Speech Recognition Algorithm for Detecting Mispronunciation for Research with Children

Dana Kalaaji

EPFL Master Thesis

18 September 2024

Automatic Speech Recognition

Definition: Automatic Speech Recognition (ASR)
→ Converts speech into text

Evolution: From traditional to end-to-end ASR
→ Improved performance

Motivation: Promising opportunity for application in educational context

→ Literacy development

Challenge: Models primarily trained on adult data
→ Underperforms with children's speech

Thesis Objective

Problem:

- Reading assessment methods → labour-intensive and time-consuming

Thesis Goal:

Automate mispronunciation detection

- Real-time model to classify pronunciation of each (pseudo) words from predefined lists
- Binary classification: Correct / Incorrect

Purpose:

Integrate into speech evaluation pipelines

- Alleviate professional's manual workload
- Allow them to focus on analysing errors

Challenges in Child Speech Recognition: Overview

Unique Speech Signal Characteristics

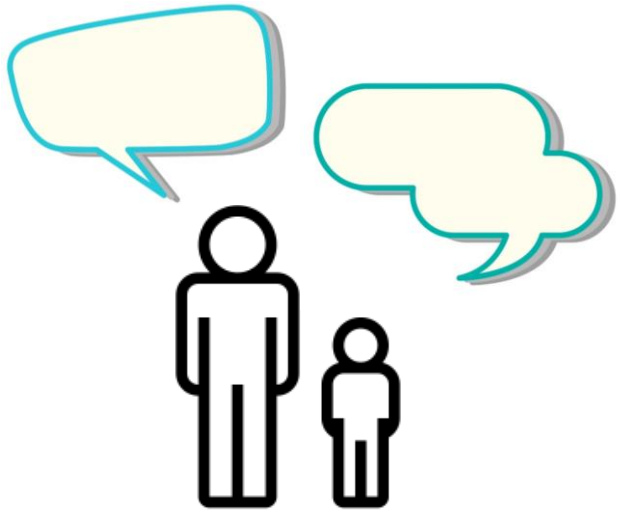
Increased Inter and Intra Variability

Dataset Scarcity

Unique Child Speech Signal Characteristics

Higher fundamental frequency

Shifted formant frequencies due to smaller vocal tracts



More non-verbal vocalizations

Lower articulation rate

Higher incidence of mispronunciations and disfluencies

Longer mean phone duration

Longer and less strategic use of pauses

Increased Variability in Child Speech

- Greater inter + intra-speaker variability (spectral & temporal) → Decreases with age
- Challenges in ASR for younger speakers:
    - More adaptation data needed for younger children
    - Persistent performance gap for younger children

Child Speech Data Scarcity

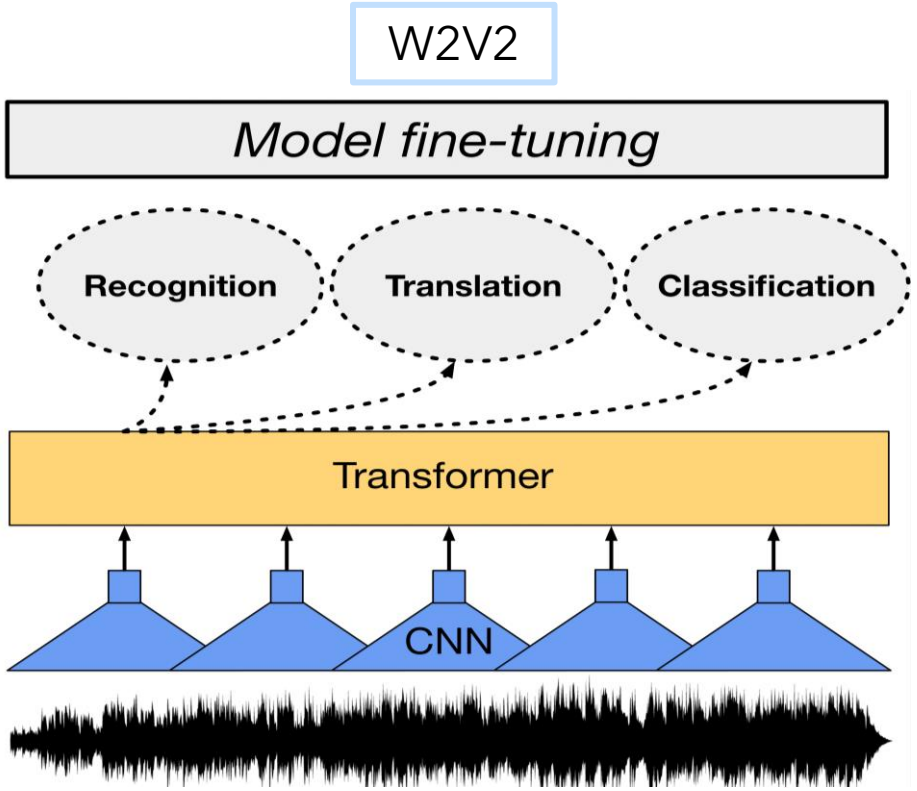
- Lack of datasets for children's speech, particularly for languages other than English

End-to-End Models and Child Speech Recognition

- Child Speech: end-to-end models outperform traditional ASR
- CNN-based E2E acoustic models → learn formant info invariant to acoustic differences
- Challenges persist → Performance for child speech is
  - 10 to 19 times lower without adaptation
  - 6 times lower with adaptation

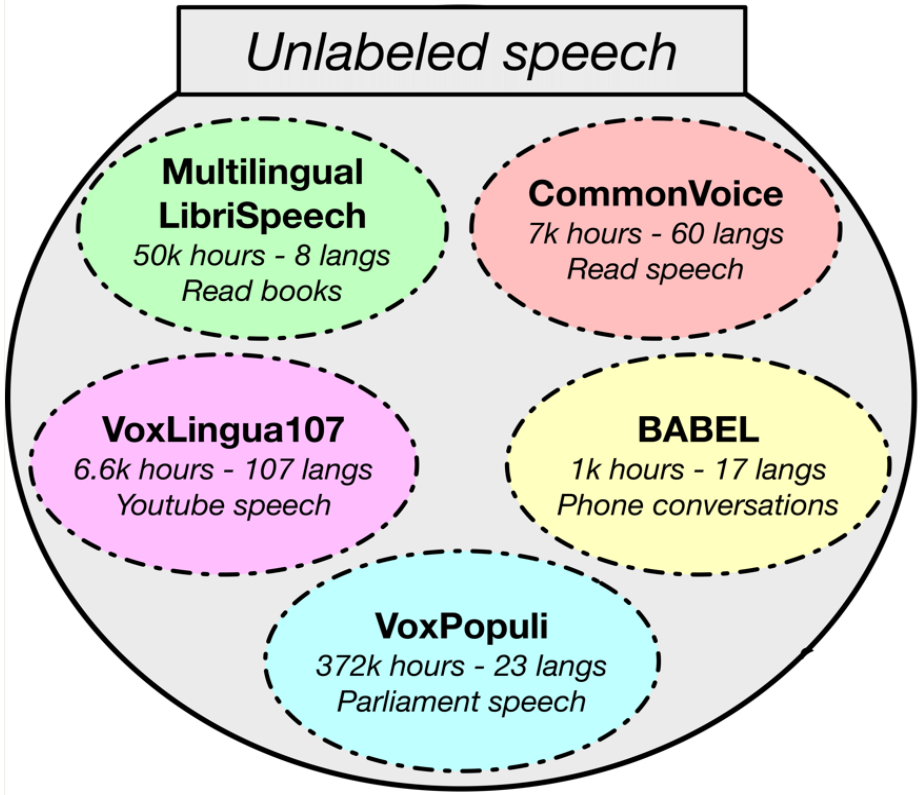
State of the Art ASRs

Whisper	Wav2Vec2
Fully Supervised	Self-Supervised
Advantages	
Zero-shot performance	Good if low ressource
Robust (accents, noise, …)	LM independance



State of the Art ASRs

	WhisperX	W2V2 XLS	W2V2 VOX	W2V2 Phoneme
Training Hours	680k	436k / 1.5k	100k / 750	22.8k / 940
Languages	100	128 / Fr	23 / Fr	Fr
Datasets	Diverse scenarios, accents, …	Diverse scenarios	VoxPopuli / Common Voice	VoxPopuli / Common Voice



Reading Task Dataset: The Decoding Task

- 155 participants → 7-8 years old children
- Decoding task: 18 different lists of 12 (pseudo) words
- 3 groups of comparable difficulty (A, B, C)
    - 6 lists per group: 2 easy, 2 complex, 2 pseudo
- Total audio duration: 6.5 hours

Easy List
nuit métal joue escalade valet tente jaloux couleur fossé noix balade reptile
Complex List
piège femme secret finit mille cerf jardin précision dix lieux million débarquement
Pseudo List
suf fari juit lumèce goix munon donte tondé toir rombage jeur brète

Audio Quality Challenges

Background noise

Microphone buzz

Disfluencies

Partially cut/
inaudible recording

Variations in pauses
& reading speeds

Transcription & Accuracy Scoring Methodology

Clinician transcription

- Transcribe pronunciation of sounds
  → not correct spelling & inconsistent
- Mark insertions & sounding out

Clinician score

- 2: Correct
- 1: Almost correct
- 0: Incorrect
- NA: Not applicable

Example

Reference:	bucurelle daveau ninoie trefonet tiège ronjeau meuil trufondulence hupeur mive gralation turlème
Transcription:	{comments} bu.turelle davo ninoi trefonet tchiètche ron ronge ronjo mail trufodulence {inaudible} dralation turlè{v}.me
Accuracy:	0 2 2 2 0 2 0 0 NA NA 0 1

Performance Evaluation: Metrics

TP	Clinician score = ASR score = correct
TN	Clinician score = ASR score = incorrect
FP	Clinician score =incorrect & ASR score = correct
FN	Clinician score =correct & ASR score = incorrect

→ Objective: Maximize TP, TN, minimize FN, FP

2 scenarios → Need flexibility !

- Post-hoc Analysis: low FP rate (< 5%) → Trust "correct" label
- Real-time Task Assessment: low FN rate → Trust "incorrect" label

Clinician's score mapping for binary classification
- 2, 1 → Correct (1)
- 0, NA → Incorrect (0)

Performance Evaluation: Method

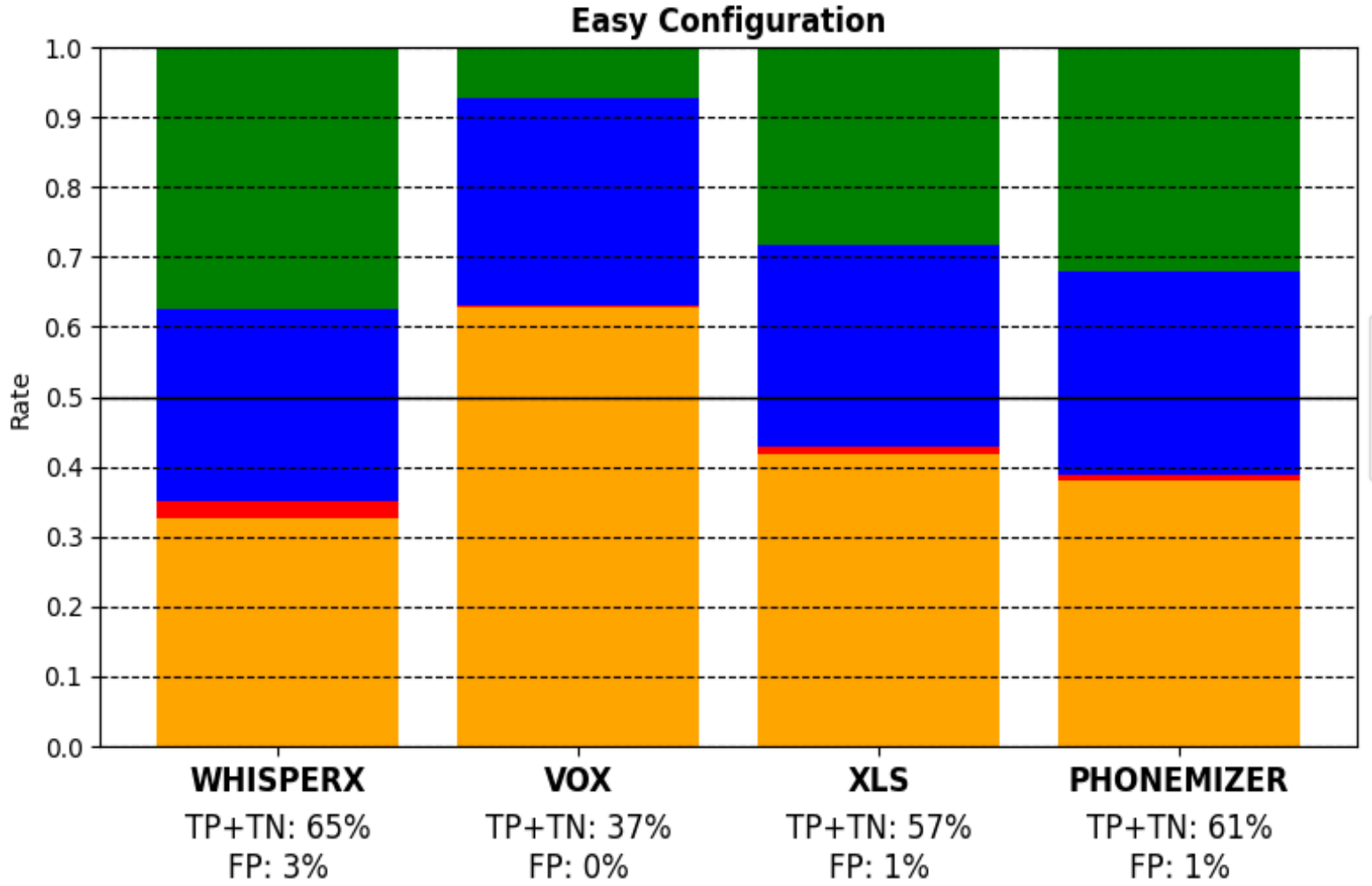
Method

1. Get ASR Transcription
2. ASR Accuracy Scoring → 1 if match reference, 0 else
3. Clinician Performance Comparison → TP, TN,FP, or FN for each item

