Université Paris Cité

Nicolas Ballier

‘Calibrating’ Whisper probability scores for subtokens with Phonetic posteriorgrams

Grenoble, 6th March 2026 LIDILEM / Maison des Langues

Joint research with Maelle Amand, Taylor Arnold, Maelle Bourbon, Léa Burin, Tori Fullerton, Gina-Anne Levow, Siyu Liang, Adrien Méli, Behnoosh Namdarzadeh, Sara Ng, Erin Pacquetet, Artem Saloev, Mehak Soleil, Chloé Scholent, Guillaume Wisniewski, Richard Wright & Jean-Baptiste Yunès (also made possible through a CNRS deputation at LLF)

Université Paris Cité

ALTAE
Langues et cultures connectées

Outline of the Presentation

- A quick introduction to Whisper
- Previous results on Whisper scoring method : how to use subtoken probability for scoring?
- The posteriorgram method : Preliminary results (Mehak Soleil): Calibrating Whisper probability scores for subtokens with Phonetic posteriorgrams : from probability-scoring to mispronounciation diagnosis
- The subtokenisation bias (encoder level) / the conditional probability bias (decoder level)
- Next steps and Discussion : Whisper vs. what's perceived in the data ?

- Segmental analysis : A roadmap (Janus WP2.1) for subtoken level scoring and potentially mispronunciation detection and diagnosis module
- Tiny vs. tiny.en models
- Fine-tuning suggestions

Whisper training (Radford 2023)

Multitask training data (680k hours)

English transcription

"Ask not what your country can do for ⋯"

Ask not what your country can do for ⋯

Any-to-English speech translation

"El rápido zorro marrón salta sobre ⋯"

The quick brown fox jumps over ⋯

Non-English transcription

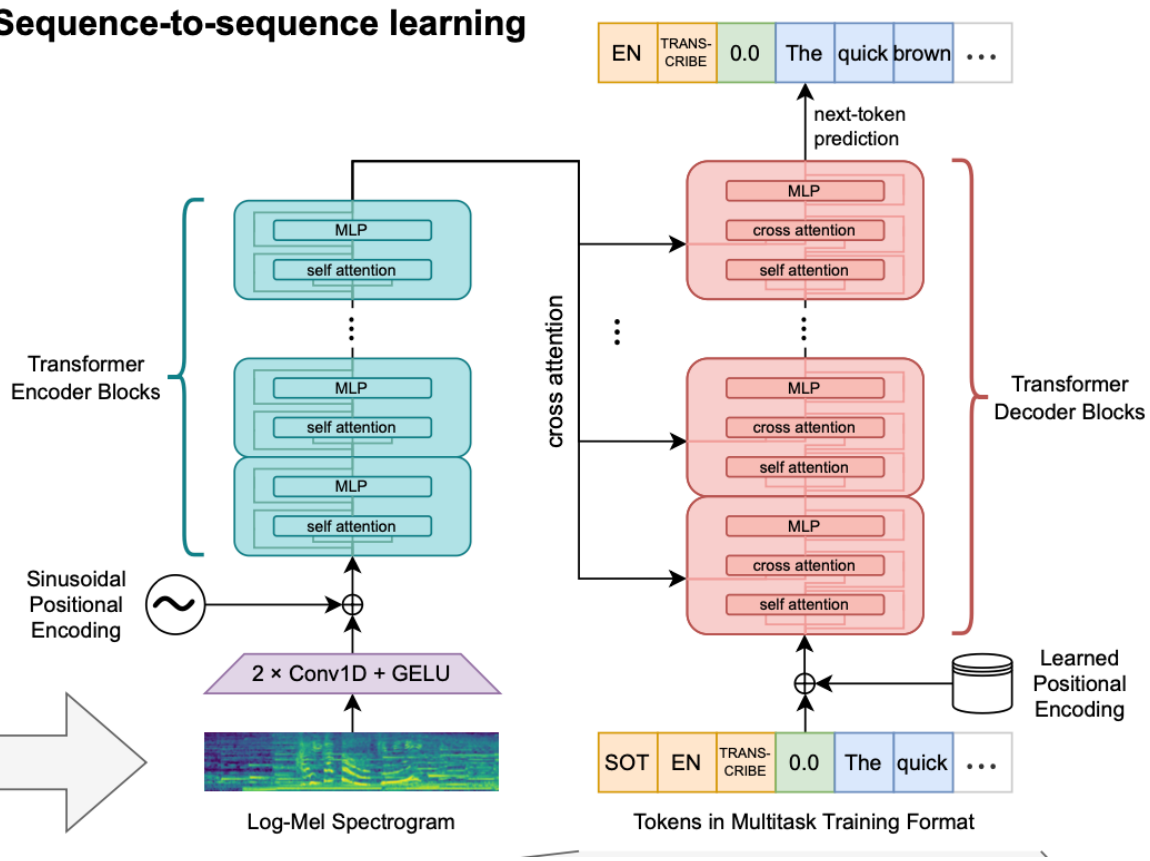
"언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯"

언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯

No speech

(background music playing)

∅



Main parameters of the Whisper models (Radford 2022 + Whisper github)

Size	Parameters	Required VRAM	Relative speed
tiny	39 M	1 GB	32x
base	74 M	1 GB	16x
small	244 M	2 GB	6x
medium	769 M	5 GB	2x
large	1550 M	10 GB	1x
large-v2	1550 M	10? GB	1?x

Table 1: Whisper models tested for this experiment

https://huggingface.co/models?search=openai/whisper

The large-v3 model is trained on 1 million hours of weakly labeled audio and 4 million hours of pseudo-labeled audio collected using large-v2.
https://github.com/openai/whisper/discussions/1762

Initial intuition working on translation and transcription : Interlanguage Retranscription & Named Entity Recognition (NER) Issues

Chomsky

- expected model /ˈtʃɒmski/ French realisation [ʃɔmski] for <Chomsky>
- Different interpretations of different models:
- *Je me ski* (tiny)
- *J'aime ce qui* (base)
- *James Key* (medium)
- *Jomski* (large)
- *Jamsky* (small/large-v2)

Ballier, N. Namdarzadeh, B. Zimina, M. and Yunès, J.-B. (2023) Translating Dislocations or Parentheticals : Investigating the Role of Prosodic Boundaries for Spoken Language Translation of French into English, Proceedings of Machine Translation Summit XIX Vol. 2: Users Track, 119-132. https://files.sciconf.cn/upload/file/20230827/20230827195133_32318.pdf

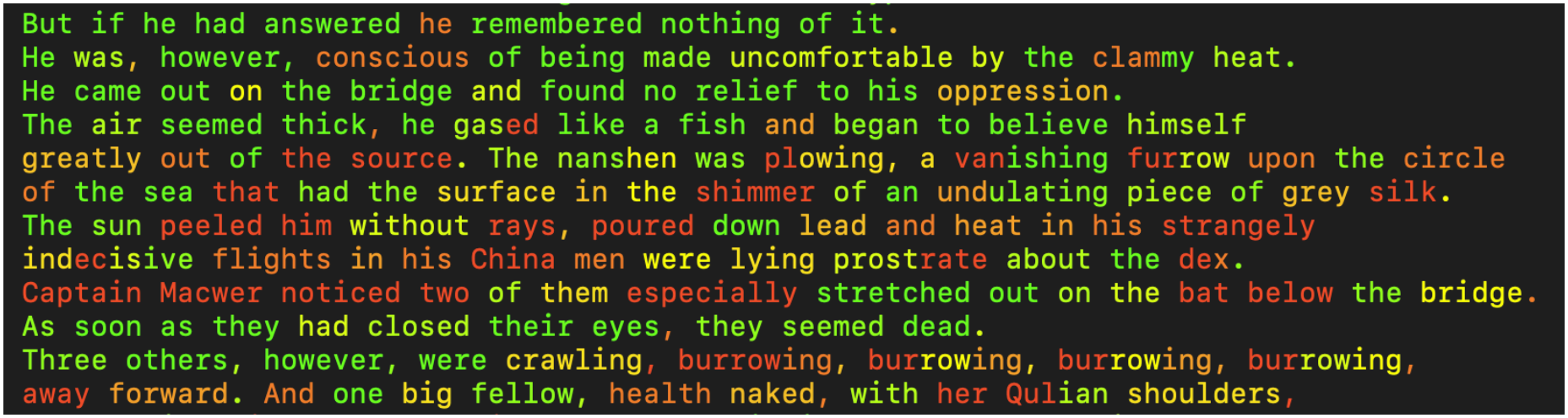
Plausible uses of Whisper for segmental analysis

- Language detection feature for A1 identification (work in progress for A2)
- average subtoken probability score for level/CEFR correlates
- Levenshtein distance as robust measure / correlate to levels
- Tiny/ tiny.en more sensitive to learner variation

- To be more systematically tested for spontaneous speech : Delta between tiny.en (sensitivity to distorsion) and medium for « reference » transcription

-> Papers on global scoring

Analysing confidence scores with C++ implementation of Whisper (Gerganov 2022)



Herculean

https://github.com/ggerganov/whisper.cpp https://github.com/jbyunes/whisper.cpp

Tiny.en

```
[00:00:00.000 --> 00:00:09.920]  [_BEG_] This is the story of how to manage the top of the Ever
est on the 29th of May 1953 and come back safely to their friends below.[_TT_496]
```

tiny

```
[00:00:00.000 --> 00:00:10.000]  [_BEG_] This is the story of how to manage the top of the Everes
t on the 29th of May 1953 and come back safely to their friends below.[_TT_500]
```

large v3

```
[00:00:00.040 --> 00:00:10.260]  [_PREV_][_PREV_][_PREV_][_PREV_][_PREV_][_PREV_][_PREV_][_PREV_][_
PREV_][_PREV_][_PREV_][_PREV_][_PREV_][_PREV_][_PREV_][_PREV_][_PREV_] wake the[_TT_513]
```

largev1

```
[00:00:00.000 --> 00:00:07.400]  [_BEG_] This is the story of how two men reached the top of the Ev
erest on the 29th of May 1953[_TT_370]
[00:00:07.400 --> 00:00:10.000]   and come back safely to their friends below.[_TT_500]
```

