Fouille de données vocales sur les langues africaines

Maguelonne TEISSEIRE, TETIS Inrae
Damien NOUVEL, ERTIM Inalco
Valentina FEDCHENKO, ERTIM Inalco

Dan HOU

Contexte & Objectif :

– Veille épidémiologique basée sur les médias africains (radios locales)

– Langues peu dotées (wolof, peul, ewondo) : peu de ressources ASR

– Objectif : fouille de données ciblée (thèmes agricoles, sanitaires)

Méthodologie : transcription automatique (ASR) + fouille de données ciblée

Audio → [ASR Model] → Texte → [Classification mBERT] → Thématique

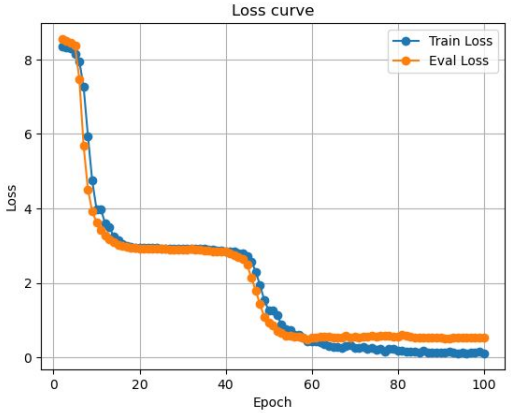
Données : corpus VoxPopuli (français) – audio + transcription, corpus text agriculture vs. non-agriculture

Modèles de transcription testés : Wav2Vec2, Whisper, HuBERT

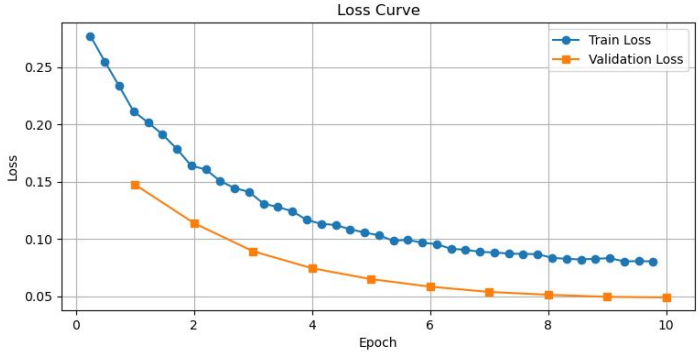
Classification des transcriptions : mBERT pour identifier la thématique

Expérimentations :

🎧 Audio → 📄 Texte :
facebook/wav2vec2-large-xlsr-53



📄 Texte → Classification :
mBERT



gold_text : et de donner ainsi des réponses à de nombreuses questions dans tous les domaines politiques liés au tourisme.

prédiction : i doner ainsi des réponses sur de nombreuses questions dans tous les domaines politiques li liés au tourisme.

(0 : non-agriculture)

WER (Word Error Rate) = 27.78%

Résultats actuels :

– Modèle ASR wav2vec2 fine-tuning
– Modèle mBERT efficace

Perspectives : VoxPopuli → ESTER2 (français) : bruit, musique, parole spontanée

↑

Étape avant langues africaines (Wolof, Peul…)