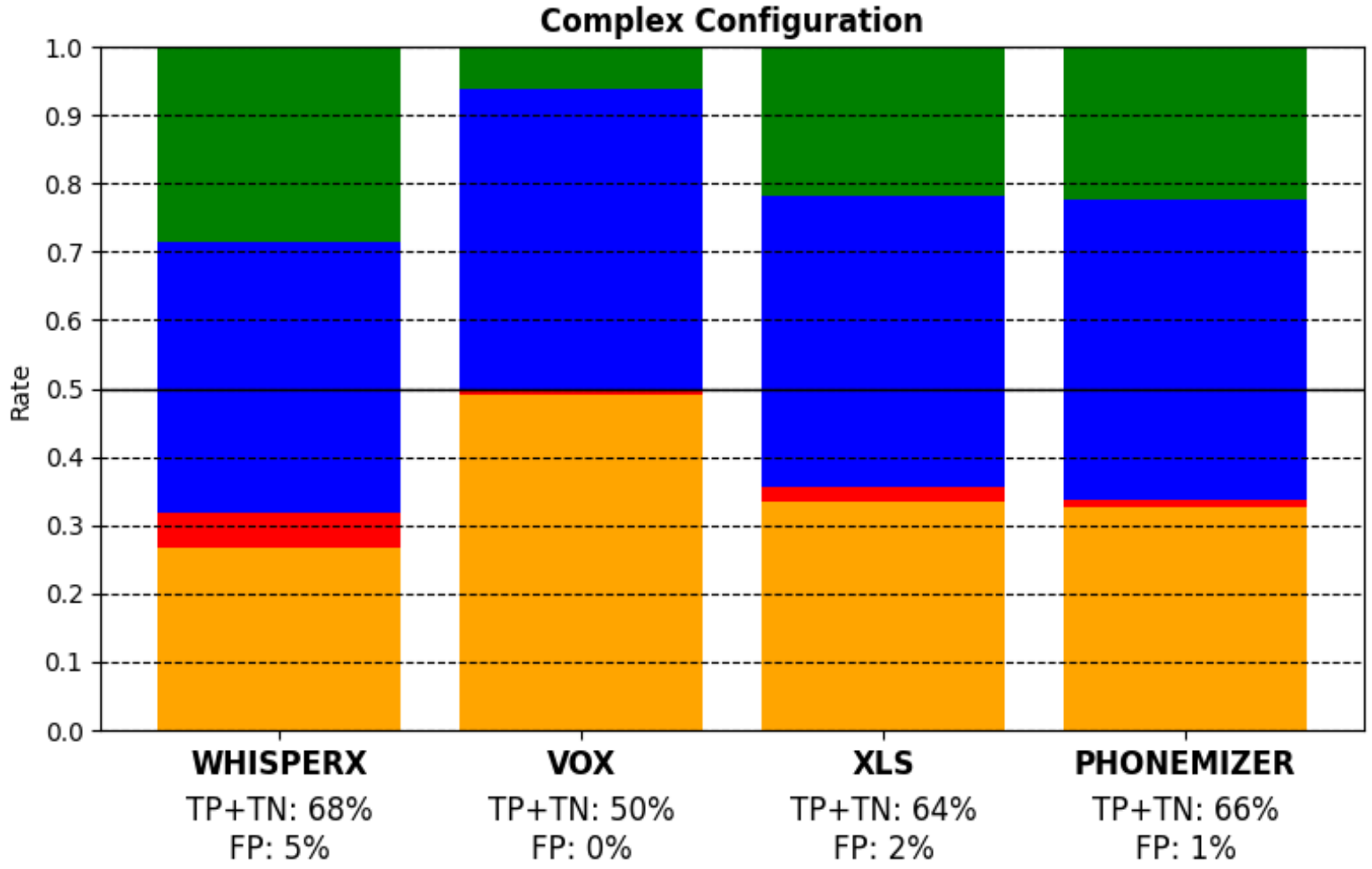
Example

Reference Text:	toit régal juin escapade navet pente bandit chaleur carré choix parade fossile
Clinician Transcription:	toi rédale join es esgabade navet pente mendi chaleur charré choix parade fossile
ASR Transcription:	toi ridal juin estabadnavé pantmendi chaleur tharéchoix parade fossiles
Clinician Accuracy:	1 0 1 0 1 1 0 1 0 1 1 1
ASR Accuracy:	0 0 1 0 0 0 0 1 0 0 1 0
Performance:	FN TN TP FN FN FN TN TP TN TN TP FN