Probing Whisper scores with C++ implementation

```
[_BEG_] 0.977773    0   0
 mais   0.429366    0   24
 je 0.988376    24  36
 rev    0.997742    36  54
iens    0.995006    54  78
 sur    0.992805    78  95
 ce 0.821359    95  108
 problème   0.991321    110 164
 qui    0.69173 164 180
 est    0.973068    180 196
 un 0.979191    196 207
 problème   0.966143    213 284
,   0.368668    284 284
[_TT_142]   0.0282332   284 284
 voilà  0.786231    284 319
,   0.610747    319 330
 d  0.965854    330 336
'   0.999668    336 339
être    0.999371    346 370
 chez   0.997543    374 394
 moi    0.993733    394 415
,   0.576415    416 416
 combien    0.698848    424 444
```

Reverse engineering : the meaning of Whisper's special subtokens

- • 50,255 linguistic subtokens, corresponding to English words or fragments for French or graphemes for languages like Persian;
- • special tokens, some of them corresponding to boundaries of the Transformer: the end of text and end of sentence subtokens 50257 *[_EOT_]* and 50258 *[_SOT_]*;
- • 100 extra-tokens labelled *[_ex- tra_token_50259]* to *[_extra_token_50359]*;
- • 7 special tokens are also acknowledged in the literature such as 50360 *[_SOLM_]*, 50361 *[_PREV_]*, 50362 *[_NOSP_]*, 50363 *[_NOT_]* and 50364 *[_BEG_]*. *[_BEG_]* cor- responds to the beginning of the 30 second window when the sound file is processed by Whisper;
- • 1,500 out-of-vocabulary OOV subtokens from *[_TT_1]* to *[_TT_1500]*. they correspond to temporal subtokens

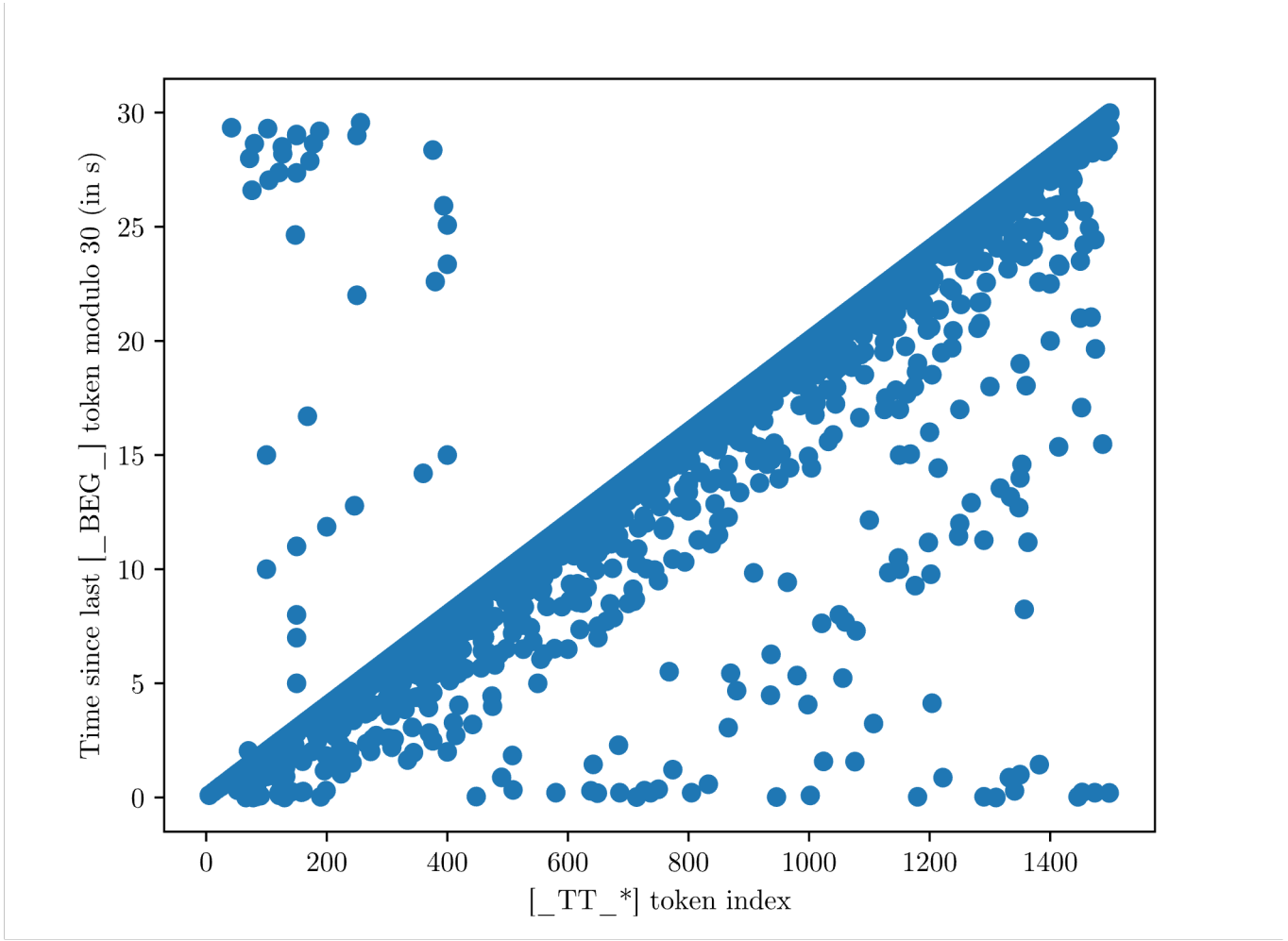


Figure 4: Token indices vs modulated time

MAIN RESULTS (3 papers in one slide)

corpora

- ANGLISH levels predicted
- ISLE levels predicted

tasks

- Language identification task (probability of identification English / L1)
- Mean scores for transcription task

- Metrics : Levensthein distance to expected transcription

Scoring the ANGLISH corpus

Table 5 Means and standard error per level in the ANGLISH data

Group	Mu	SE
FR1	0.87	0.01
FR2	0.89	0.01
GB	0.94	0.00

Scoring the ANGLISH corpus

Table 6 Confusion matrix of the prediction of levels with the algorithm k-means with k = 3 based on linguistic subtokens

Pred	Group		
	FR1	FR2	GB
FR1	13	6	2
FR2	5	11	0
GB	2	3	18

« affordance » : ability to capture (mis)realisations locally at the subtoken level

	1.Je me ski:	ʒ(ə)m(ə)ski
	2.J'aime ce qui:	ʒɛms(ə)ki
	3.James Key:]	ʒɛmski
[ʃɔmski]	4.Jomski:	ʒɔmski
	5.Jamsky:	ʒamski

Holistic probability scores vs. Detailed scores for subtokens

-Different phonetic-subtoken mappings for different models

Graphematic affordance: what's in the graphemic representation ('holes' in the Whisper dictionary / net)?

- JANUS WP 2.1
- (pilot) phonological neighbour density

Phonological neighbourhood density (WiP)

9	**509**	You	33	cou, You, pou, yo, you, vou, sou, Lou, Yo, Dou, lou, bou, Hou, yol, yog, rou, Yok, gou, Vou, YOUR, Cou, Rou, Nou, Tou, Sou, fou,
10	**510**	here	38	were, where, her, There, Her, hero, phere, here, hers, mere, bere, Where, Hero, there, dere, vere, hele, Bere, Here, gere, Herz, ere, Hee,
11	**511**	her	61	ber, per, fer, ther, he, her, ier, mer, der, er, ner, ger, wer, ER, hr, cer, Her, Er, zer, He, uer, TER, ker, har, here, cher, Hey, yer, hern, hes, jer, hee, ER,
12	**512**	some	14	somet, come, same, home, esome, som, Somet, Home, Some, dome, sme, somm, Sole, COME
13	**513**	oug	25	ong, ous, ough, og, ug, oup, oun, oud, oul, org, OU, oung, Our, OUT, Out, Ug, OUR, zug, oux, ogg, oue, jug, bug, OG, OUL
14	**514**	ak	75	ah, ck, ag, ap, K, ake, ank, alk, av, ark, ek, az, ik, ai, au, aw, AN, aj, AS, AL, ask, AY, aa, AC, AP, sk, AA, aks, akt, AD, aki, An, aka, ae, mak, AB, al
15	**515**	ard	52	are, ars, ord, ark, ary, ward, arn, ird, ared, ari, aud, ald, arm, And, AD, arp, arl, erd, rd, ARR, AND, Are, ORD, yard, ART, Aud, Ad, aru, uard, aid, ha
16	**516**	going	5	doing, going, Doing, goin, Gong
17	**517**	un	94	us, U, und, In, fun, um, An, run, On, unt, oun, US, sun, Um, gun, unf, UN, Uh, ur, tun, pun, Us, AN, Up, Sun, unc, bun, IN, UK, hu
18	**518**	ment	20	ent, ments, ient, ement, rent, men, ment, Ent, mente, met, gent, zent, nent, mont, Men, Ment, Ent, meno, mens, sent
19	**519**	think	8	thing, Thank, thank, thick, thin, thinks, think, Thing
20	**520**	pe	98	te, fe, pr, ye, He, pre, per, po, spe, ke, Ye, Be, ph, Se, ope, ve, Re, De, Le, je, ge, pie, Ne, ce, pa, Per, Ke, ple, Pa, pen, Spe, Fe
21	**521**	end	41	ond, ens, end, und, iend, ene, eng, enn, endo, ena, rend, ED, eed, And, ened, ende, eno, eld, End, enda, erd, AND, ND, UND, ENN, eni, En, pen
22	**522**	(	61	J, ë, [, 2, K, U, â, Ã, V, z, ê, i, 3, ×, ', Ñ, 4, 5, Î, í, Z, Q, Ø, 6, Ù, 7, 8, 9, X, Â, $, *, ?, ,, #, &,], Å, +, =, -(,), %, Õ, ((, (", |,
23	**523**	cause	4	cause, lause, ause, caust
24	**524**	tim	56	time, im, him, sim, tem, tit, Sim, tip, Im, dim, Tom, Kim, Him, aim, Time, Jim, tie, tam, Tem, tym, til, Tam, ti, tir, Tit, Tik, tin, rim, L
25	**525**	ast	54	ost, act, ass, ase, St, ait, ash, aut, att, AS, alt, cast, ask, rast, St, asc, ST, ast, asy, akt, ST, fast, agt, asi, EST, adt, asm, ART, last, amt, At, Ass, U

