusion

- Technologie “fini”
- Plusieurs approches
- Réponds à différents besoins
- Plus d’automatisation = moins de contrôle humain