Ensembles of Hybrid and End-to-End Speech Recognition

Aditya Kamlesh Parikh, Louis ten Bosch, Henk van den Heuvel
Center for Language and Speech Technology (CLST), Radboud University, Nijmegen, The Netherlands

Introduction

- **Challenge in ASR:** Achieving optimal performance for under-resourced languages is difficult because standalone ASR systems cannot achieve optimal performance in isolation.
- **Ensemble Technique:** We explored ensemble learning techniques, using **ROVER** (Recognizer Output Voting Error Reduction), to combine hybrid ASR with end-to-end ASR.
- **Method:** We used alignment and confidence scores to ensemble the ASR systems.
- **Results:** For the **Irish** language (with less than 15 hours of training data), the ensemble approach achieved ~14% Word Error Rate Reduction (WERR) on the primary test set and ~20% WERR on noisy and imbalanced test data.

ROVER Ensemble Technique

Example
Hypothesis 1: ['this been very interesting']
Hypothesis 2: ['this has been way interesting']

Output from alignment module:
Alignment 1: ['this', '@', 'been', 'very', 'interesting']
Alignment 2: ['this', 'has', 'been', 'way', 'interesting']

Output from Voting module:
Confidence 1: ['this', 1.0], ['@', 0.5], ['been', 0.91], ['very', 0.65], ['interesting', 0.95]
Confidence 2: ['this', 0.99], ['has', 0.9], ['been', 0.98], ['way', 0.16], ['interesting', 0.56]
Final output: this has been very interesting

Confidence Estimation

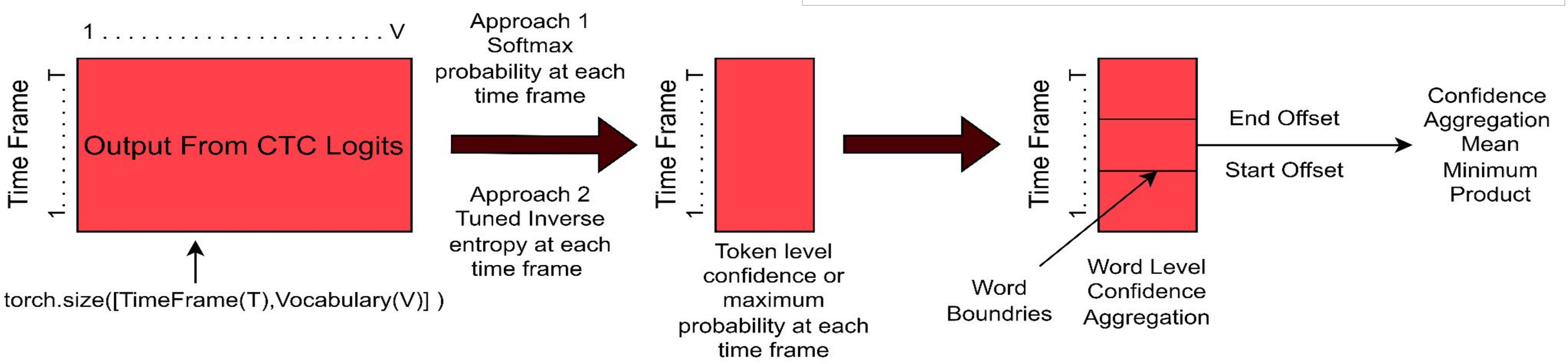
Confidence estimation in hybrid ASR:
Based on Minimum Bayes Risk (MBR)

$$w^* = \arg\min_w \sum_{w'} p(w'|x) L(w, w')$$

$p(w'|x)$ Probability of the word sequence w′ given the audio signal x

$L(w, w')$ Levenshtein distance between the two word sequences w and w'

End-to-End ASR Confidence estimation



Dataset Selection For Irish

Dataset	#Utterances	Duration	#Word Tokens	#Word Types
CV Train	4097	4.1h	27880	2341
LA Irish	1122	1h	11360	3542
GF Irish	1947	8.4h	48929	9866
CV Test	513	0.5h	3423	1109
CV Invalidated	282	0.3h	2230	707

Overview of the Irish datasets used. The abbreviations **CV**, **LA** and **GF** denote Common Voice, Living Audio and Google Fleurs, respectively.

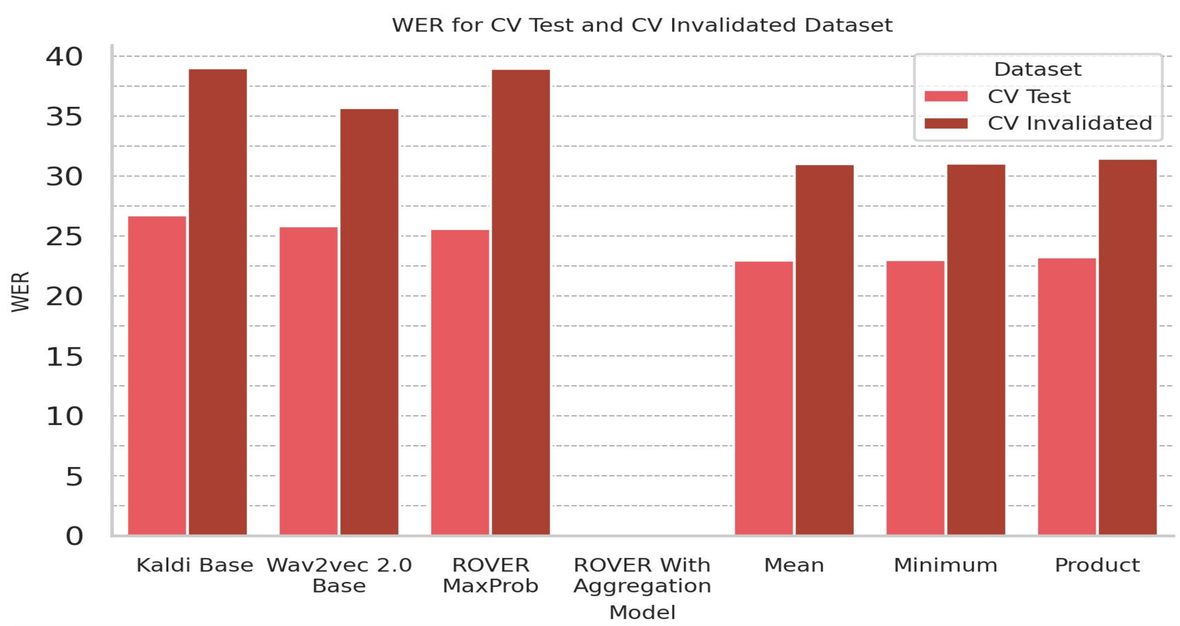
Experiments

Hybrid ASR model: Kaldi based model created based on “mini librispeech recipe”.

End-to-End ASR model: Wav2vec 2.0 XLS-R model is fine-tuned on 300 million parameters checkpoint. We used greedy decoding means simply picked up the best hypothesis at each time step.

Combined both models with ROVER, Tuned Renyi’s entropy with temperature scaling.

Results



Discussion

Advantages: Enhanced robustness against hallucination and improved generalization capabilities.
Disadvantage: Not suitable for real-time decoding.