WiP : « phonetic » neighbours in alternative predictions

[_BEG_]	0.947713	0	0	[_TT_12]	0.00670132	0	0	[_TT_11]	0.00531413
Obs	0.43384	6	7	observing	0.281372	6	7	"	0.114933
erving	0.995759	31	78	er	0.00194947	31	78	erve	0.000467226
the	0.990482	78	104	The	0.00234761	78	104	a	0.00119722
steady	0.944887	104	153	study	0.0333635	104	153	Stead	0.00799184
fall	0.961571	168	191	Fall	0.00802978	168	191	fall	0.00721409
of	0.993961	191	199	the	0.00107607	191	199	in	0.000427133
the	0.969816	209	234	Bar	0.00718367	209	234	bar	0.00361852
bar	0.426446	234	260	Bar	0.33805	234	260	b	0.0446175
ometer	0.937619	260	307	omet	0.0159721	260	307	o	0.0136897
,	0.871901	323	323	Captain	0.0520907	323	323	kept	0.00540585
Captain	0.867363	363	379	captain	0.0336688	363	379	Cap	0.00512762
Mack	0.321873	392	410	Mac	0.179011	392	410	Mag	0.0842532
worth	0.510859	410	446	wer	0.105311	410	446	were	0.0928428
thought	0.727912	446	496	fought	0.212631	446	496	followed	0.00587244
,	0.526715	501	502	there	0.308247	501	502	"	0.028238
there	0.580201	553	553	"	0.135482	553	553	[_TT_250]	0.013906

Phonetic sensitivity : subtoken transcription robustness

- Pilot study : VOT (Ballier & Fullerton, 2024, Fullerton & Ballier, to be resubmitted)

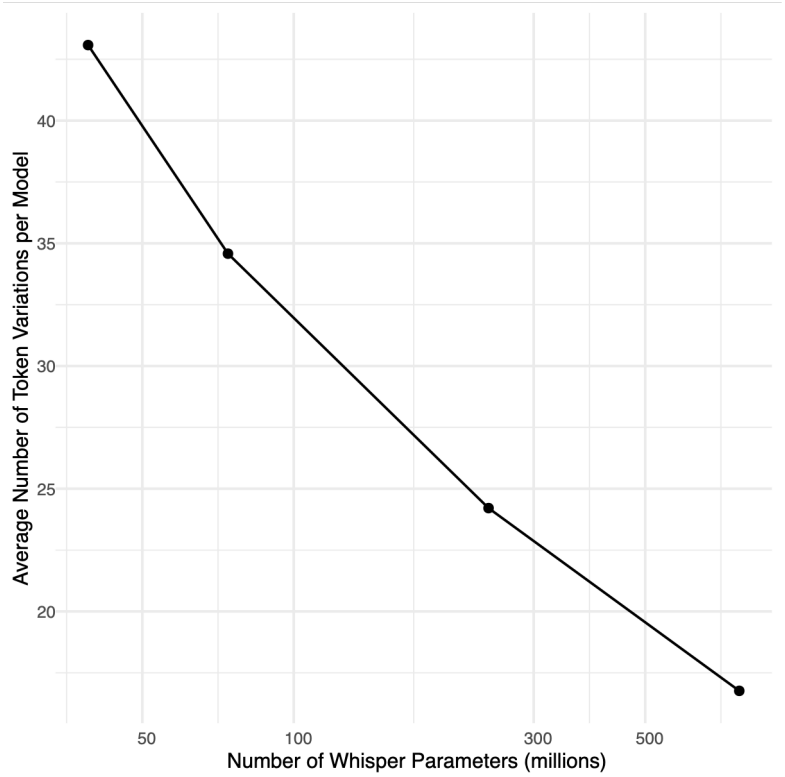
- Calibration studies on the signal-to-subtoken mapping : investigating the paradigm: multilingual vs. Native model sensitivity (retranscriptions of the same .wav input)

Using probability as a proxy (work in Progress: Maelle Bourbon & colleagues)

Model sensitivity to compounds (Fullerton, in progress)

model	avg_correct	avg_prob	n
<chr>	<dbl>	<dbl>	<int>
small	0.389	0.123	36
medium	0.371	0.314	35
medium.en	0.361	0.196	36
tiny	0.361	0.271	36
tiny.en	0.143	0.0693	35
large-v2	0.139	0.282	36
small.en	0.139	0.453	36
base	0.0833	0.254	36
base.en	0.0556	0.398	36
large	0.0556	0.311	36
large-v1	0.0556	0.311	36
large-v3	0.0294	0.0973	34

Role of Size in models for sensitivity?
(character error rate)



INCLSP2023

Disc: Speech tokenisers and the issue of discretisation of speech / descriptors

- « criterial feature » (Hawkins & Buttery, 2010) for speech ??
- WER as criterial feature -> Subtoken error rate as feedback ?
- From WER to phonological features (Stafford et al. submitted)

- Discrete speech phenomenon for CEFR boundaries ?
- RQ Matching speech (sub)tokens with criterial feature?
- Subtoken prediction and accuracy of the prediction : CALIBRATION
- -> subtoken (vs. Phone representation)

Model calibration

(Ballier et al. 2024)

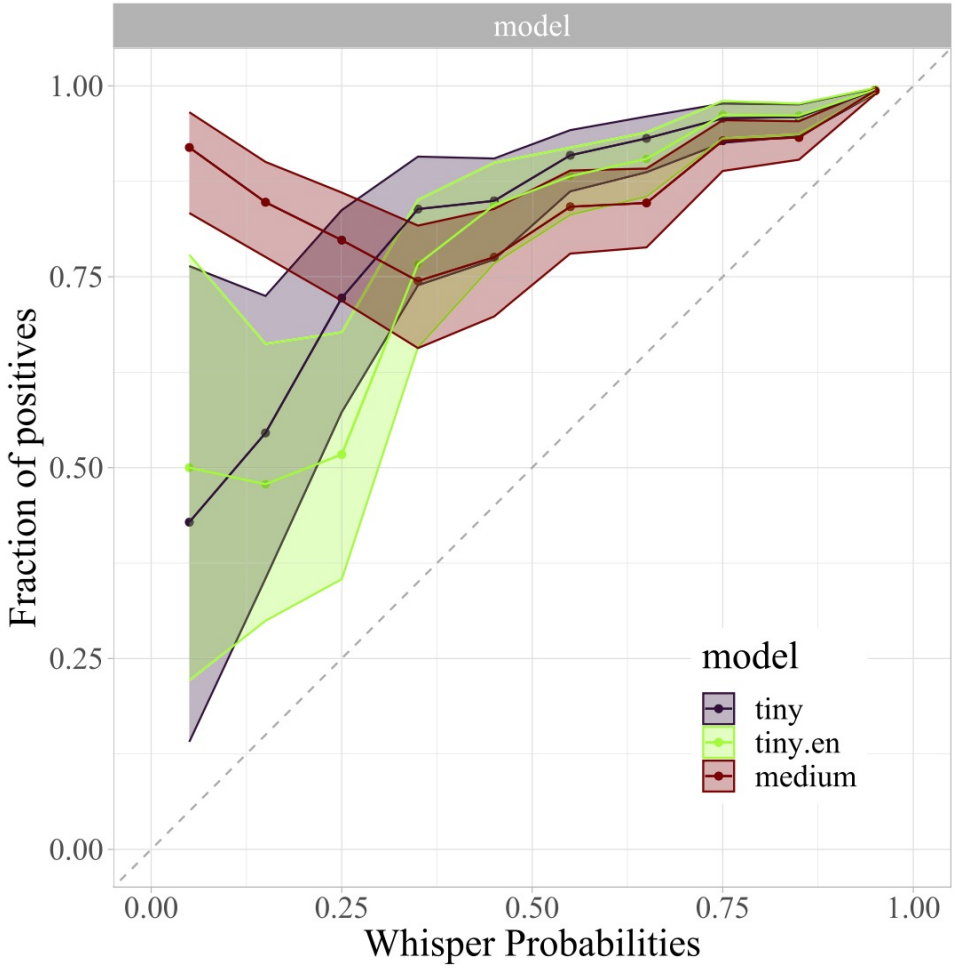
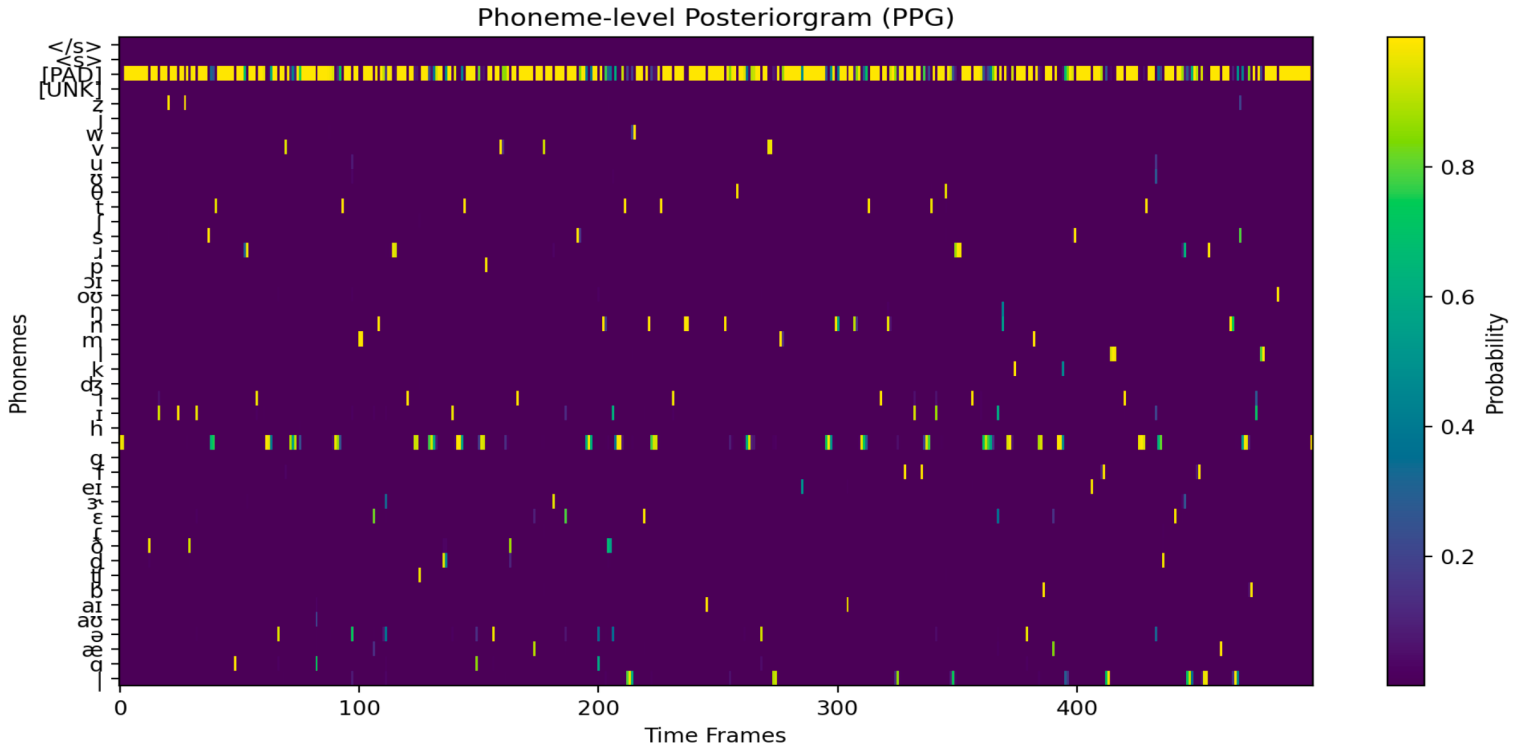