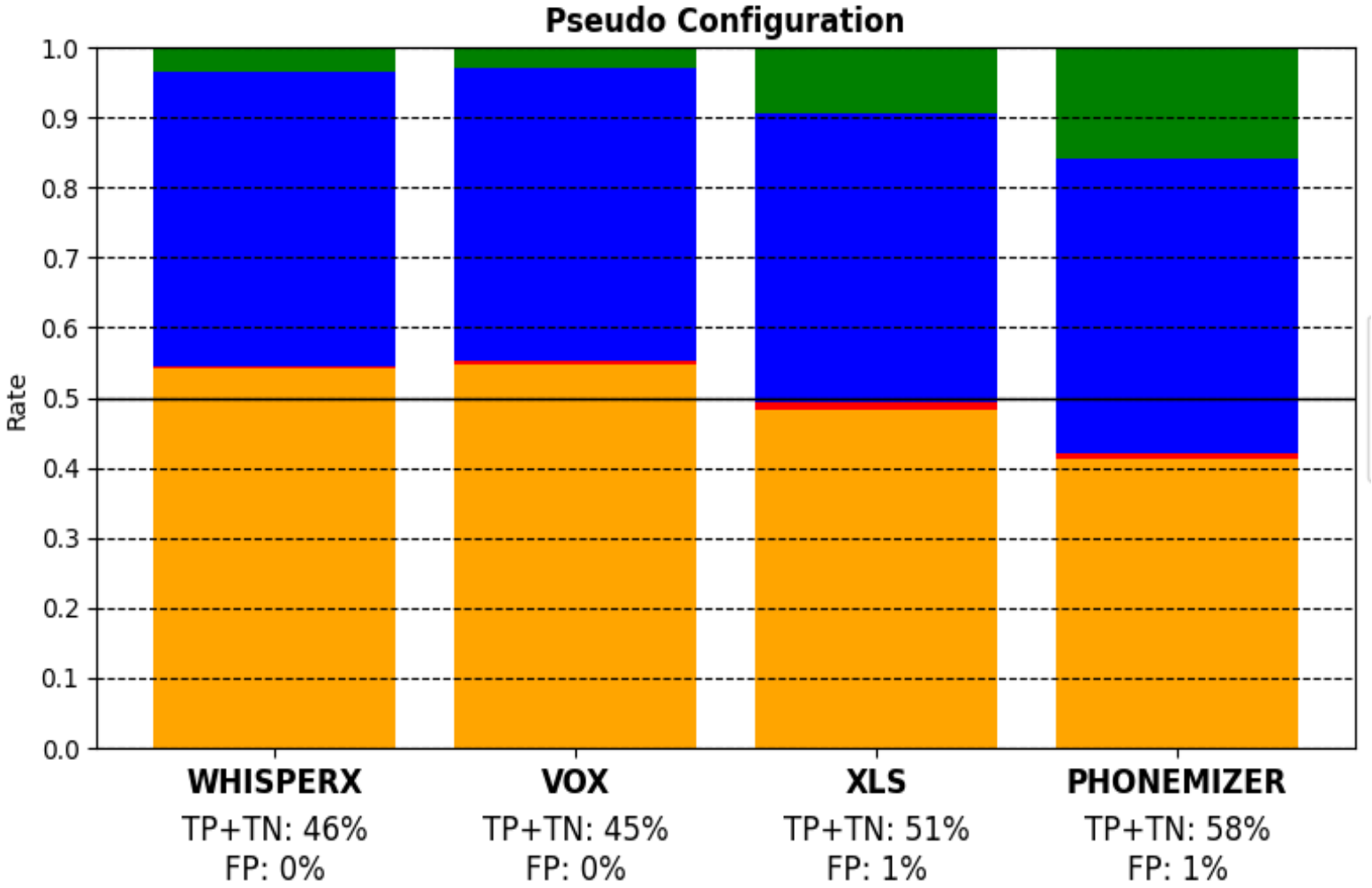
Performance on Easy Lists



Performance on Complex Lists



Performance on Pseudo Lists

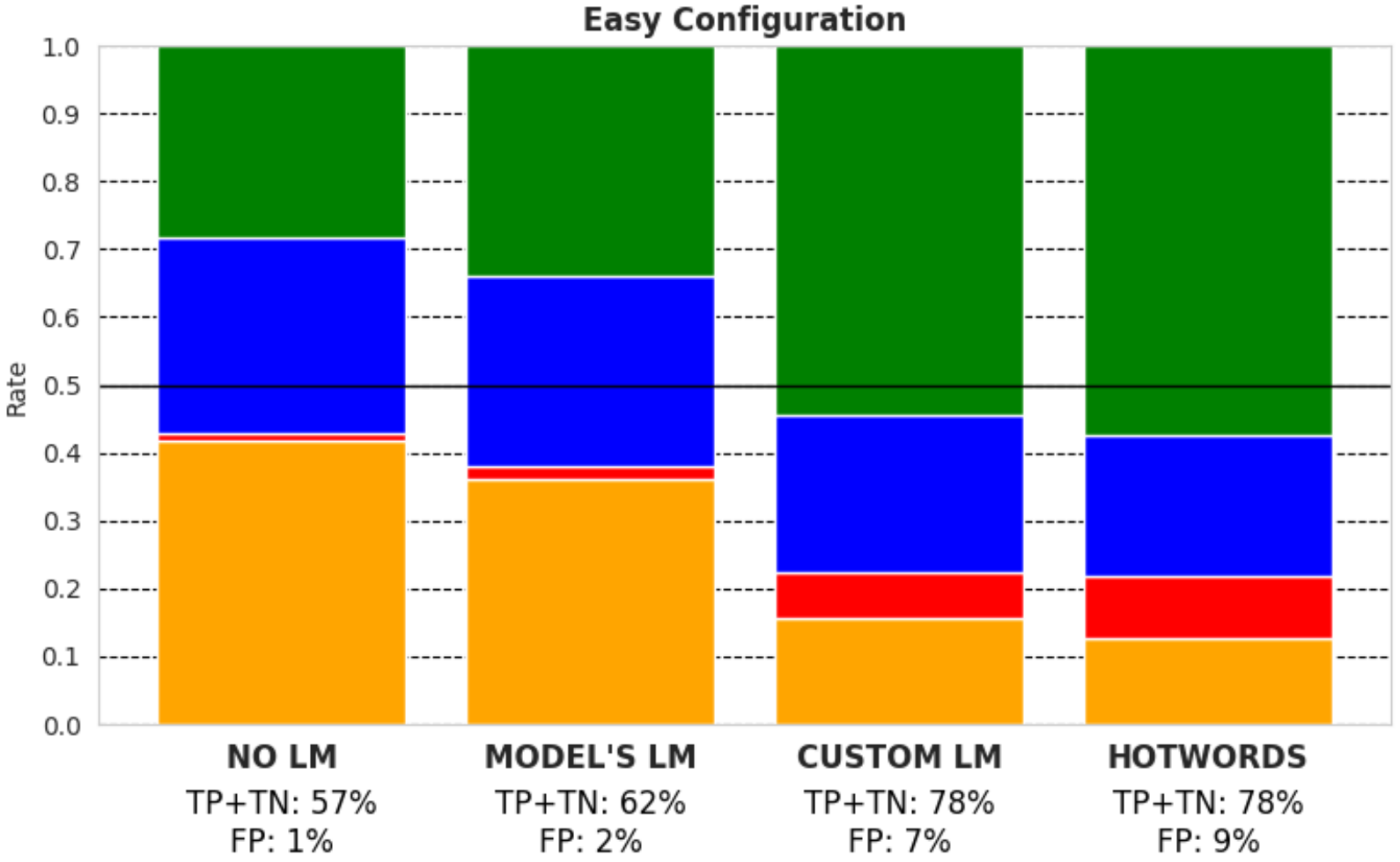


Performance on the Dataset: Example

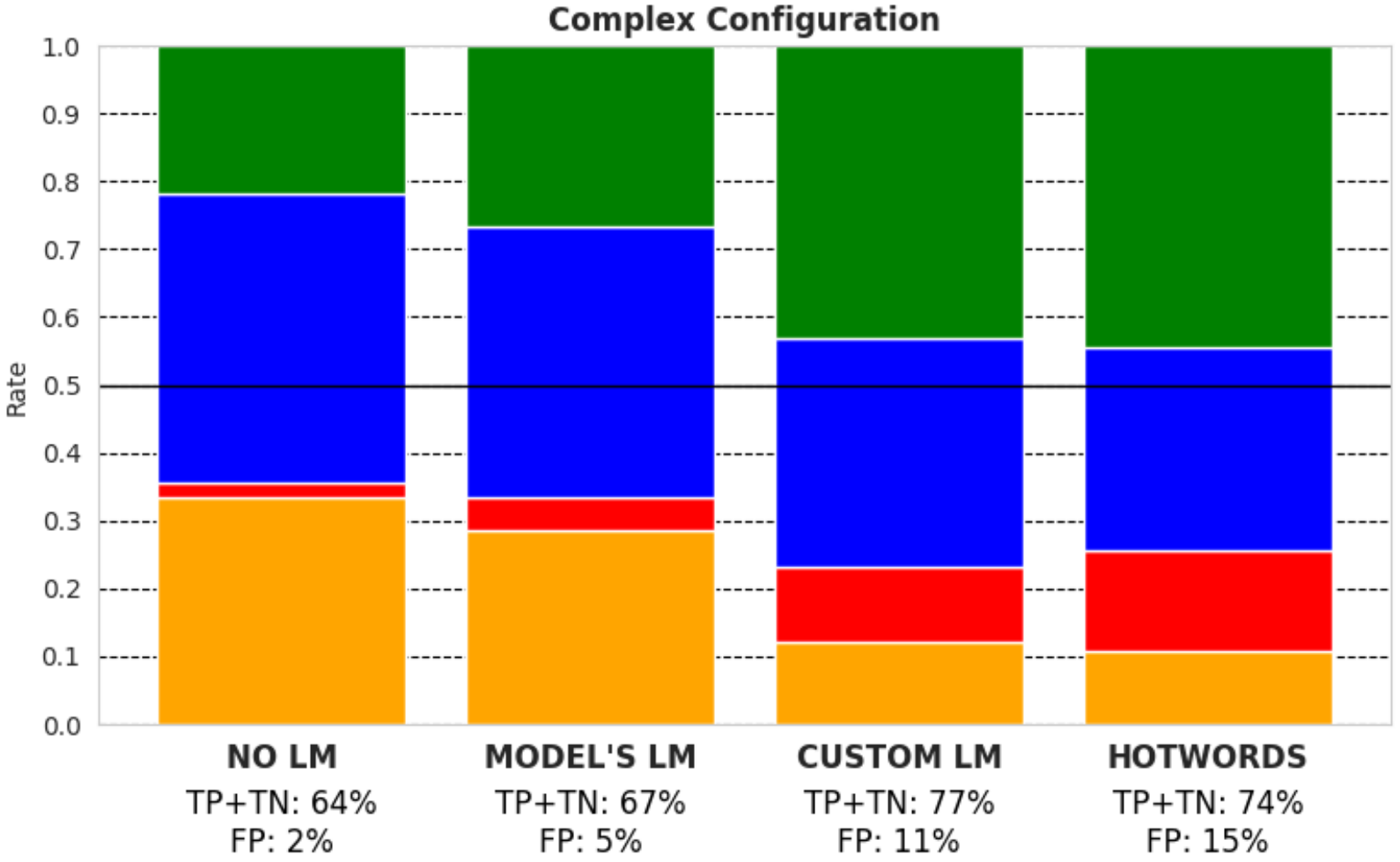
Reference	bucurelle daveau ninoie trefonet tiège ronjeau meuil trufondulence hupeur mive gralation turlème
Clinician Transcription	brou bu.curelle davo ni.noi tréchfoné tiège ron.jeau mieu tronfondulence uper mèive car gralation turlem lé {cut recording}
WhisperX	De Querelle, Davos, Nînois, Trésifonée, Thièges, Rongeau, Mieux, Tronçon du Lens, Jupiter, Mives, Quelle relation tu relèves ?
VOX	por u urrel davo ninoirttrafone tiegeronjemiur troufon du lonpur milzegagralation turllaenl
XLS	bbeucurel daveau ni noir tréfoné tiège ron jau mieux troufon du lance uper mive gagralation turlame

Table 4: Comparison of the performance of WhisperX, Vox and XLS on a pseudo words list

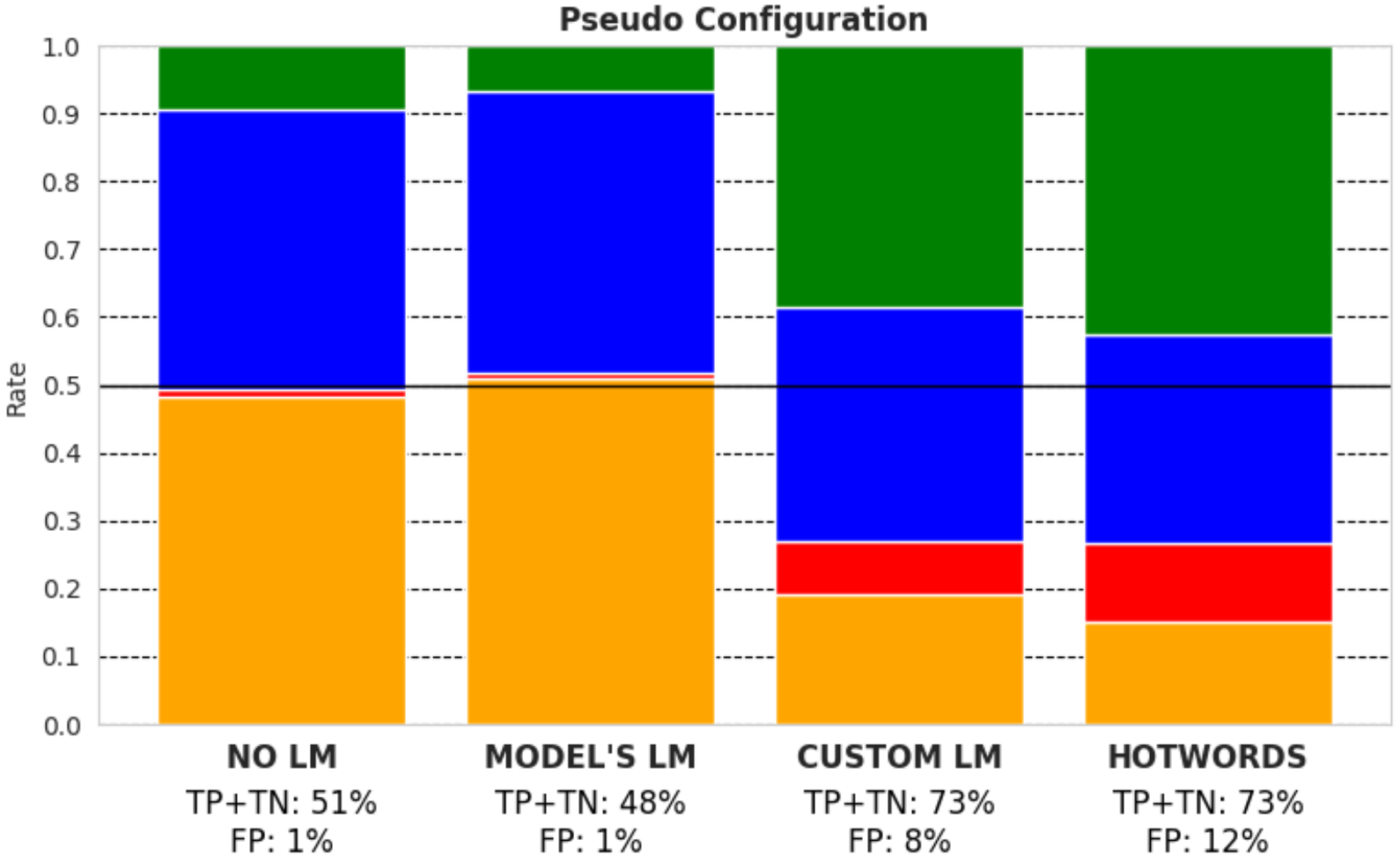
LM & Hotwords Integration: Easy Lists



LM & Hotwords Integration: Easy Lists



LM & Hotwords Integration: Easy Lists



LM & Hotwords Integration: Example

Reference	bucurelle daveau ninoie trefonet tiège ronjeau meuil trufondu-lence hupeur mive gralation turlème
Clinician Transcription	brou bu.curelle davo ni.noi tréchfoné tiège ron.jeau mieu tronfon-dulence uper mèive car gralation turlem lé {cut recording}
XLS no LM	bbeucurel daveau ni noir tréfoné tiège ron jau mieux troufon du lance uper mive gagralation turlame
XLS + LM	deux cures daveau ni noir trépané tiège rongea mieux trou fond du lanceur mive gagralation turlan
XLS + Custom LM	rof rari tuit cutice bumon bante perné loi bonbonfange je tetraime
XLS + Hotwords	bucurelle daveau ninoie trefonet tiège ronjeau meu trufondu-lance hupeur mive gralation turlème

Table 5: Influence of a LM and Hotwords for the XLS model on a pseudo words list

Problem: Ambiguity in Text-Based Approach

- Example: 'orchestre'
  - Correct: /ɔʁkɛstʁ/
  - Possible mispronunciation: /ɔʁʃɛstʁ/ 'or-ch-estre'

- Example: 'aout' :
  - Correct: /ut/
  - Possible mispronunciation: /aut/ 'a-out'

→ Text outputs doesn't distinguish between pronunciations

Challenges for the Phonetic transcription Approach

Statement: Wav2Vec2 Phonemizer competes despite less training data + outperforms for pseudo-words

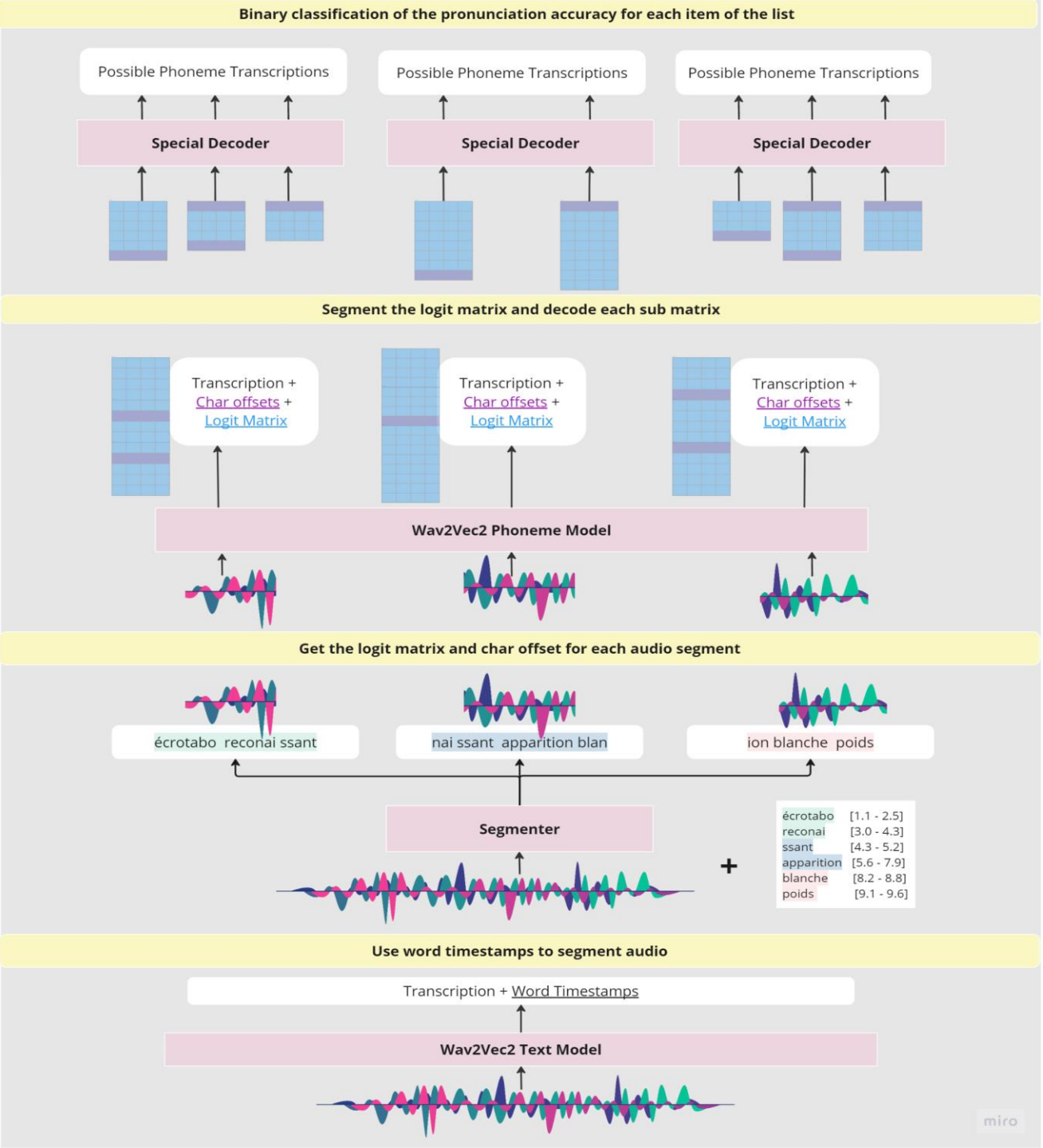
Problem: text-to-phonetic conversion not reliable → **Can not finetune !!**

1) Encoding based on known mappings → bad with non-words, misspelling, accents

Pronunciation	Phonetic Encoding	Clinician Transcription	Convertor's Output
Correct	/ɔʁkɛstʁ/	orkestre	/ɔʁkɛstʁ/
Incorrect	/ɔʁʃɛstʁ/	orchestre	/ɔʁkɛstʁ/
Pronunciation	**Phonetic Encoding**	**Clinician Transcription**	**Convertor's Output**
Correct	/apaʁisjɔ /	apparission	/apaʁisjɔ /
Incorrect	/apaʁitjɔ /	apparition	/apaʁisjɔ /

