Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Étude de la variabilité de la prononciation dans des plongements de grands modèles audio. Effets du locuteur et de la L1 en français L2

Maxime Fily[1] Martine Adda-Decker[2] Guillaume Wisniewski[3]
(1) INALCO / ERTIM – Équipe de Recherche en Textes, Informatique, Multilinguisme
(2) Université Sorbonne Nouvelle / LPP – Laboratoire de Phonétique et de Phonologie
(3) Université Paris-Cité / LLF – Laboratoire de Linguistique Formelle
maxime.fily@inalco.fr, martine.adda-decker@sorbonne-nouvelle.fr,
guillaume.wisniewski@u-paris.fr

RÉSUMÉ

La variation en parole française native et non-native est traitée avec une méthode *low-resource* basée sur une comparaison des représentations acoustiques wav2vec2/XLSR-53 brutes, utilisant des transcriptions phonétiques fines effectuées par annotateurs experts. Les méthodes de z-scoring et de normalisation temporelle sont explorées pour évaluer les informations phonétiquement analysables. En adaptant le *Dynamic Time Warping* aux plongements, nous comparons des enregistrements phonologiquement similaires de locuteurs natifs et non-natifs et l'effet sur les plongements et les MFCCs de la variabilité inter- et intra-locuteur / de parole native vs. non-native. Ce travail sur les représentations montre que les représentations sont locuteur-dépendantes. Afin de mieux aborder la variabilité de la prononciation L2, une normalisation temporelle permet de séparer les facteurs de fluidité et de précision dans la prononciation L2. Cela montre que wav2vec2 contient des informations phonétiques fines telles que la prononciation non-native. Nous montrons par ailleurs que les plongements encodent temporellement l'information phonétique.

ABSTRACT

Investigating speaker pronunciation variability in speech embeddings : speaker and L1 effects on French as a Second Language

Speech variation between native and non-native speakers of French is addressed with a low-resource method based on a frame-wise comparison of wav2vec2 embeddings using fine-grained phonetic transcriptions by experts as baseline. z-scoring and time normalisation are explored to assess the phonetically analysable information. Adapting a Dynamic Time Warping method to speech embeddings, we compare phonologically similar recordings of native vs. non-native French speakers. The question is whether XLSR-53 embeddings are more robust than MFCCs to inter-speaker vs. intra-speaker variability, or to native speech vs. L2-speech. Results suggest that the model allows phonetically meaningful correlative analyses. Working on raw embeddings shows however that the representations are not speaker-independent, so with a view to address issues in relationship with L2 pronunciation variability, we show that normalising time provides a way to separate fluency and accuracy effects in L2-speech. This shows that wav2vec2 encapsulates time-dependent phonetic information in the embeddings, including speaker accent.

MOTS-CLÉS : Similarités cosinus, parole, acquisition L2, recherche-par-exemple (QbE-STD), méthode non supervisée.

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 2 : articles déjà publiés, pages 15–16.

KEYWORDS: Cosine similarities, speech, L2-acquisition, Query-by-Example Spoken Term Detection (QbE-STD), unsupervised method.

ARTICLE ACCEPTÉ À : SPEAKABLE@LREC 2026 : Workshop on Speech Language Models in Low-Resource Settings : Performance, Evaluation, and Bias Analysis, Palma, Mallorca, Spain, May 11-16, 2026..
URL : https://hal.science/hal-05577663