Fig. 4 Calibration curve for three Whisper models for the transcription of the learner #003 from the ISLE corpus

The posteriorgram method for calibration??

- Morrison et al. 2024 : ProMonet library

The posteriorgam method (Morrison et al. 2024 : ProMonet library)



Posteriorgram outputs (probabilities)

Phone	0.040	0.120	0.200	0.280	0.360	0.440	0.520
|	1.41675e-05	8.025782e-06	1.1889407e-05	4.9290436e-05	6.353135e-05	5.6095734e-05	0.00015720323
ɑ	6.4684646e-06	5.0322196e-06	2.5595696e-06	1.2072704e-05	9.581176e-06	8.374674e-06	1.6998612e-05
æ	6.179055e-06	3.5160394e-06	1.9930599e-06	1.8697245e-05	3.342115e-05	1.7433247e-05	3.962736e-05
ə	1.6525522e-05	1.4157539e-05	8.236401e-06	1.9069099e-05	0.0003121891	2.0440755e-05	0.00043182672
aʊ	2.7179724e-06	5.4161404e-07	3.225302e-07	2.619914e-05	5.347693e-06	1.1503414e-05	1.5836924e-05
aɪ	5.2215933e-06	2.2750266e-06	1.291359e-06	1.3572434e-05	1.2055012e-05	1.1174292e-05	2.1428408e-05
b	8.557055e-06	1.2283849e-06	1.3532631e-06	0.00013904998	6.8305403e-06	1.4366044e-05	1.711816e-05
ʧ	5.858904e-06	3.0096345e-07	2.3019643e-07	1.539746e-05	8.5339425e-06	1.3186154e-05	1.584889e-05
d	1.6169854e-05	2.2551917e-06	3.0452347e-06	0.0043227645	1.0210753e-05	2.6559612e-05	2.819538e-05
ð	8.705647e-06	3.5581243e-06	4.515888e-06	0.24329224	1.4316149e-05	7.4184856e-05	6.375192e-05
ɾ	5.4942348e-06	2.8278134e-06	2.1249386e-06	8.578579e-05	2.5096406e-05	4.390361e-05	5.1906696e-05
ɛ	3.866231e-06	5.166748e-06	2.5549186e-06	3.944099e-05	0.00034792177	1.7871815e-05	0.00019141177
ɝ	8.147061e-06	3.2271455e-06	2.3377606e-06	1.5555206e-05	2.634221e-05	3.1412113e-05	6.670572e-05
eɪ	2.910147e-06	1.8000112e-06	9.178467e-07	2.5349114e-05	7.793983e-05	2.0191861e-05	4.930426e-05
f	4.2076836e-06	1.3058539e-06	8.620347e-07	1.8040157e-05	6.054404e-06	2.7812175e-05	2.8459272e-05
g	1.02922295e-05	1.4516426e-06	1.5207668e-06	0.00010273239	1.0083642e-05	1.7250602e-05	2.314935e-05
	0.49176887	2.6278572e-05	0.0002658261	0.000110059584	7.7537e-05	8.332099e-05	8.164508e-05
h	9.019695e-06	5.1848056e-06	4.4042163e-06	0.00010186448	1.4635561e-05	2.006758e-05	2.4296573e-05
ɪ	2.3714358e-05	2.343887e-05	1.3432419e-05	3.1253545e-05	0.23521666	6.417903e-05	0.2480456
i	9.3563385e-06	8.278052e-06	4.1575618e-06	3.3993434e-05	0.013343346	4.492764e-05	0.0009398903

Predictions per time frame (20ms)

ŋ	oʊ	ɔɪ	p	ɹ	s	ʃ
1.544112092233263e-05	6.075181772757787e-06	5.1315596465428825e-06	2.1265992472763173e-05	8.99688802746823e-06	1.644592521188315e-05	1
9.503902219876181e-06	4.807642199011752e-06	2.495818534953287e-06	1.1387987797206733e-05	1.0222643140878063e-05	1.4482448932540137e-05	1
8.250188443525985e-07	1.4675269994768314e-06	2.8231522719579516e-07	1.0710507467592834e-06	8.516043635609094e-06	3.847515927191125e-06	2
1.552646494928922e-06	2.6670975330489455e-06	4.2191442162220483e-07	1.6400827007601038e-06	4.914290002488997e-06	5.896574748476269e-06	2
9.033171863848111e-07	2.1279915927152615e-06	4.275620142379921e-07	1.0919033002210199e-06	1.034534398058895e-05	6.21473191131372e-06	3
1.193460775539279e-06	2.2864098809805e-06	4.2907547026516113e-07	1.1577593568290467e-06	6.523353022203082e-06	5.50144477529102e-06	2
1.0131338967767078e-06	3.159894504278782e-06	4.5761956357637246e-07	1.015105908663827e-06	1.0528985512792133e-05	5.5850741773610935e-06	1
8.279184271486884e-07	1.8972001498696045e-06	4.0518148125556763e-07	9.705681804916821e-07	7.0763740041002166e-06	4.858017746300902e-06	2
8.827798865240766e-07	1.8799153167492477e-06	3.3696025525387086e-07	1.1157457038279972e-06	7.105594249878777e-06	4.224464191793231e-06	1
6.12147800893581e-07	1.3203982689447002e-06	2.6985347290064965e-07	8.491903713547799e-07	4.874432761425851e-06	3.4513122955104336e-06	1
6.841773370069859e-07	8.937168445299903e-07	2.2244672948090738e-07	7.872504852457496e-07	5.1881588660762645e-06	2.814881554513704e-06	1
7.067014280437434e-07	9.67960318121186e-07	2.8175108468531107e-07	7.424221166729694e-07	7.728494892944582e-06	2.587354401839548e-06	1
0.00014117256796453148	0.00010603619739413261	4.362903200672008e-05	0.00013202661648392677	0.00012429514026734978	0.00015499316214118153	0
5.417243755800882e-07	1.0493512263565208e-06	3.0401955086745147e-07	4.85877080791397e-07	1.0145362466573715e-05	2.6923150926450035e-06	1
6.95234632530628e-07	1.0997696335834917e-06	3.7298178767741774e-07	6.43767975816445e-07	1.493003946961835e-05	3.6982503388571786e-06	1