2) Encoding based on litteral sound → bad with silent letters & special pronunciations

- 'lisoie' encoded as /li**s**wa/ and not /li**z**wa/
- 'cerf' encoded as /sɛʁ**f**/ and not /sɛʁ/



ASR for Mispronunciation Detection: Pipeline Overview

1. Hotword-Enhanced Wav2Vec
2. Audio Segmentation
3. Phoneme-Level Wav2Vec2
4. Advanced Decoding Process

→ Combine strength of word & phoneme level ASRs

→ ASRs trained on adult speech = accessible

Audio Segmentation

Challenges in Wav2Vec2 Phoneme Transcriptions

- Item boundary delimitation
- Phoneme omissions and errors

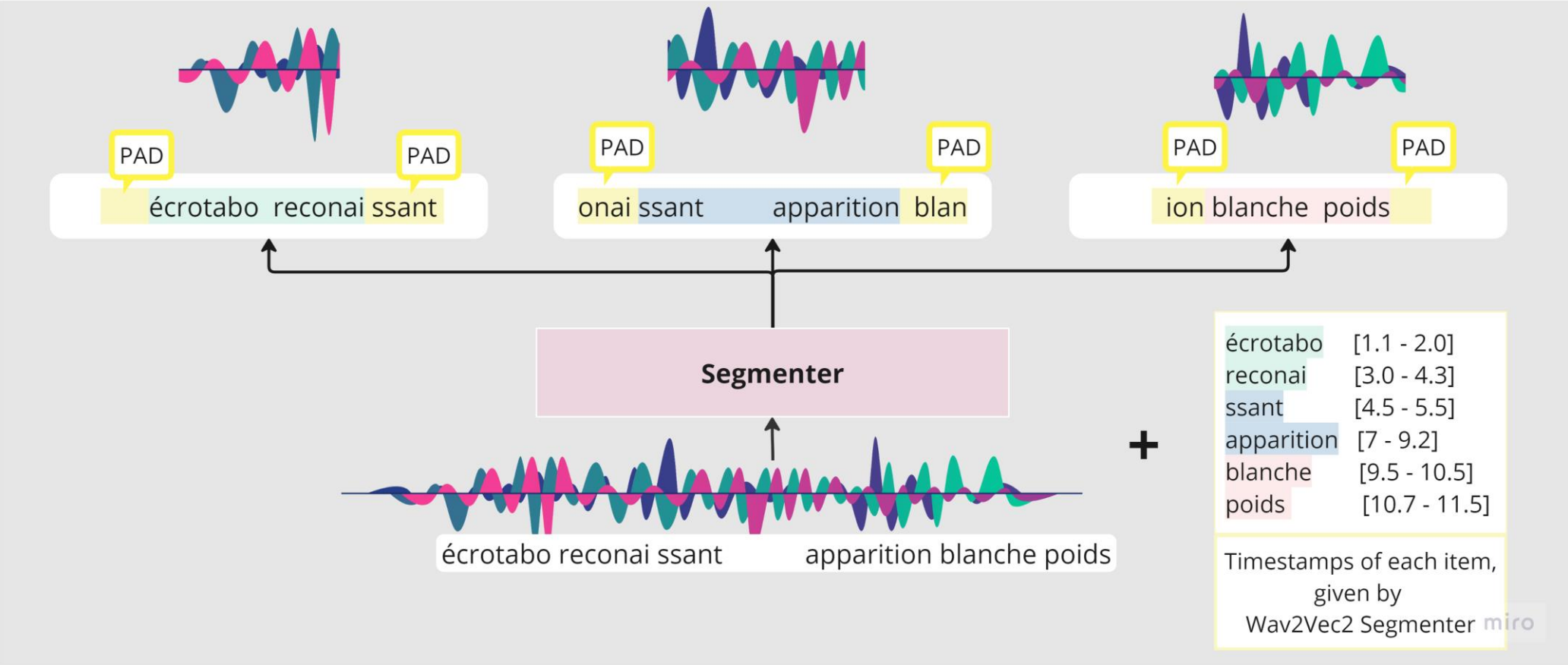
Possible Reasons

- Trained on 5-6s samples vs. 20-30s in dataset
- Variable children's speech patterns

Segmentation Attempts

- Split audio into 5s chunks with 1s overlap → decreased performance
- Revealed high variability in speech content within segments

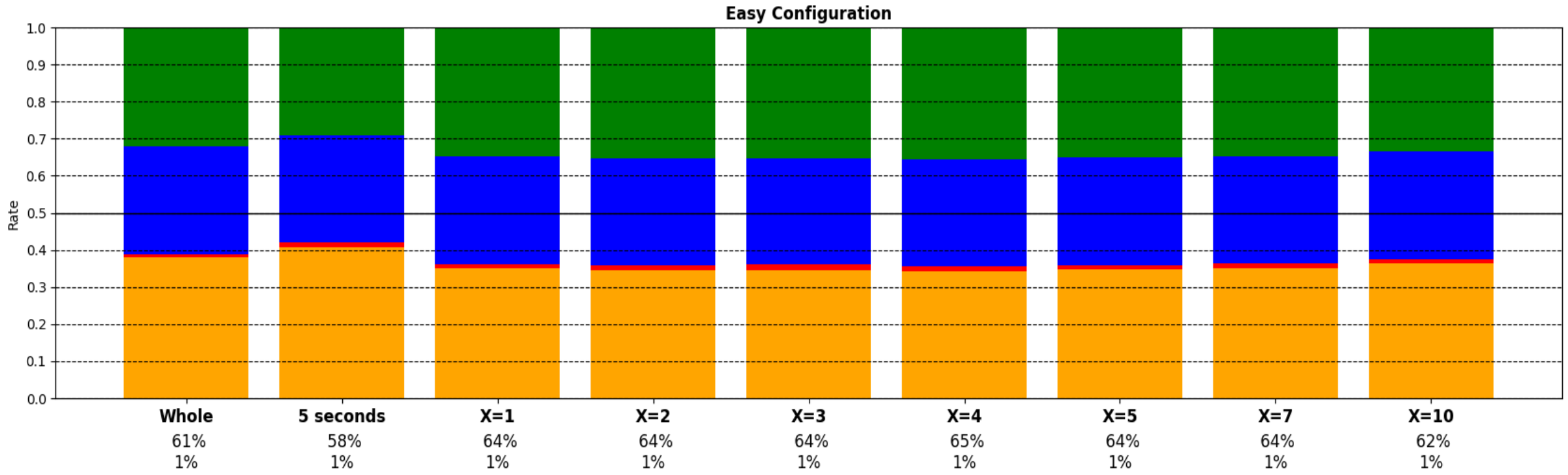
Segmentation Strategy



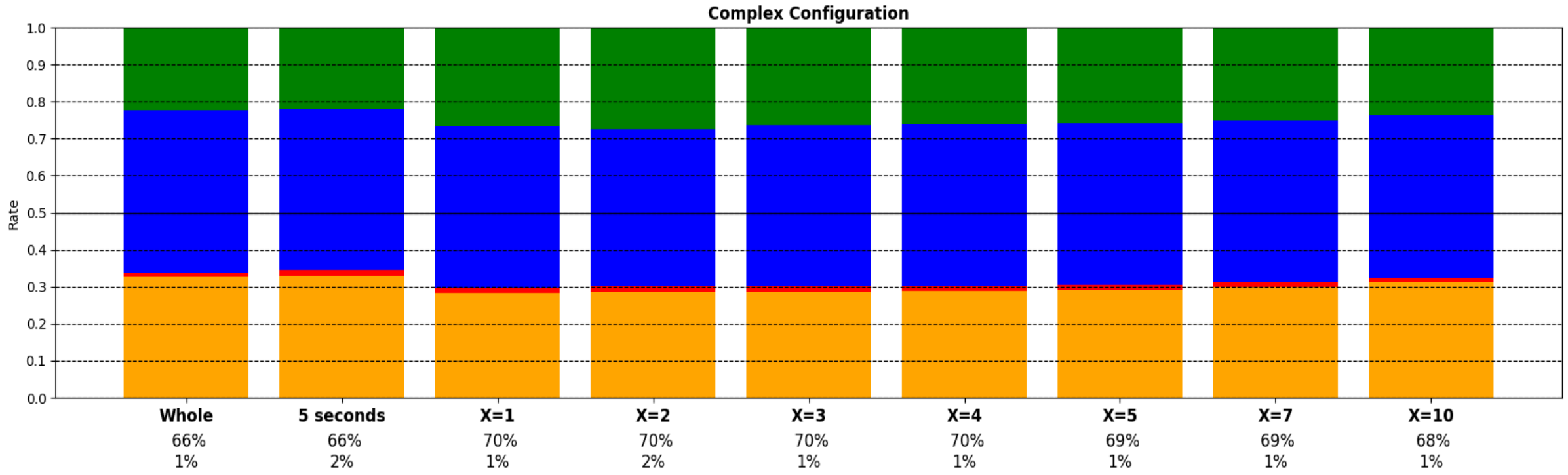
→ Segment audio based on actual speech content + 1s Pad

→ Optimal segment duration: 4 s of actual speech

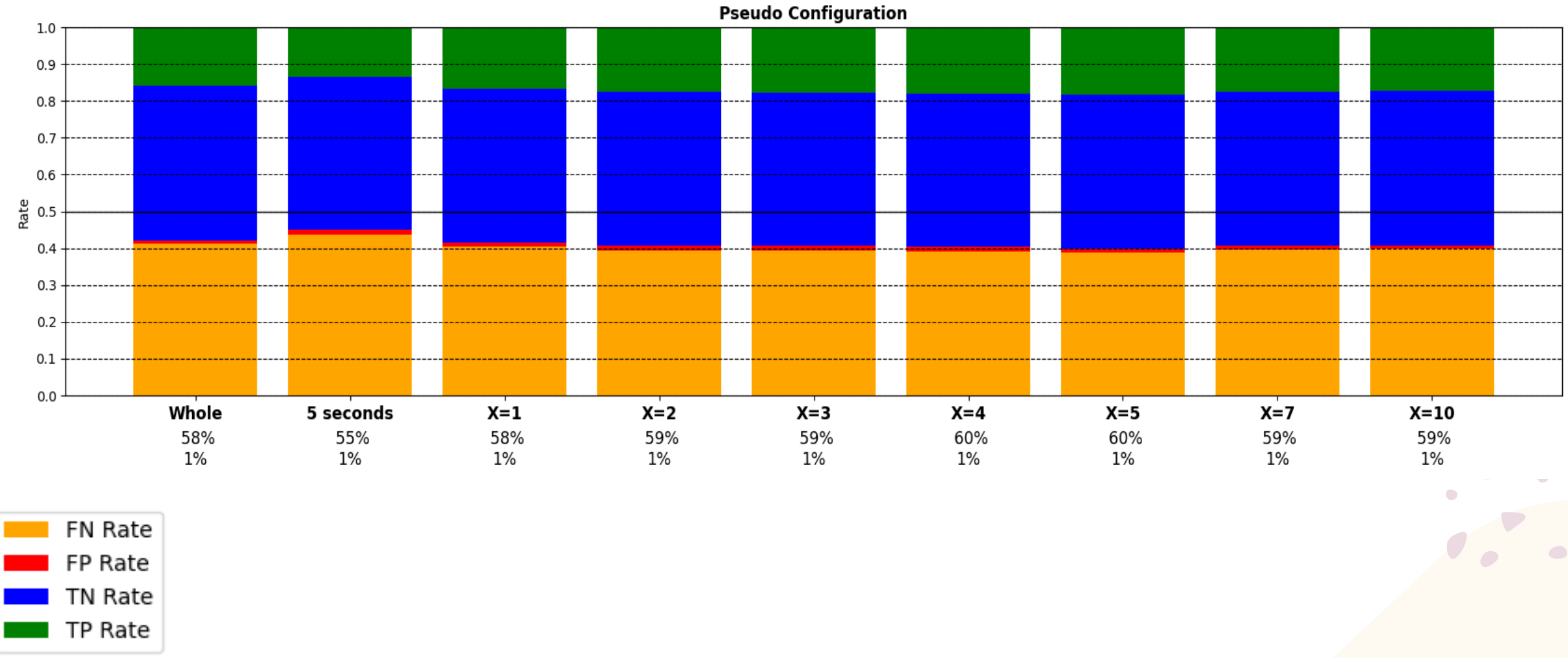
Segmentation Strategies Approaches: Easy Lists



