

Nordic AI Meet 2026

P3: LLMs, Language, Evaluation, Privacy & Trustworthy AI

6. From Sign to Speech: A Low-Resource Route for Icelandic Sign Language

Authors: Leroux Baptiste¹; Torfi Þórhallsson¹; Jón Guðnason¹

Affiliations: 1. Department of Engineering, Reykjavik University, Iceland

Abstract: Icelandic Sign Language is used by a small signing community and, unlike widely studied sign languages such as ASL, has no dedicated tooling for automatic translation. Existing sign-language representation and translation models are trained on high-resource sign languages and large-scale corpora; applied directly to ISL, they transfer poorly, both because ISL is linguistically distinct and because available ISL video is limited and lacks sign-level annotation. This gap motivates our work: building a pipeline that translates ISL into an audio modality. Our data comes primarily from RUV, the national broadcaster: news broadcasts in which a dedicated interpreter appears alongside the spoken newscast (~300h, under active preprocessing). This source does not provide sign-level gloss annotation, ruling out standard supervised fine-tuning and requiring indirect validation. We therefore ask whether a self-supervised video encoder, V-JEPA 2.1, already captures the structure our task needs. Using video only, its dense features are articulator-selective, with hands and face separating cleanly from the torso and background. Adding audio, cross correlating the interpreter's motion envelope against speech activity yields a single stable peak across 100 broadcast videos: the interpreter consistently trails the speaker by 1.3 seconds. This is, to our knowledge, the first reported audio-to-signing offset for ISL. The value is consistent with the only closely comparable published setting. The short, stable lag is consistent with the scripted, predictable nature of news voice-over, and our estimate is obtained without glosses, transcripts, or manual alignment. We ask whether those two independently trained frozen encoders represent shared content. Neither model was trained on sign language, yet across 100 videos the two feature spaces correlate well above chance: their temporal self-similarity structures align in 96 of 100 files ($\rho=0.10$, $\rho_z=5.4$), far exceeding a permutation null. This indicates that spoken content may be partially recoverable from generic visual features, though the signal is modest and could in part reflect shared activity rather than semantic content. The correspondence is strong enough to retrieve the matching spoken text in every file. In parallel, we build the audio-synthesis side. We extract features with mHuBERT-147 and train a KMeans model with 100 units on 40% of the corpus to predict discrete units, which feed a Unit HiFi-GAN vocoder. The vocoder is trained on 30% of the dataset for 150K steps. Comparing the 100K and 150K checkpoints on the full test set gives near-identical resynthesis quality, with mean WER of 0.441 versus 0.445. These results are promising, with further gains likely as training and data scale. Taken together, these results indicate that meaningful ISL-specific structure may be recoverable without sign-level labels, provided evaluation is designed to isolate style and timing from semantic content, and that V-JEPA 2.1 features are a sound basis for downstream task