PPGs for English: three implementations, one pronunciation model

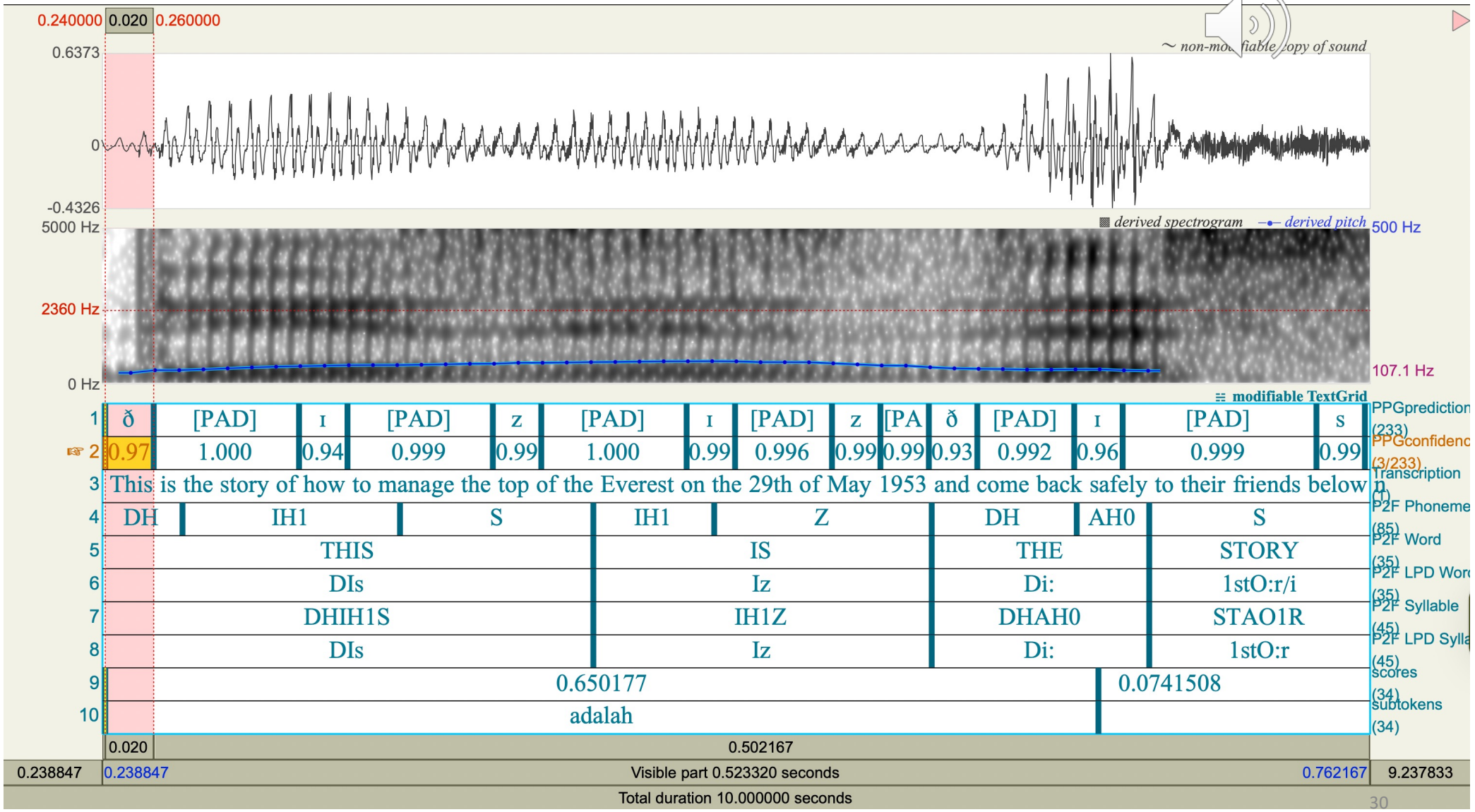
- TRAINING DATA : TIMIT

- GenAm Model

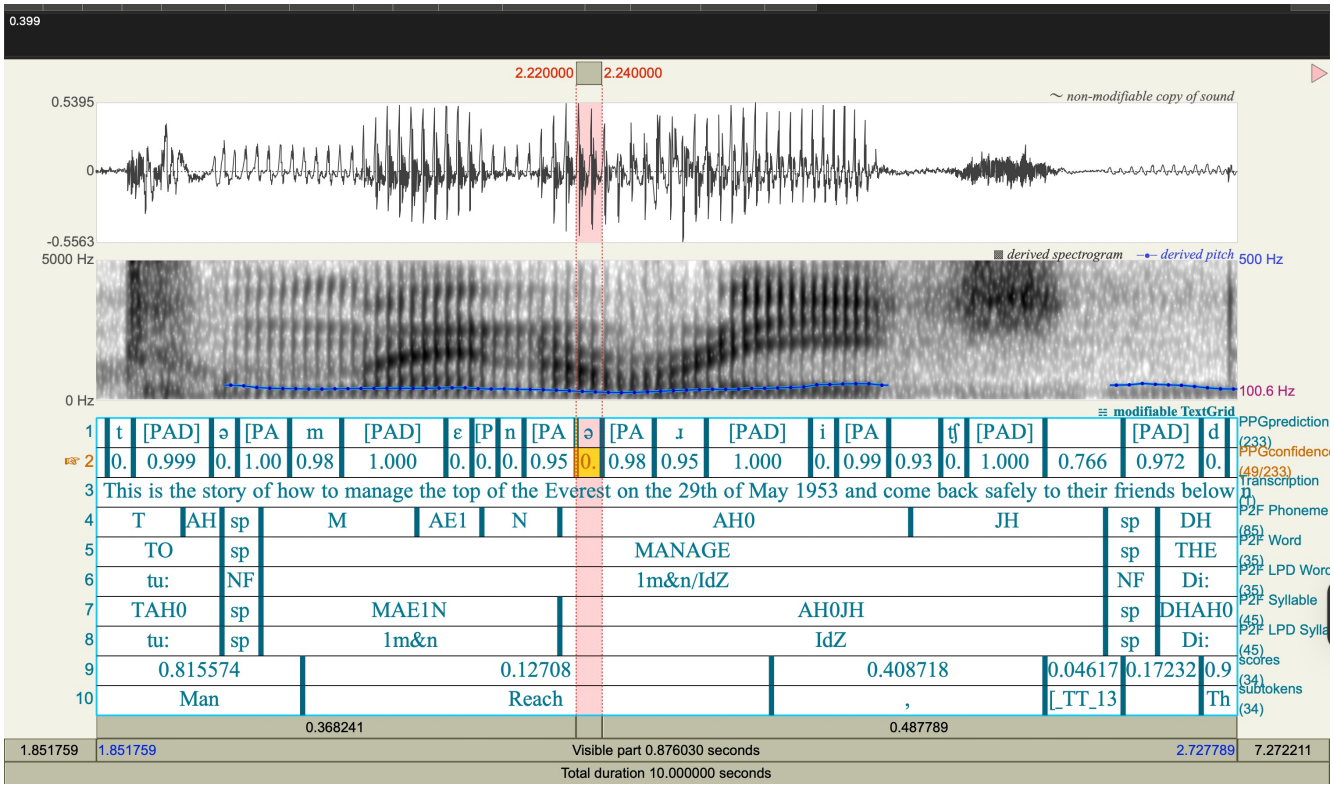
The pipeline

- Whisper scores -> time alignment -> Tier
- PPG -> 20 ms frames -> PPG prediction + scores tier
- Whisper -> syllable (PEASYV pipeline with PFA forced aligner, Ballier & Méli 2023)

- Comparison with https://plspp.univ-grenoble-alpes.fr/???


0.240000
0.020
0.260000
0.6373
0
-0.4326
~ non-modifiable copy of sound
5000 Hz
2360 Hz
0 Hz
derived spectrogram
derived pitch
500 Hz
107.1 Hz
modifiable TextGrid
PPGprediction (233)
PPGconfidenc (3/233)
Transcription (1)
P2F Phoneme (85)
P2F Word (35)
P2F LPD Word (35)
P2F Syllable (45)
P2F LPD Syll (45)
scores (34)
subtokens (34)
1 ð [PAD] ɪ [PAD] z [PAD] ɪ [PAD] z [PA ð [PAD] ɪ [PAD] s
2 0.97 1.000 0.94 0.999 0.99 1.000 0.99 0.996 0.99 0.99 0.93 0.992 0.96 0.999 0.99
3 This is the story of how to manage the top of the Everest on the 29th of May 1953 and come back safely to their friends below
4 DH IH1 S IH1 Z DH AH0 S
5 THIS IS THE STORY
6 DIs Iz Di: 1stO:r/i
7 DHIH1S IH1Z DHAH0 STAO1R
8 DIs Iz Di: 1stO:r
9 0.650177 0.0741508
10 adalah
0.020
0.502167
0.238847
0.238847
Visible part 0.523320 seconds
0.762167
9.237833
Total duration 10.000000 seconds

How two men reached / how to manage/ Out to man reach



Interim conclusions (Sohail, in progress)

- Heavier Padding than expected for posteriorgrams
- Alignment issues with a vengeance

(Top) PPG prediction

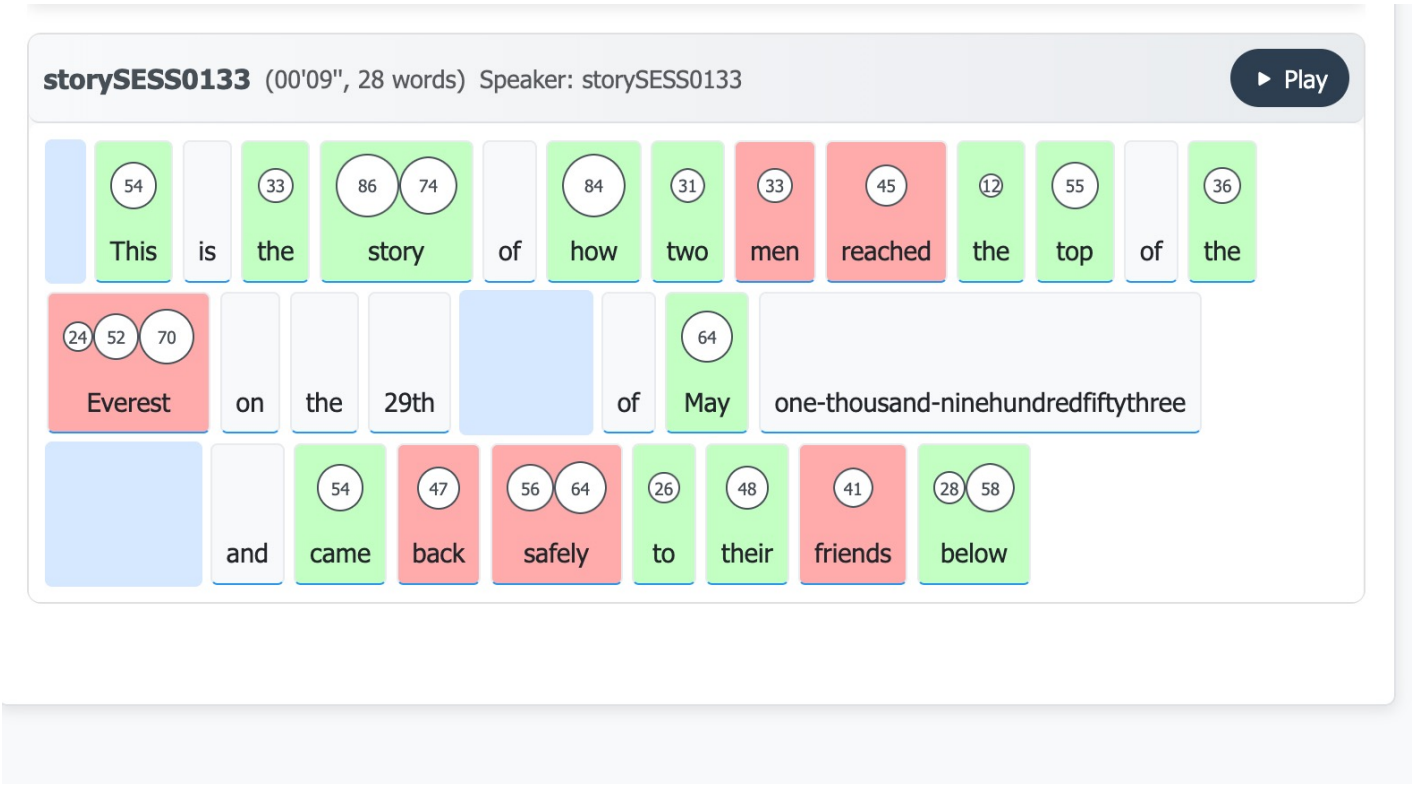
- This is the story of how two men reached