Segmentation Strategies Approaches: Complex Lists



Segmentation Strategies Approaches: Pseudo Lists



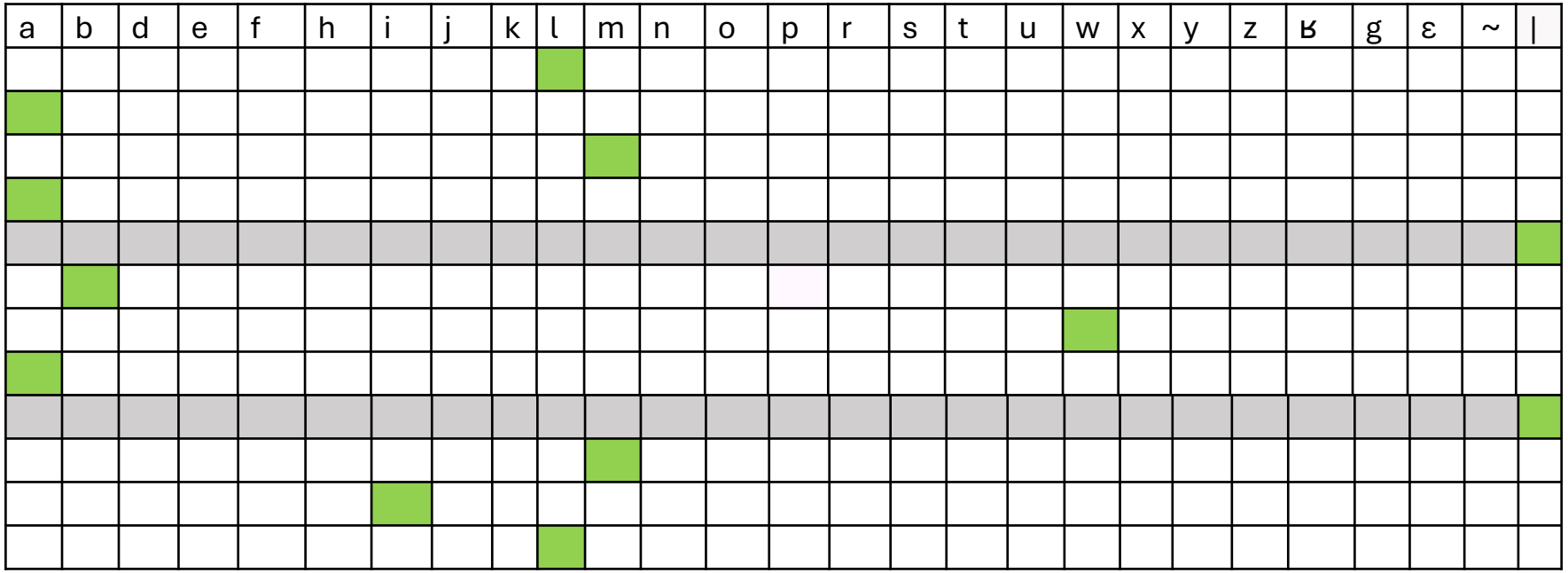
Original Decoder

Task: **lama poids mille**

Target: **[lama] [pwa] [mil]**

Utterance: **lama pwa mil**

True Score: **[1, 1, 1]**



Logit Matrix:

- Dimensions: [Number of timeframes × 59 characters]
- 1 timeframe = 20 ms
- Values: confidence for each character at each timeframe

Wav2Vec2 Decoder + Classification :

- Argmax for each timeframe
- Check for each list if an element matches a target

Limitations: Only considers single most probable character

➔ **'lama bwa mil'** ➔ **[1, 0, 1]**

Decoder 1

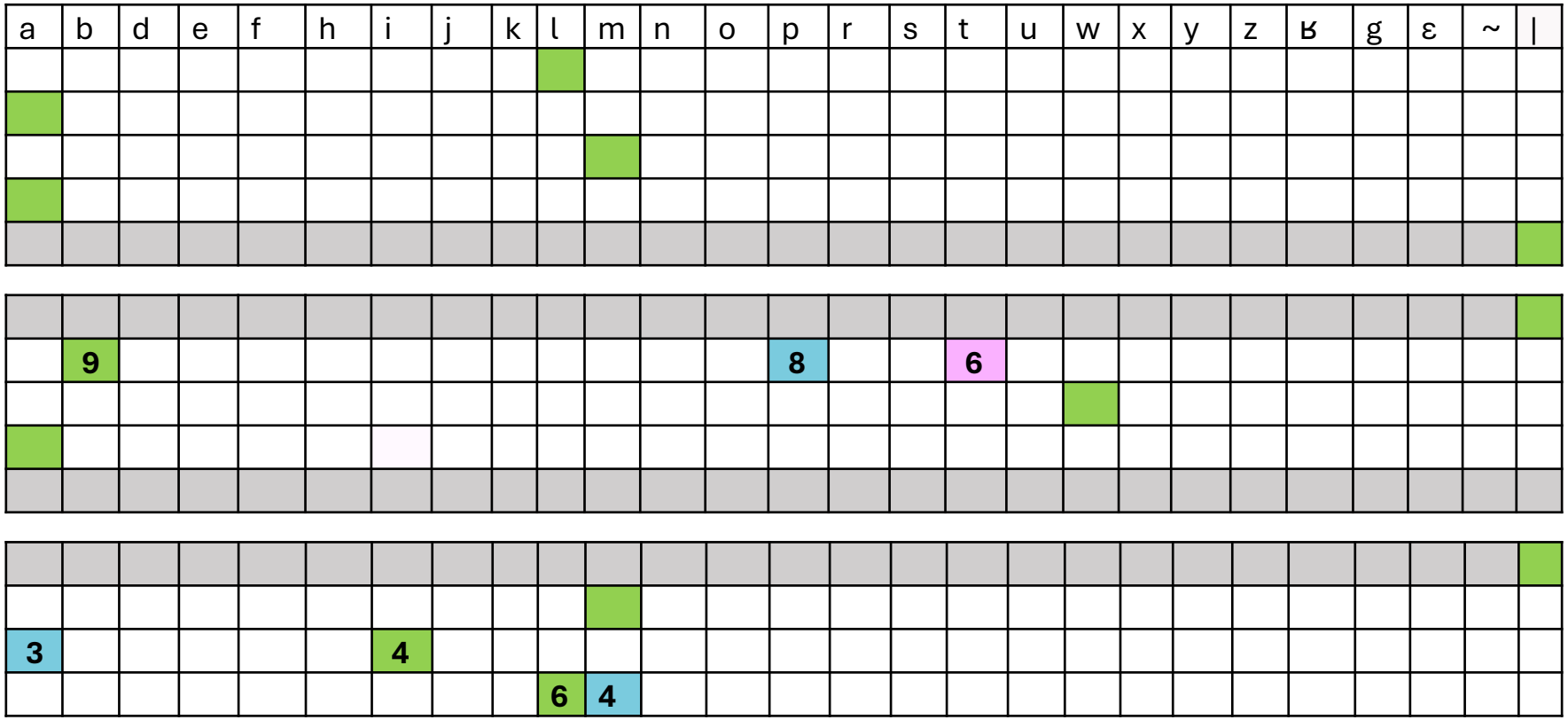
Task: **lama poids mille**

Target: **[lama] [pwa] [mil]**

Utterance: **lama pwa mil**

True Score: **[1, 1, 1]**

Threshold: **t = 2**



Improvement: Consider multiple probable phonemes at each timeframe

1. Segment logit matrix based on char offset
2. For each timeframe select phonemes within threshold **t** of top **k** candidate
3. Generate all plausible phoneme sequences (Cap predictions at **M** per submatrix)

   → [lama] [bwa, bwi, pwa, pwi] [mim, mam, mil, mal]

1. Compare against reference phonemes

   → **[1, 1, 1]**

Decoder 1 limits

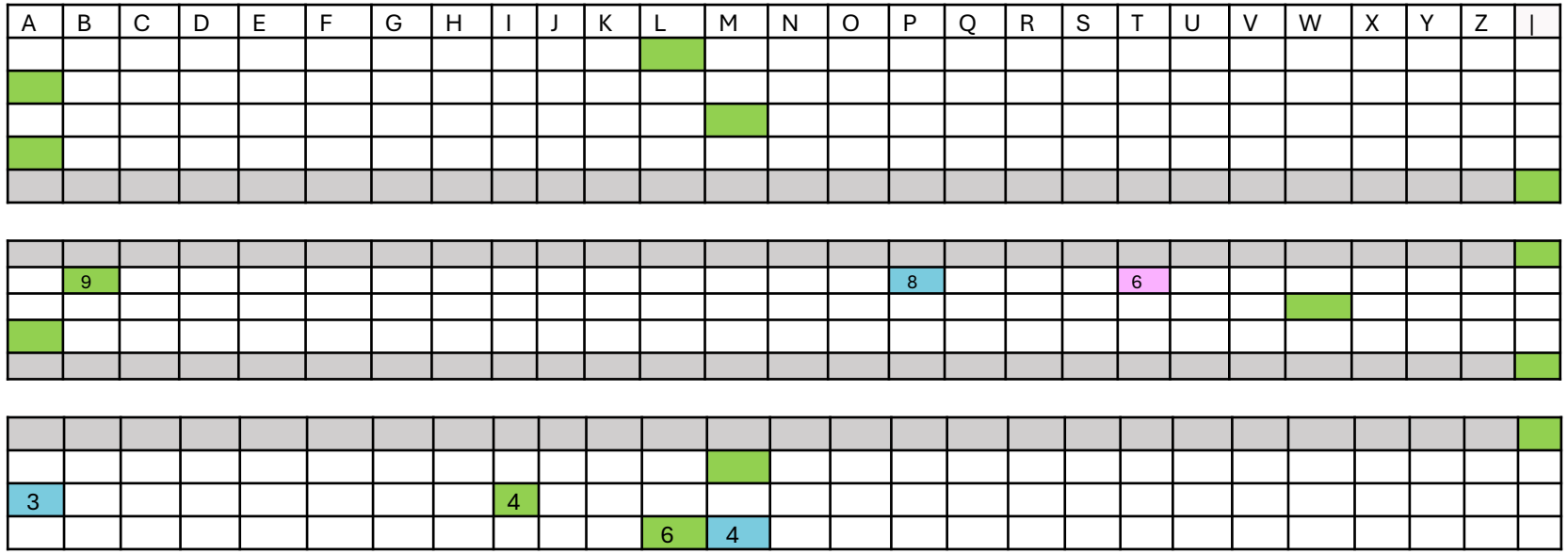
Task: **lama poids mille**

Target: **[lama] [pwa] [mil]**

Utterance: **lama pwa mil**

True Score: **[1, 1, 1]**

Threshold: **t = 2**



Problem: bad if spaces not well placed

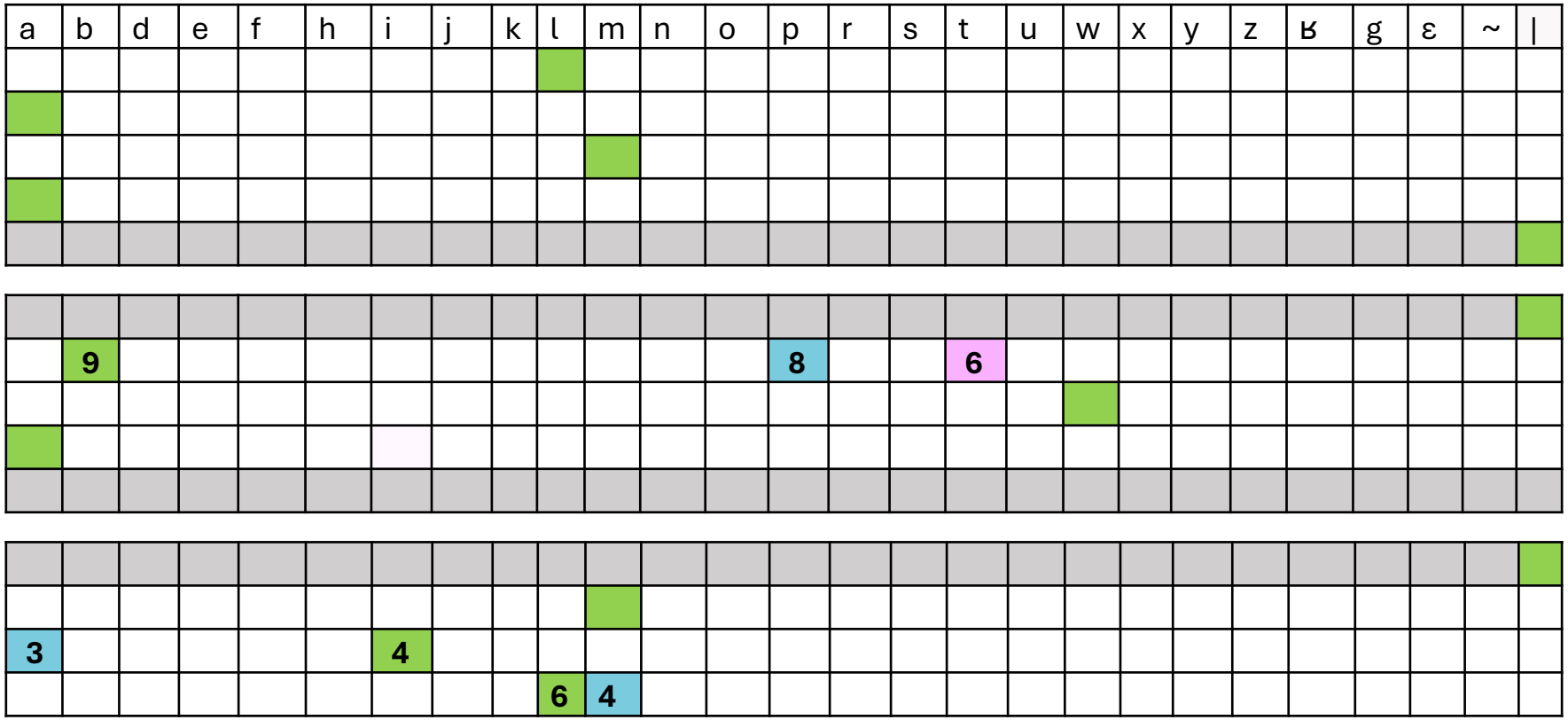
Missing space → Segment = **lama pwamil**

1. Segmentation: [lama] [pwamil]
2. Possible candidates: [lama] [bwamim, bwamam, …, pwamim, pwamam, pwamil, pwamal]
3. Matching & Result: **[1, 0, 0]**

Extra space → Segment = **lama pw a mil**

1. Segmentation: [lama] [pw] [a] [mil]
2. Possible candidates: [lama] [pw, bw] [a], [mim, mil, mam, mal]
3. Matching & Result: **[1, 0, 1]**

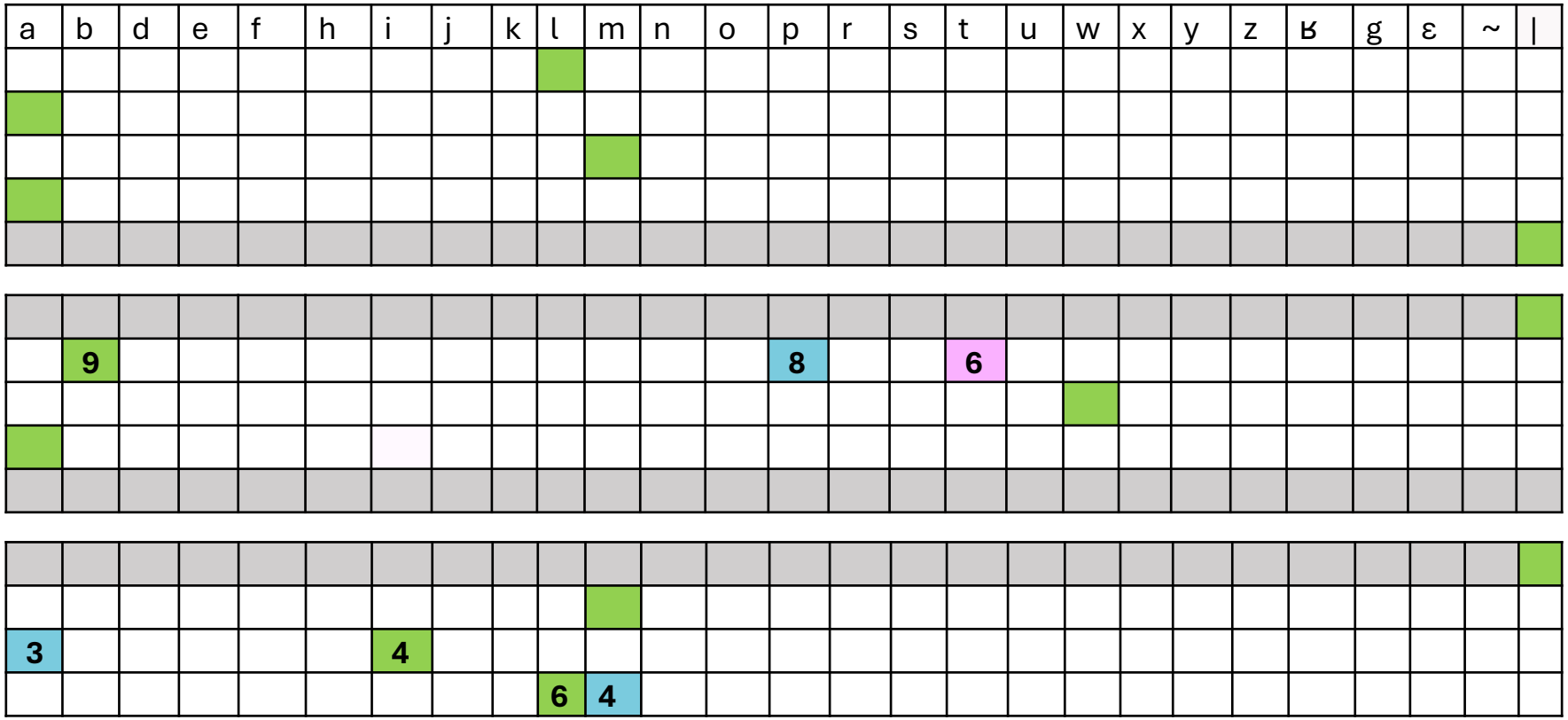
Decoder 1



Problem: Bad if spaces not well placed (speech rate)

1. Segment logit matrix based on char offset
2. For each timeframe select phonemes within threshold **t** of top **k** candidate
3. Generate all plausible phoneme sequences (Cap predictions at **M** per submatrix)
4. Check if an option is equal to a target

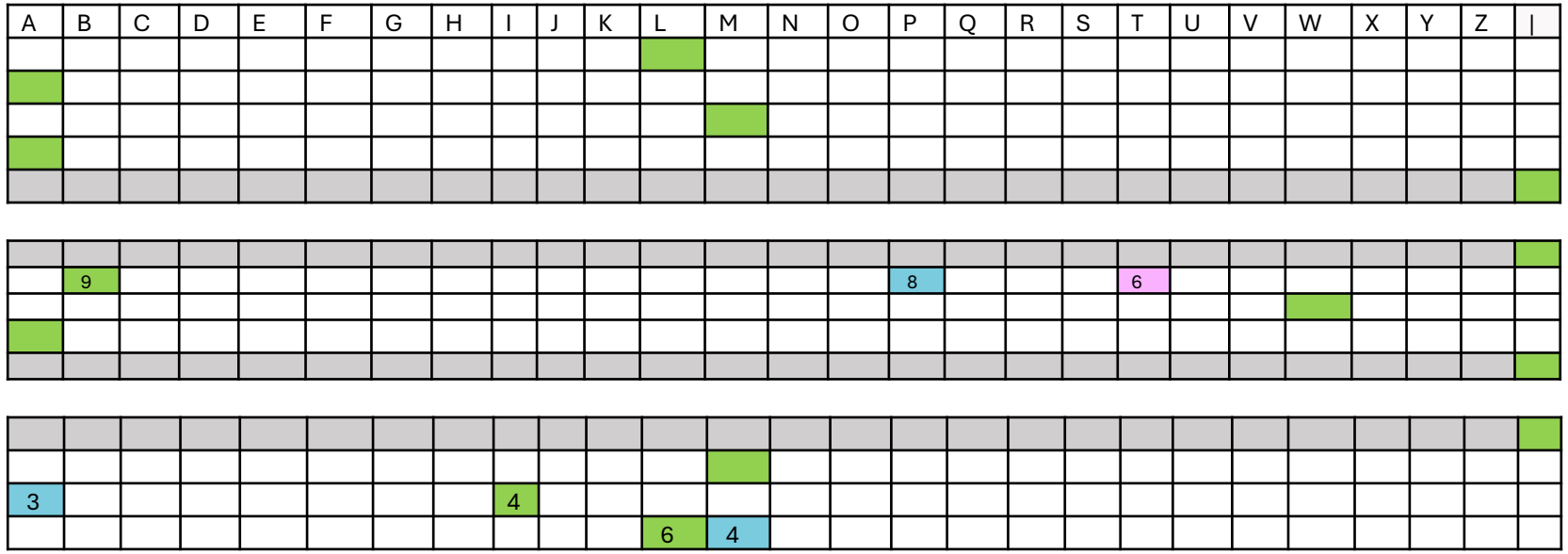
Decoder 2



Improvement: Handles cases where spaces not well placed (speech rate)

1. Segment logit matrix based on char offset
2. For each timeframe select phonemes within threshold **t** of top **k** candidate
3. Generate all plausible phoneme sequences (Cap predictions at **M** per submatrix)
4. Create a sliding window of 3 concatenated predictions
5. Check if target is contained in any concatenation

Decoder 2
Adress the limits



Task: **lama poids mille**

Target: **[lama] [pwa] [mil]**

Utterance: **lama pwa mil**

True Score: **[1, 1, 1]**

Threshold: **t = 2**

Missing space → Segment = **lama pwamil**

1. Segmentation: [lama] [pwamil]
2. Possible candidates: [lama] [bwamim, bwamam, …, pwamim, pwamam, pwamil, pwamal]
3. Concatenation: [lamabwamim, lamabwamam, …, lamapwamal]
4. Matching & Result: **[1, 1, 1]**

Extra space → Segment = **lama pw a mil**

1. Segmentation: [lama] [pw] [a] [mil]
2. Possible candidates: [lama] [pw, bw] [a], [mim, mil, mam, mal]
3. Concatenation: [lamapwa, lamabwa] [pwamim, bwamim, pwamil, bwamil, pwamam, …]
4. Matching & Result: **[1, 1, 1]**

Decoder 2 limits

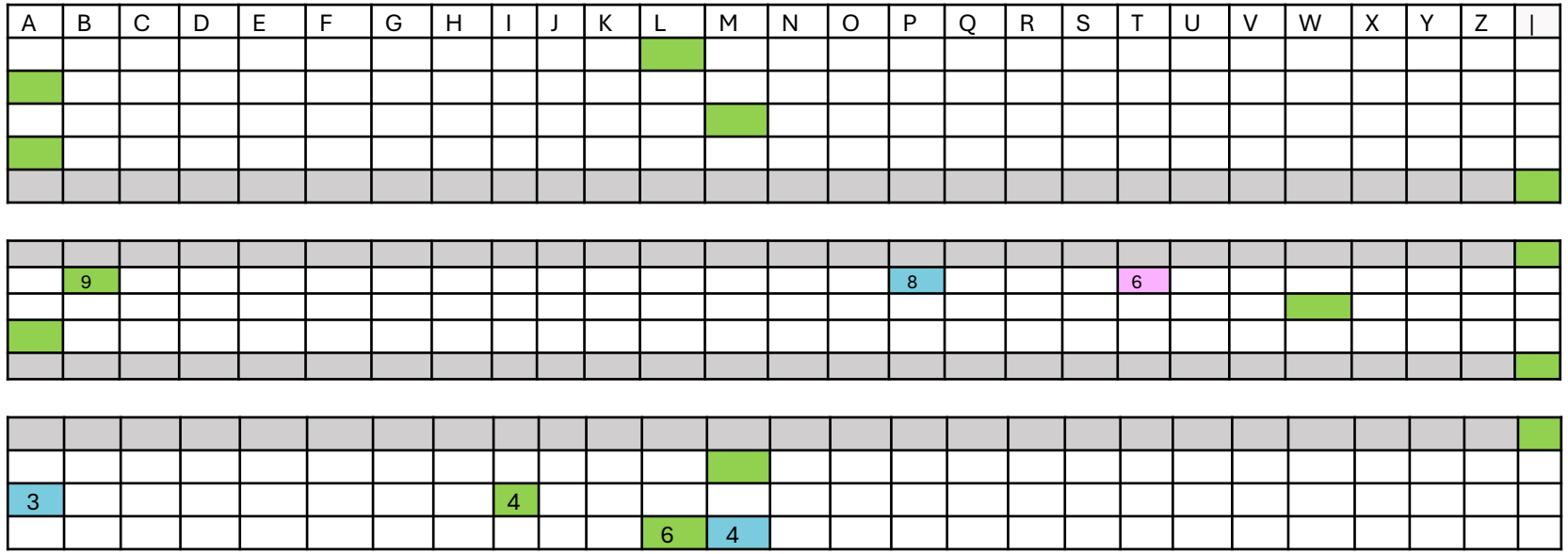
Task: **poids mille aout**

Target: **[lama] [pwa] [ut]**

Utterance: **lama pwad aut**

True Score: **[0, 1, 0]**

Threshold: **t = 2**

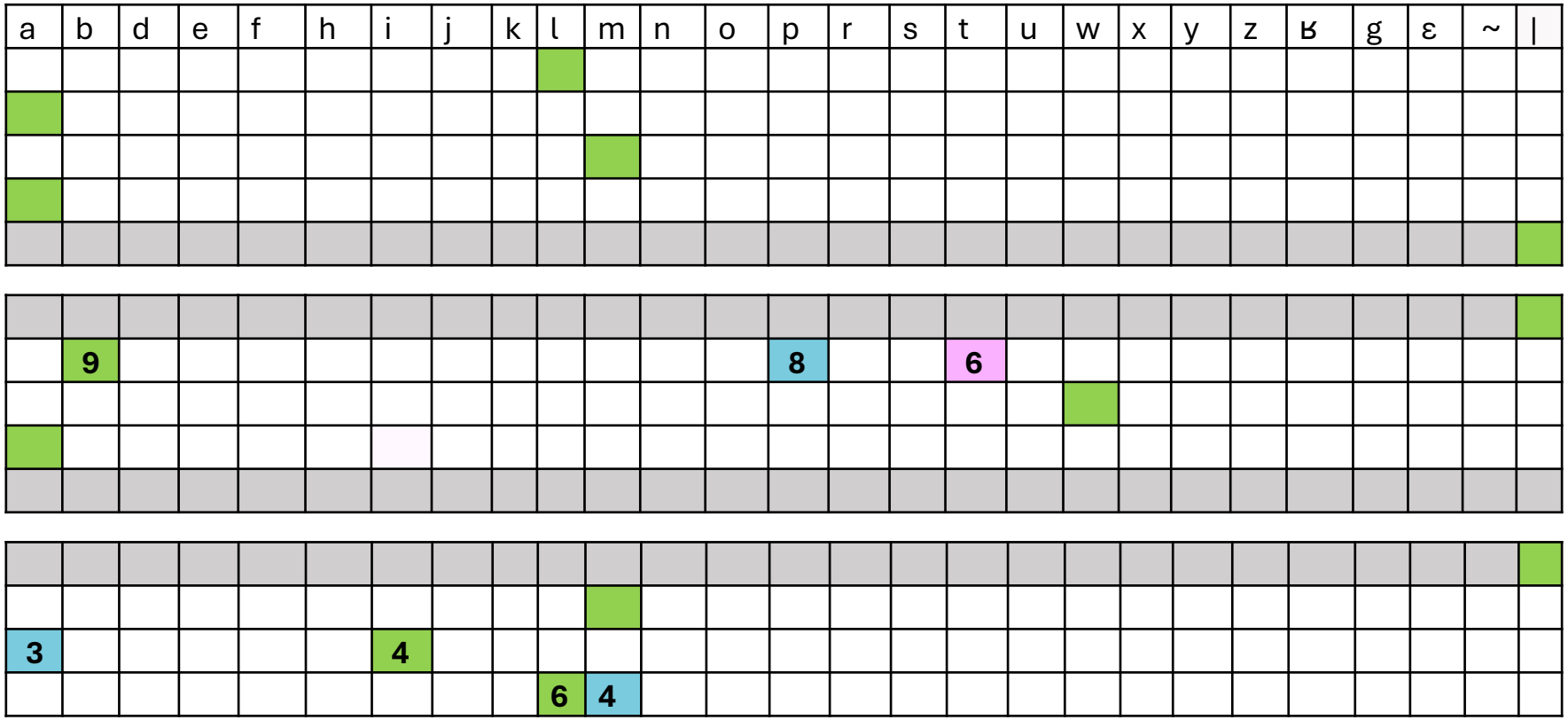


Problem: Bad if mispronunciation is at the beginning or end of a word

Segment = **pw ad mil aut**

1. Segmentation: [pw] [ad] [mil] [aut]
2. Possible candidates: [pw, bw] [ad] [mil] [aut]
3. Concatenation: [pwadmil, bwadmil], [admilaut]
4. Matching & Result: **[1, 1, 1]**