- two men -> [tə]
- /ri:tʃt/ -> [rɪtʃt]

['ðɪzɪzðɪs tɑɹi əv ɑ təmɛnəɹi ʧ
dɪ tɑ pəvðiævɜ˞ɛs ɑnðɪ t wɛn tinaɪnθ
əv meɪ naɪn tin fɪf tɪθ ɹi ɪn kəm bæ seɪf li
tə dɛɹ f ɹæn s bɪloʊ']

PPG_predictions
ð | ɪ | z | ɪ | z | ð | ɪ
ɑ | ɹ | i
ə | v
ɑ
ɑ
t | ə | m
m | ɛ | n | ə | ɹ
i | ʧ
d | ɪ
ɑ | p
ə
i
æ | v | ɜ˞ | ɛ
ɑ | n | ð | ɪ
t | w | ɛ | n | t | i | n | aɪ
n | θ | ə | v

Sylvain’s visualisation https://plspp.univ-grenoble-alpes.fr/



Stress pattern ??

Everest 'ev ər ɪst -əst, -ə rest

DISCUSSION 2: What should we flag?? Mispronunciation Detection and Diagnostic vs. Intelligibility issues

- of How (H dropping)
- two /tu:/ (= too) / tə/   (to)
- *rich* / *reached*   /ri:tʃt/ -> [rɪtʃt]
- affix dropping

MDD vs intelligibility ??

- safe /seɪf/ -> monophthong /sɛf/

NEXT STAGES

- Tokenisation issues (March 26th) workshop ->
- Calibrating Whisper scoring
- Probing the acoustic representations with n-best approach
- Phonological neighbourhood (cont.)

- Fine-tuning with subtoken re-allocation

N-best approach (tiny model)

- First
- 5
- Best
- Candidates

	I	J	K	L	M	N	O	P	Q
s	Subtoken1	Prob 1	Subtoken2	Prob 2	Subtoken3	Prob 3	Subtoken4	Prob 4	Subtoken5
1	[_BEG_]	0.93341	[_TT_8]	0.00911632	[_TT_4]	0.00528275	[_TT_10]	0.00455067	[_TT_12]
1	This	0.942185	this	0.0364853	"	0.0062538	The	0.00205148	These
1	is	0.995068	the	0.00156444	story	0.00098175	was	0.000626417	Is
1	the	0.972572	a	0.0175595	story	0.00546502	The	0.000860622	"
1	story	0.994918	Story	0.002873	history	0.000262246	stories	0.000260899	"
1	of	0.990276	how	0.0027305	about	0.00126481	,	0.000926382	to
1	how	0.917817	How	0.0418	"	0.0182314	'	0.00471078	HOW
1	to	0.825446	To	0.0556725	two	0.0275425	T	0.00984758	too
1	manage	0.398405	man	0.371861	men	0.112434	Man	0.0205292	outreach
1	the	0.675932	each	0.085194	to	0.0682969	reach	0.0594017	reached
1	top	0.919941	Top	0.0561463	tops	0.00319518	"	0.00273084	topic
1	of	0.986171	-	0.00382384	the	0.00224424	[_TT_150]	0.00110867	on
1	the	0.874644	Everest	0.103592	The	0.0104397	"	0.00086076	this
1	Everest	0.925034	average	0.0134937	A	0.0043945	Av	0.00398479	ab
1	on	0.895711	,	0.0163938	[_TT_196]	0.00788363	[_TT_195]	0.00621277	[_TT_200]
1	the	0.752617	29	0.208646	May	0.00412582	20	0.00359524	June
1	29	0.972814	twenty	0.0070889	20	0.00669178	23	0.00071936	28
1	th	0.968777	May	0.0114614	of	0.0113768	st	0.00195217	,
1	of	0.91058	May	0.0674201	,	0.00457459	may	0.00275712	[_TT_260]

WiP : « phonetic » neighbours in alternative predictions
(distance to prediction2 probability as a proxy for intelligibility?)

[_BEG_]	0.947713	0	0	[_TT_12]	0.00670132	0	0	[_TT_11]	0.00531413
Obs	0.43384	6	7	observing	0.281372	6	7	"	0.114933
erving	0.995759	31	78	er	0.00194947	31	78	erve	0.000467226
the	0.990482	78	104	The	0.00234761	78	104	a	0.00119722
steady	0.944887	104	153	study	0.0333635	104	153	Stead	0.00799184
fall	0.961571	168	191	Fall	0.00802978	168	191	fall	0.00721409
of	0.993961	191	199	the	0.00107607	191	199	in	0.000427133
the	0.969816	209	234	Bar	0.00718367	209	234	bar	0.00361852
bar	0.426446	234	260	Bar	0.33805	234	260	b	0.0446175
ometer	0.937619	260	307	omet	0.0159721	260	307	o	0.0136897
,	0.871901	323	323	Captain	0.0520907	323	323	kept	0.00540585
Captain	0.867363	363	379	captain	0.0336688	363	379	Cap	0.00512762
Mack	0.321873	392	410	Mac	0.179011	392	410	Mag	0.0842532
worth	0.510859	410	446	wer	0.105311	410	446	were	0.0928428
thought	0.727912	446	496	fought	0.212631	446	496	followed	0.00587244
,	0.526715	501	502	there	0.308247	501	502	"	0.028238
there	0.580201	553	553	"	0.135482	553	553	[_TT_250]	0.013906

https://github.com/jbyunes/whisper.cpp

Task-mediated and "listener-dependent" (Linda Tarrier and Lionel Fontan)

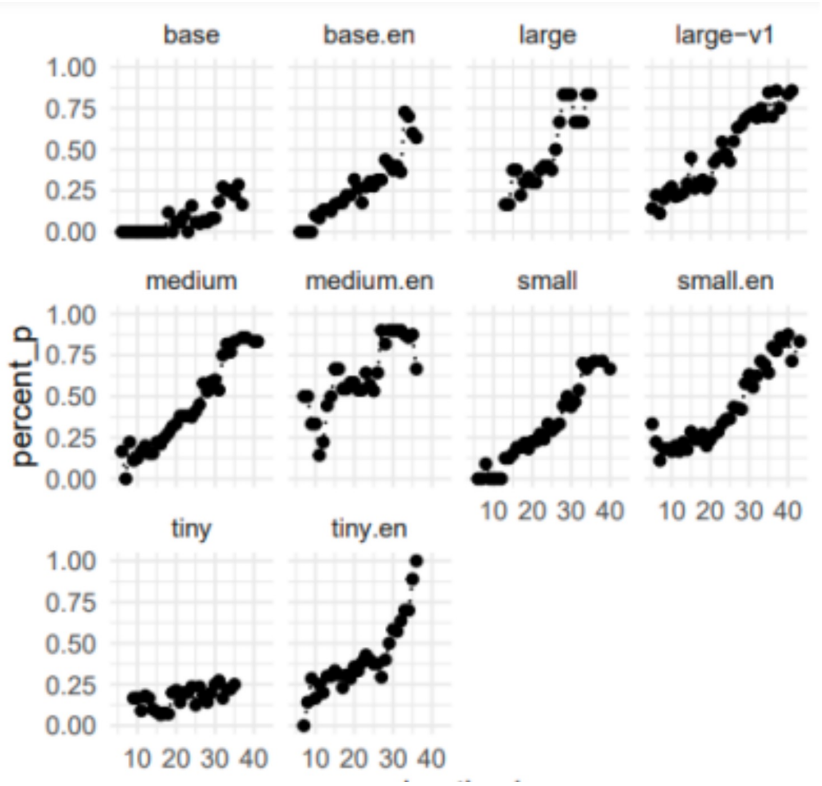
- "task-mediated" : scoring whatever task ?

-> Medium

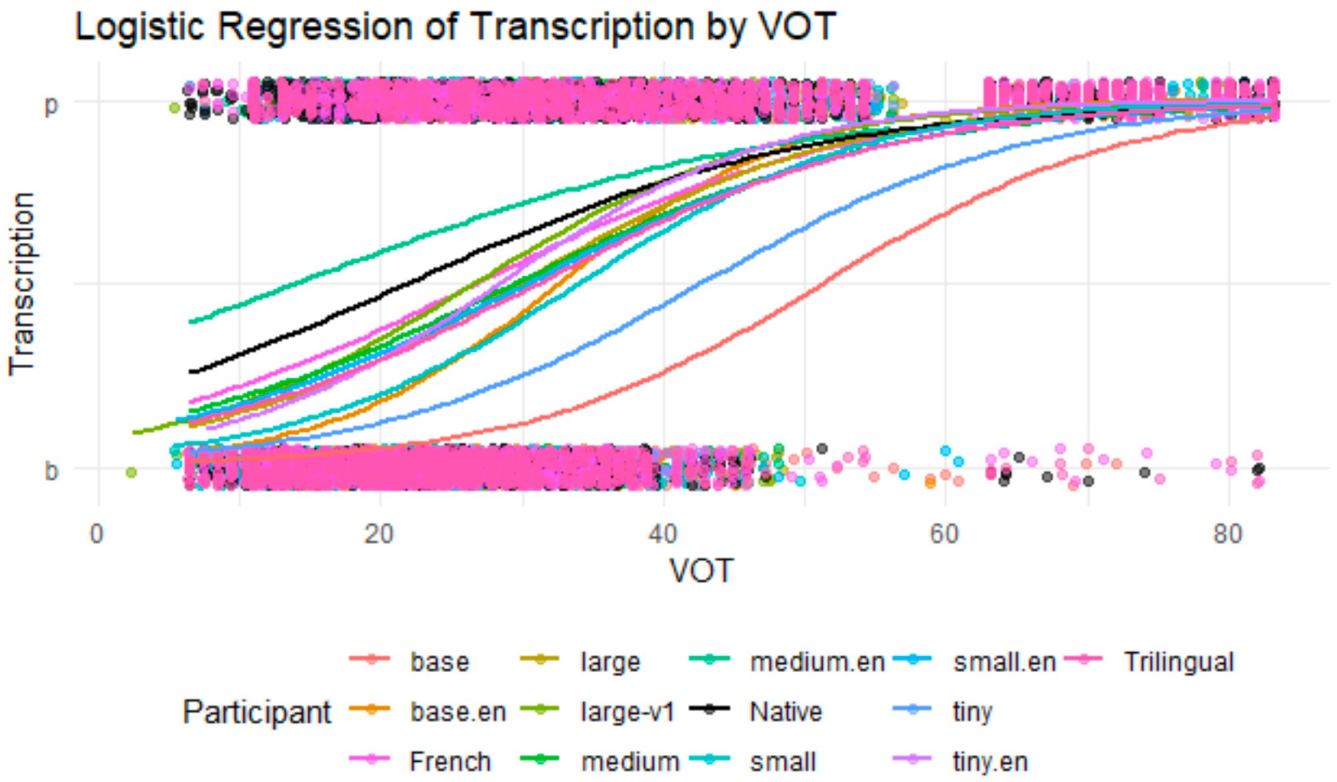
- which Whisper model should we use?

- ASR as the judge
- emulating human perception?
- From Levenshtein distance to feature distance (Stafford *et al.*, submitted)

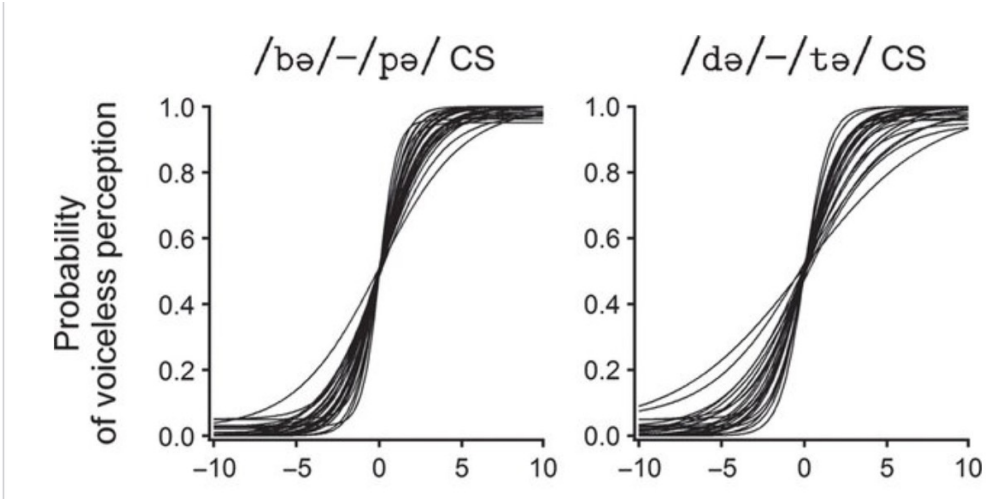
Which Whisper model to emulate comprehension?

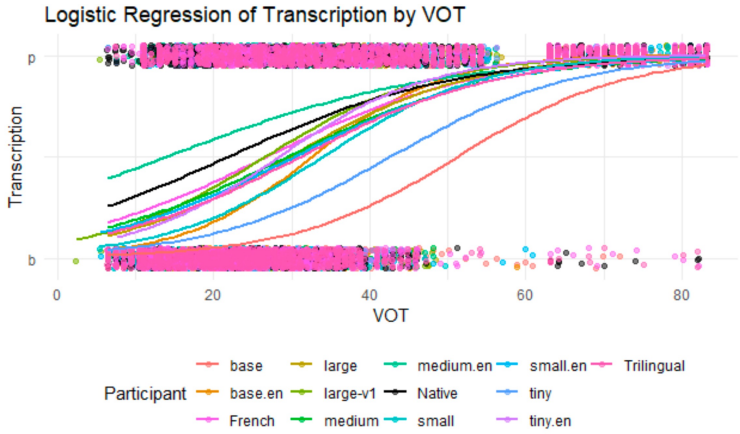


Emulating human perception? (Thurston & Ballier, to be resubmitted)



https://doi.org/10.1044/1092-4388(2011/10-0224)





Ethics guidelines for AI : (Jobin et al., 2019) (thanks to Nevya de Jong)

- Transparency: subtokenisation bias + explainability
- Justice or fairness (no…)
- Non-maleficence (non intentional, but representational harms)
- Responsibility : more research is needed, L2-scoring vs. MDD
- Privacy : local version

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature machine intelligence*, *1*(9), 389-399

Transparency

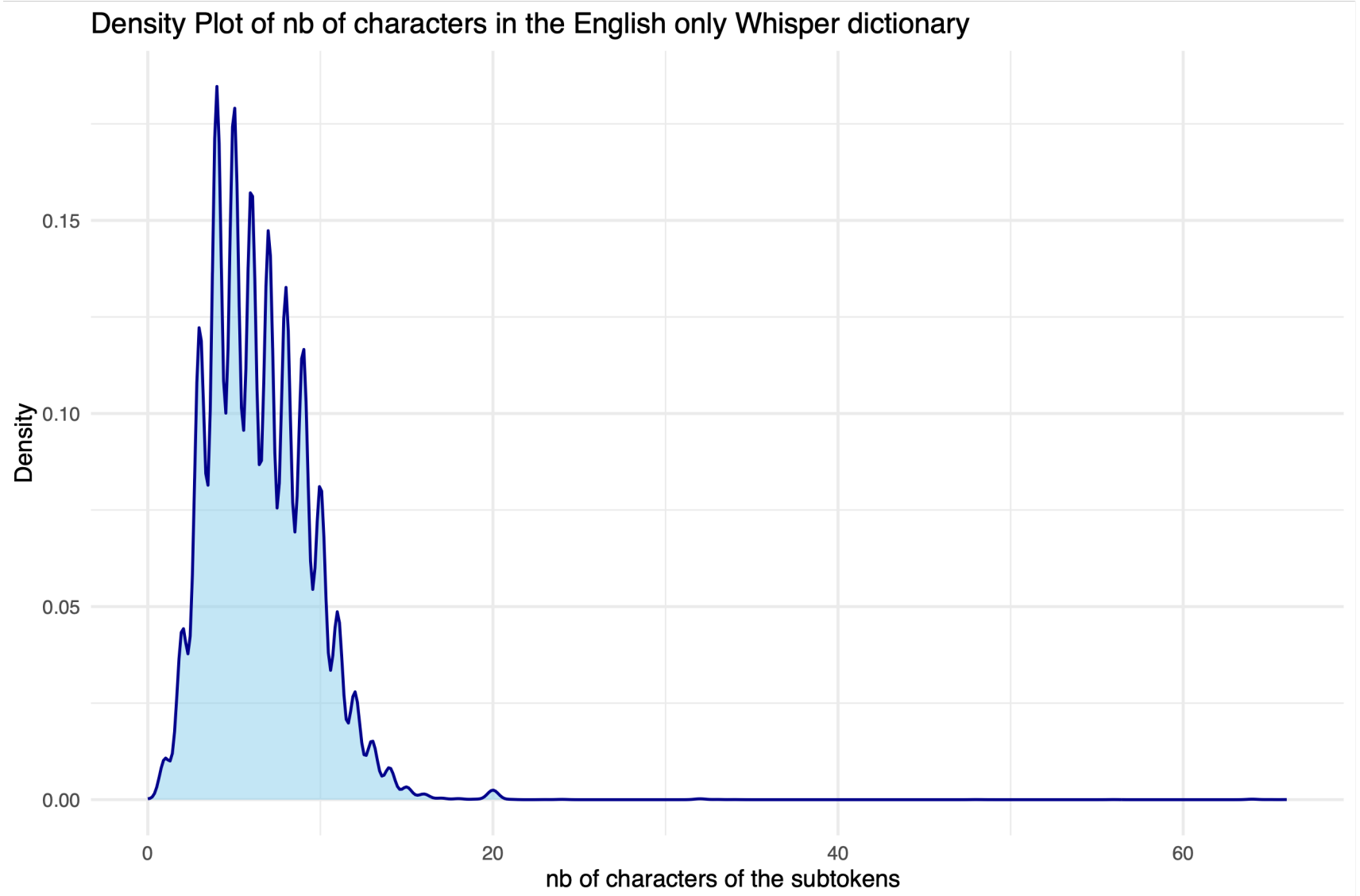
- EXPLAINABILITY : acoustic space hypothesis (Liang et al, 2026)
- (retro)convertbility of subtokens as evidence of mispronunciation
- Subtoken choices (unequal parity : prediction / length of the subtoken)
- (clearer)Division of labour between encoder and decoder
- Alignment with human judgements
- Comparable decision boundaries

Noisy dictionary for chatGPT2 (English only model)

```
   token                                                        token_length
   <chr>                                                               <int>
 1 "---------------------------------------------------------…            64
 2 "  --------------------------------------------------------…           65
 3 "  --------------------------------"                                   33
 4 ".........................................................…             64
 5 "=========================================================…            64
 6 "_________________________________________________________…            64
 7 "ÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃÂÃ…          64
 8 "  ========================================================…           66
 9 "  ********************************"                                   33
10 "--------------------------------------------------------"             56
11 "  ================================="                                  34
12 "------------------------------------------------"                     48
```