Decoder 2

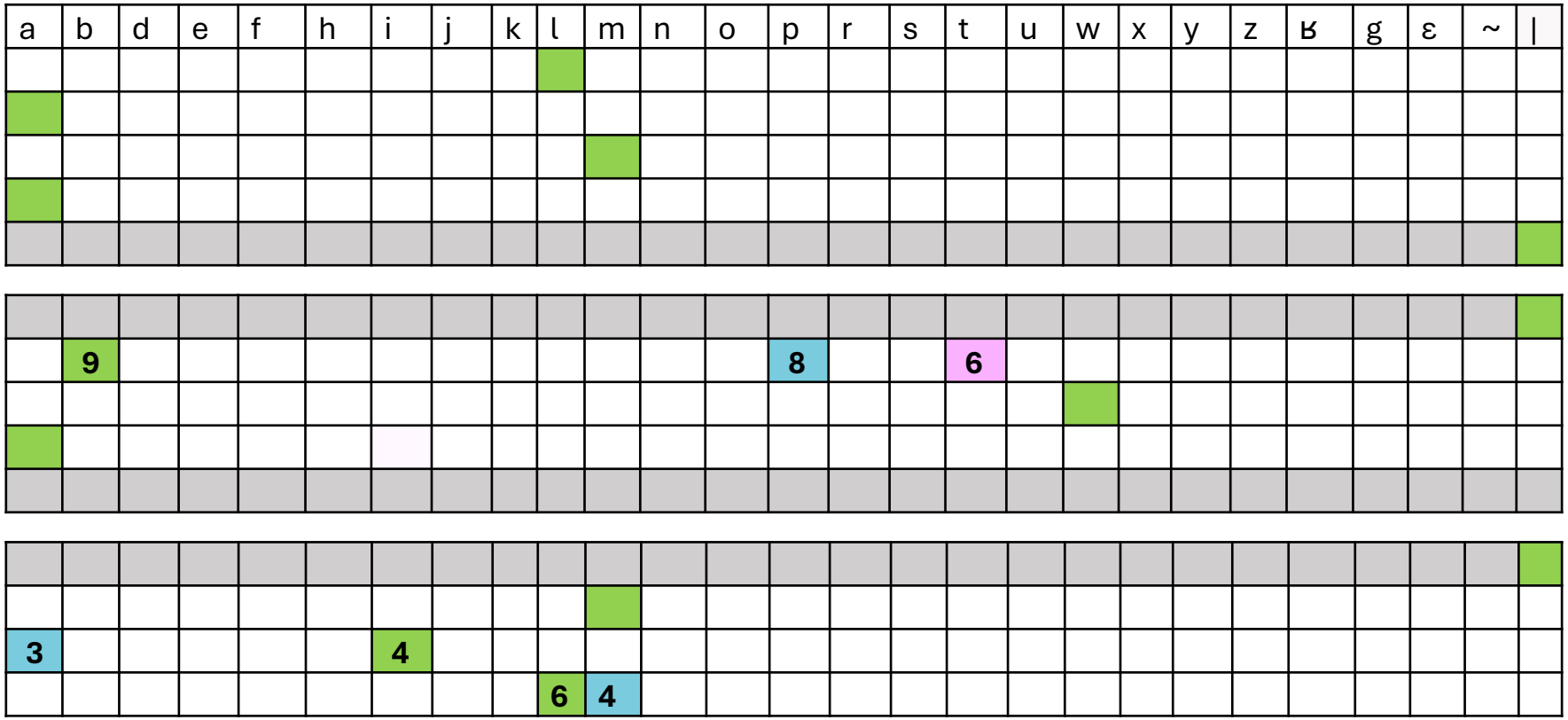


Problem: Bad if mispronunciation is at the beginning or end of a word

1. Segment logit matrix based on char offset
2. For each timeframe select phonemes within threshold **t** of top **k** candidate
3. Generate all plausible phoneme sequences (Cap predictions at **M** per submatrix)
4. Create a sliding window of 3 concatenated predictions
5. Check if target is contained in any concatenation

Decoder 3

Create Error List Beforehand !



Improvement: Handling silent phonemes pronounced at borders

1. Segment logit matrix based on char offset
2. For each timeframe select phonemes within threshold **t** of top **k** candidate
3. Generate all plausible phoneme sequences (Cap predictions at **M** per submatrix)
4. Create a sliding window of 3 concatenated predictions
5. Check if any error is contained in any concatenation
6. Check if target is contained in any concatenation

Decoder 3: Creating the Error List

Create Error List Beforehand !

Decoder 3
Adress the limits

Task: **poids mille aout**

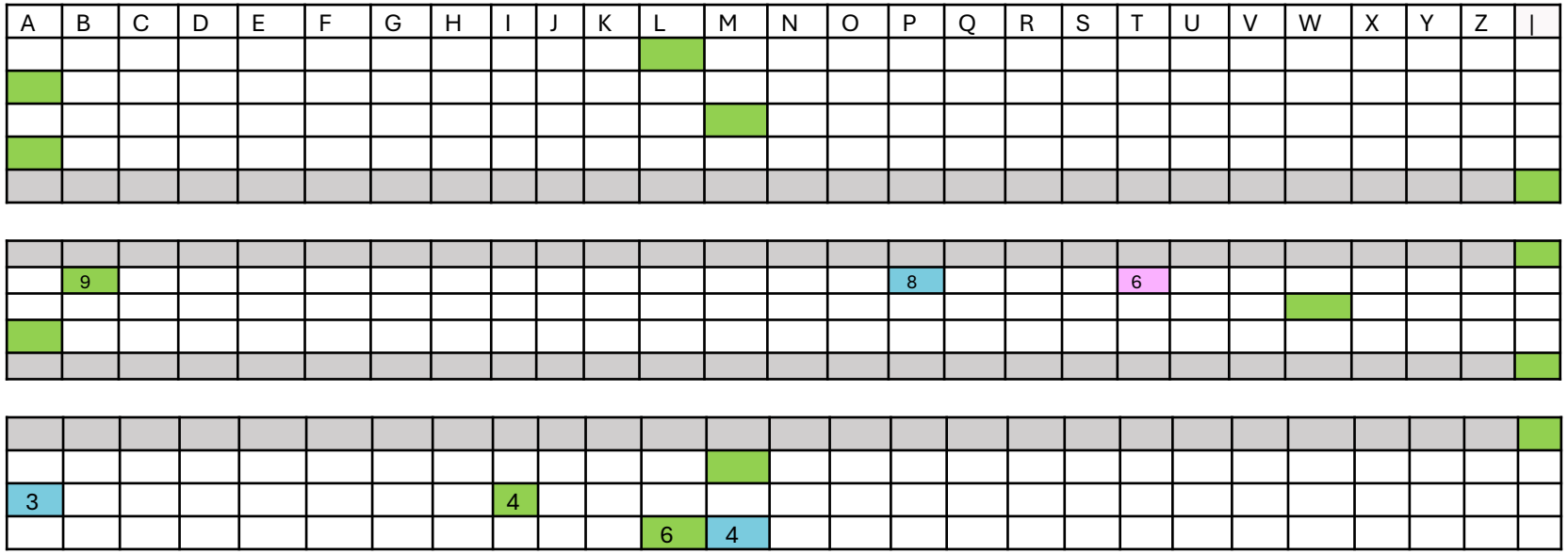
Target: **[lama] [pwa] [ut]**

Error: **[] [pwad] [aut]**

Utterance: **lama pwad aut**

True Score: **[0, 1, 0]**

Threshold: **t = 2**



Mispronunciation at the beginning or end of a word

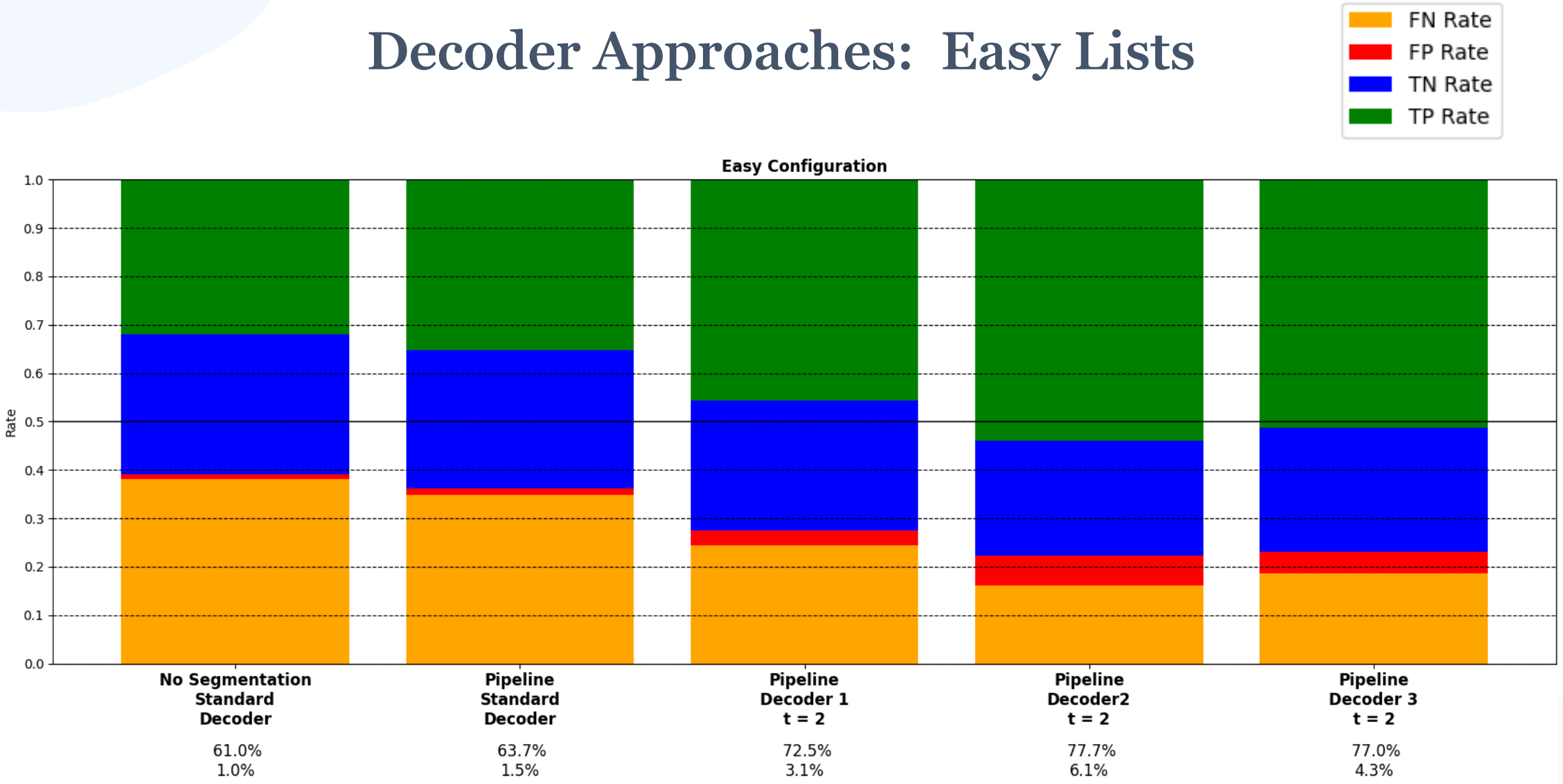
Segment = **pw ad mil aut**

1. Segmentation: [pw] [ad] [mil] [aut]
2. Possible candidates: [pw, bw] [ad] [mil] [aut]
3. Concatenation: [pwadmil, bwadmil], [admilaut]
4. Check if any error is contained in any concatenation → pwad + aut
5. Matching & Result: **[0, 1, 0]**