344 tokens not in UTF8

[241] "\xe7\x89"	"\xe8\x80"	"\xe5\x9b"	"\xe2\x9d"	"\xe5\xb7"
[246] "\xe8\xa3"	"\xe8\xbf"	"\xe9\x83"	"\xe5\xb9"	"\xe5\xbf"
[251] "の\xe9"	"\x96\x9a士"	"\xe7\x9b"	"\xe6\x89"	"\xe6\x97"
[256] "ﾞ\xd7"	"ﾟ\xd8"	"\xe9\x81"	"\xe9\x80"	" 裏\xe8"
[261] "\xe5\xa7"	" \xe2\x98"	" \xc4"	"\xe9\x87"	"\xb6\xe6"
[266] "\xa5\xb5"	"\xef\xb8"	"\xe2\x89"	"\xe6\x84"	"\xec\x9d"
[271] "\xe5\xbe"	"\xe6\xb0"	" \xe5\xa4"	"\xbb\x92"	"\xe5\xa6"
[276] "\xe9\x9b"	"\xe6\xb3"	"\xe5\xbd"	"\xe5\x91"	"\xe5\x86"
[281] "\xe6\xb5"	"\xe9\x96"	"\xb2\xbe"	"\xe5\x8d"	"\x91士"
[286] "\xe1\xb5"	"龍\xe5\xa5"	"\xe7\x9c"	"\xe0\xbc"	"\x82\xaa"
[291] "\xe1\xb8"	"\xb6\x85"	"\xe5\xba"	"\xf0\x9f\x91"	"\x88\xe8"
[296] " \xe6\x9c"	" \xe0\xa8"	" \xe2\x87"	" \xe2\x9d"	"\xe8\xbb"
[301] "\xe0\xa9"	"\xe5\xaf"	"\x82\x8e"	"\xe5\x82"	"\xe6\xa0"
[306] "\x81\xab"	"\xe5\x8e"	"\xe6\xa9"	"\xe7\xab"	"\xe5\x8c"
[311] "\xe6\x80"	"\xe7\x8b"	"\x81\x96"	"\xe1\xbd"	"\xac\xbc"
[316] "\xe9\xa3"	"\xe8\xaa"	"\xe7\xb7"	"\xab\x98"	"\xe7\x95"
[321] "\xe8\xaf"	"\xe2\x81"	"\xe6\x83"	"\xe4\xbf"	"\xeb\x8b"
[326] "\xe6\x95"	"\x8a\xb1"	"\xe7\x84"	"\xf0\x9f\x98"	"\xa9\xb6\xe6"
[331] "\xa9\xb6極"	"\xf0\x9d"	"\xe8\x83"	"\xe5\x8b"	"\xed\x95"
[336] "\xe0\xa6"	"\xad\xb7"	"\xe8\x88"	"\xe5\x87"	"の\xe5\xae"
[341] "\xe7\x90"	"\xe9\x9a"	"\x99\xbd"	" \xf0\x9f\x91"	


Density Plot of nb of characters in the English only Whisper dictionary
Density
0.15
0.10
0.05
0.00
0
20
40
60
nb of characters of the subtokens

Whisper multilingual subtoken dictionary (same for all models)

= what changes is the phon/acoustic <-> subtoken mapping

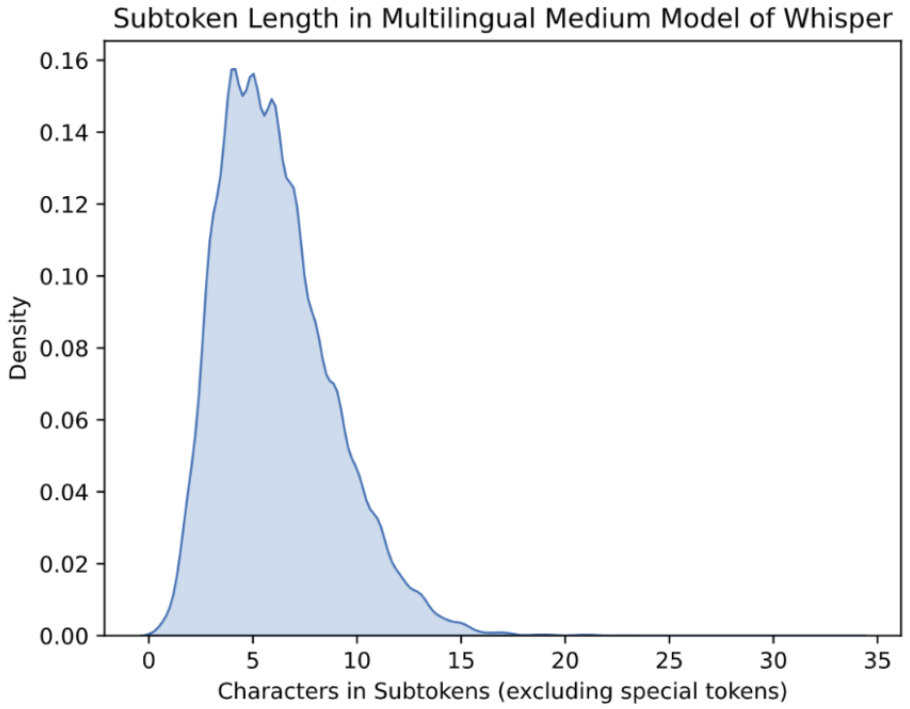


Fig. 1 Distribution of the number of characters in the Whisper subtoken dictionary

FAIRNESS: NO
(Sanabria et al. 2013)

- Compare validity

for different groups

Sanabria et al 2023

The EdAcc corpus

and Whisper WER)

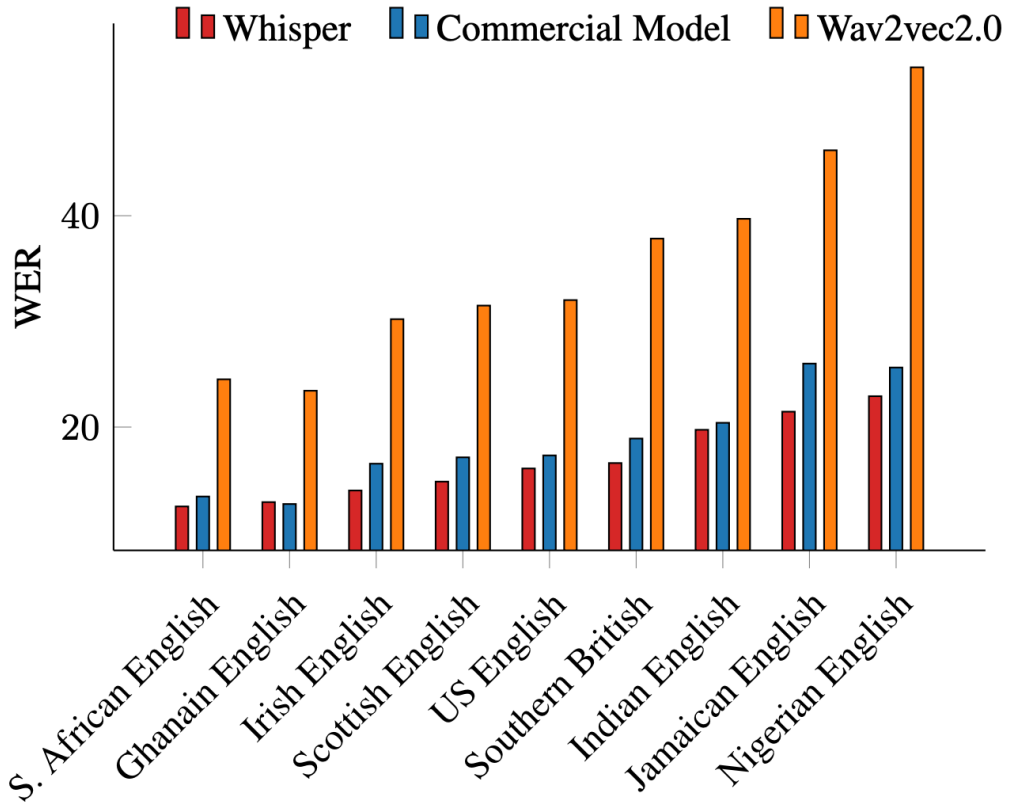


Fig. 1. WER of selected systems on conversations from the development set of `EdAcc` where both speakers has the same English variety.

A COMMON ROADMAP ?

Test suites (Janus WP 2.1)

- Holistic grading vs. Granular evaluation :
- The Speak&Improve and SpeechOcean datasets

More continua datasets for experiments to compare with human decision boundaries

- Duration ablation (30 seconds of speech) (Myssik 2011)
- More continua (pitch modification, VOT duration)

ANNOTATED datasets :

- The PARAAF corpus / dataset (UPCité) https://emmanuelferragne.com/project/paraaf/

THANKS (and CAVEATS)

THANKS AGAIN to Sylvain for organising this workshop

Radford et al : 8,500 related papers (and counting)

Many re-implementations of Whisper

The C++ customised version of Whisper (experimental modes 1-3)

https://github.com/jbyunes/whisper.cpp

Some references

- Ballier, N., Arnold, T., Méli, A., Thurston, T., & Yunès, J. B. (2024). Whisper for L2 speech scoring. *International Journal of Speech Technology*, *27*(4), 923-934.J
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature machine intelligence*, *1*(9), 389-399
- Liang, S., Ballier, N., Levow, G. A., & Wright, R. (2025). Beyond WER: Probing Whisper's Sub-token Decoder Across Diverse Language Resource Levels. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 31225-31235).
- Liang, S., Ballier, N., Levow, G. A., & Wright, R. (2026). The Limits of Data Scaling: Sub-token Utilization and Acoustic Saturation in Multilingual ASR. Paper accep53ted for LREC2026
- Sanabria, R., Bogoychev, N., Markl, N., Carmantini, A., Klejch, O., & Bell, P. (2023). The edinburgh international accents of English Corpus: Towards the democratization of English asr. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.