Decoder Evolution

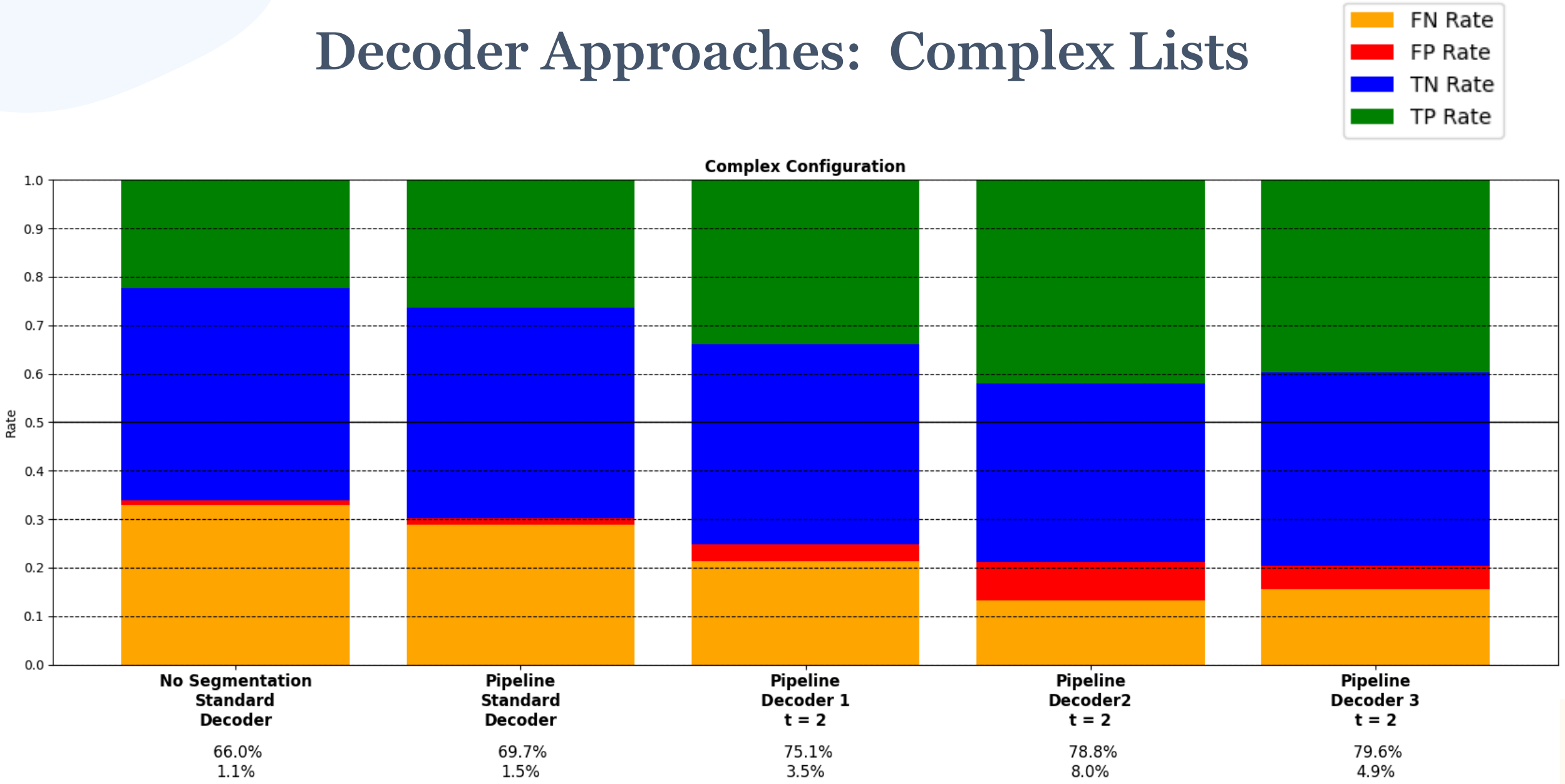
- Standard Wav2Vec2 Decoder: Limitations in handling ambiguities
- Decoder 1: Considering multiple probable outputs
- Decoder 2: Improved handling of item boundaries
- Decoder 3: Improved error detection at boundaries

Decoder Approaches: Easy Lists



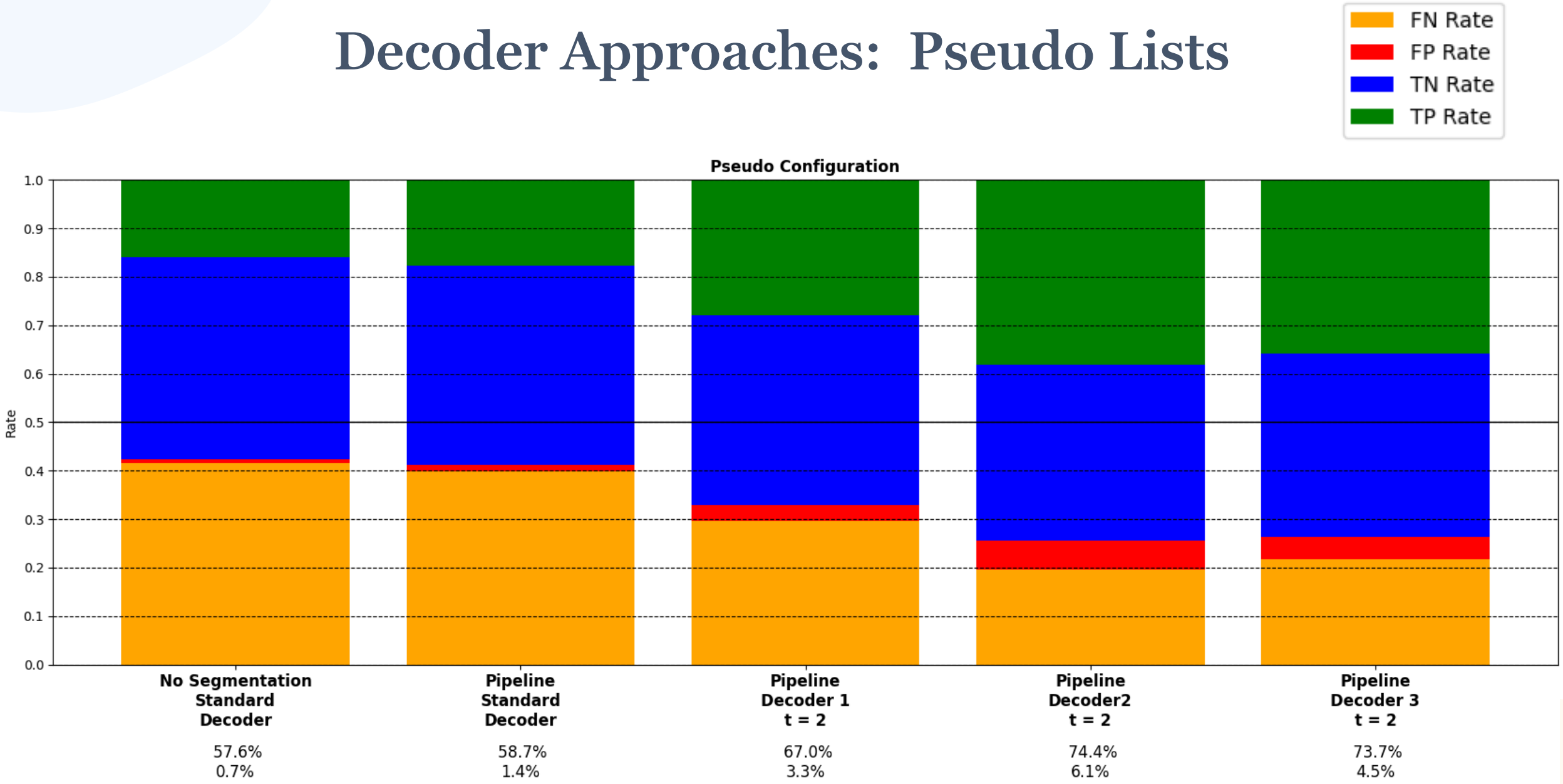
+16.0 %

Decoder Approaches: Complex Lists



+13.6 %

Decoder Approaches: Pseudo Lists



+16.1 %

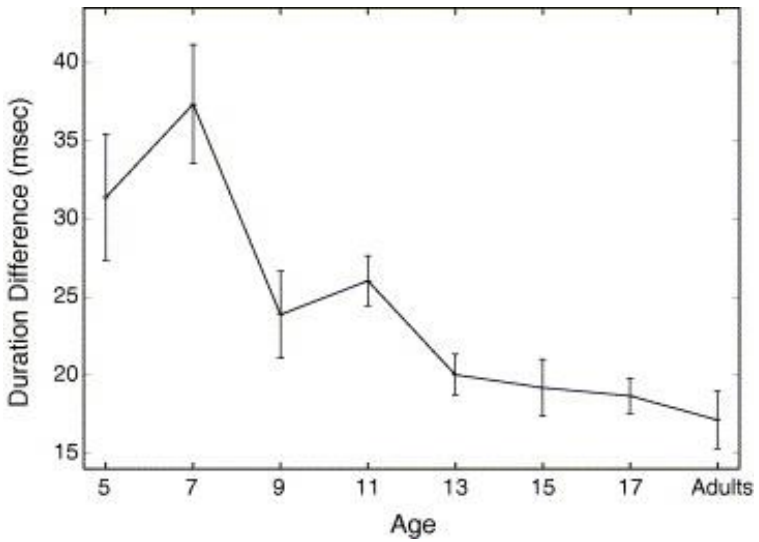
ASR for Speech Evaluation Pipeline: Final Results

- **+ 15.23%** over standard Wav2Vec2 Phoneme
- **+ 12.6 %** over standard decoder
- FP rate < 5%
- Effective in handling variability of read child speech
- Threshold **t** → Used for many speech eval settings
- Data-agnostic
- Easy to adapt

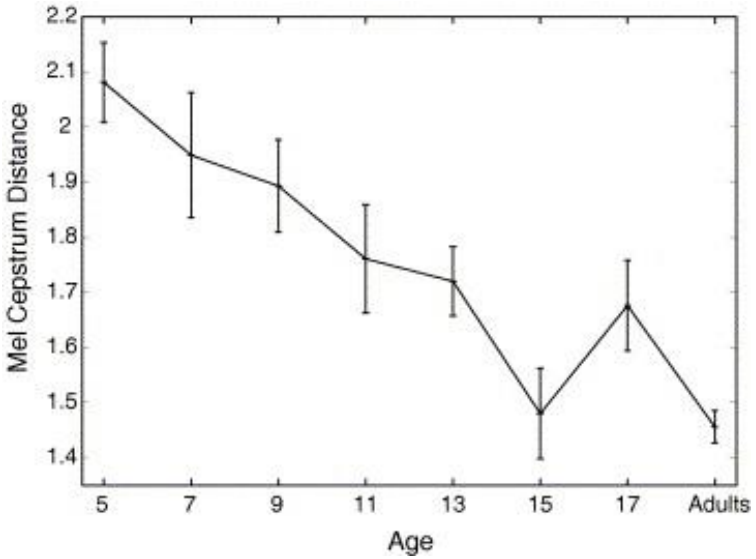
Thank you !

Question ?

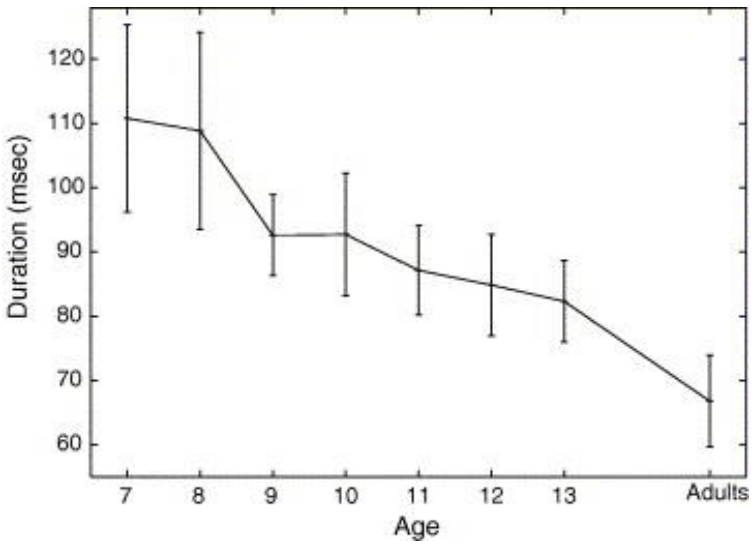
Intra- & Inter-Speaker Differences in Phoneme Variations



Temporal difference between cor-responding vowels in two repetitions of the same sentence



Spectral difference between corre-sponding vowels in two repetitions of the same sentence



Mean phone duration for children of different ages and adults

Taken from the "Acoustic variability and automatic recognition of children's speech" paper

Configuration	Num Participants	Duration (minutes)
config_A_easy_1	47	17
config_A_easy_2	47	18
config_A_complex_1	47	17
config_A_complex_2	47	22
config_A_pseudo_1	47	18
config_A_pseudo_2	47	22
config_B_easy_1	55	25
config_B_easy_2	55	25
config_B_complex_1	55	26
config_B_complex_2	55	27
config_B_pseudo_1	55	22
config_B_pseudo_2	55	27
config_C_easy_1	53	21
config_C_easy_2	53	20
config_C_complex_1	53	23
config_C_complex_2	53	28
config_C_pseudo_1	53	21
config_C_pseudo_2	53	26

Table 2: Number of participants and minutes of data per Configuration ID

Future Work

- Refinement of Wav2Vec2 Phonemizer model
- Comparative studies using enhanced models
- Evaluation on adult speech datasets (if available)
- Addressing scoring mismatches

Limitations

- Discrepancies between professional scoring and ASR operations
- Challenges with insertions within items
- Difficulties in handling sounding out followed by incorrect pronunciation
- Issues with accentuated pronunciation of final ’e’ in items