The Use of Phone Categories and Cross-Language Modeling for Phone Alignment of Panãra

Emily P. Ahn[ɑ], Eleanor Chodroff[ə], Myriam Lapierre[ɑ] & Gina-Anne Levow[ɑ]

[ɑ]*University of Washington, USA* [ə]*University of Zurich, CH*

1. Motivation

> **Automatic forced alignment aids field linguists, phoneticians, etc.**

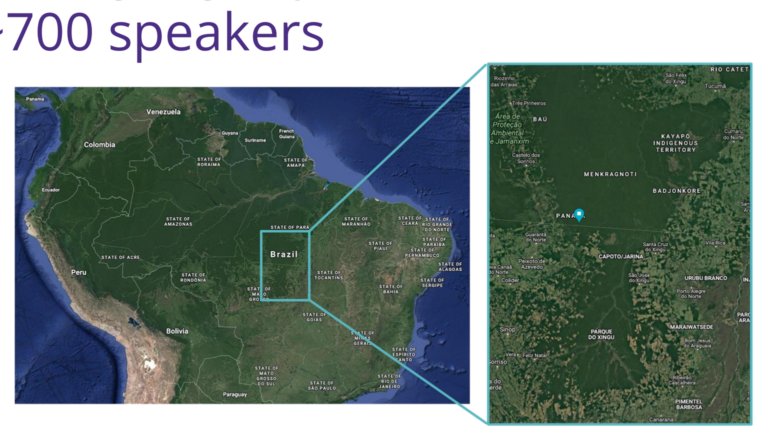
> **Goal: phone-align Panãra data in a limited data scenario via 2 strategies:**

- Cross-language modeling
  - > E.g. English/French to align Bribri[1]
- Broaden phone categories
  - > Worked for sentence alignment[4]

2. Data

Panãra (ISO: kre)

- Jê language spoken in Brazil
- ~700 speakers



Phoneme Inventory[6]

Consonants

	Bilabial	Alveolar	Palatal	Velar
Singleton obstruent	p	t	s	k
Geminate obstruent	pː	tː	sː	kː
Singleton nasal	m	n	ɲ	ŋ
Geminate nasal	mː	nː		
Approximant	w	ɾ	j	

Vowels

Short oral		
i	ɯ	u
e	ɤ	o
ɛ	a	ɔ

Short nasal		
ĩ	ɯ̃	ũ
ẽ	ɤ̃	õ

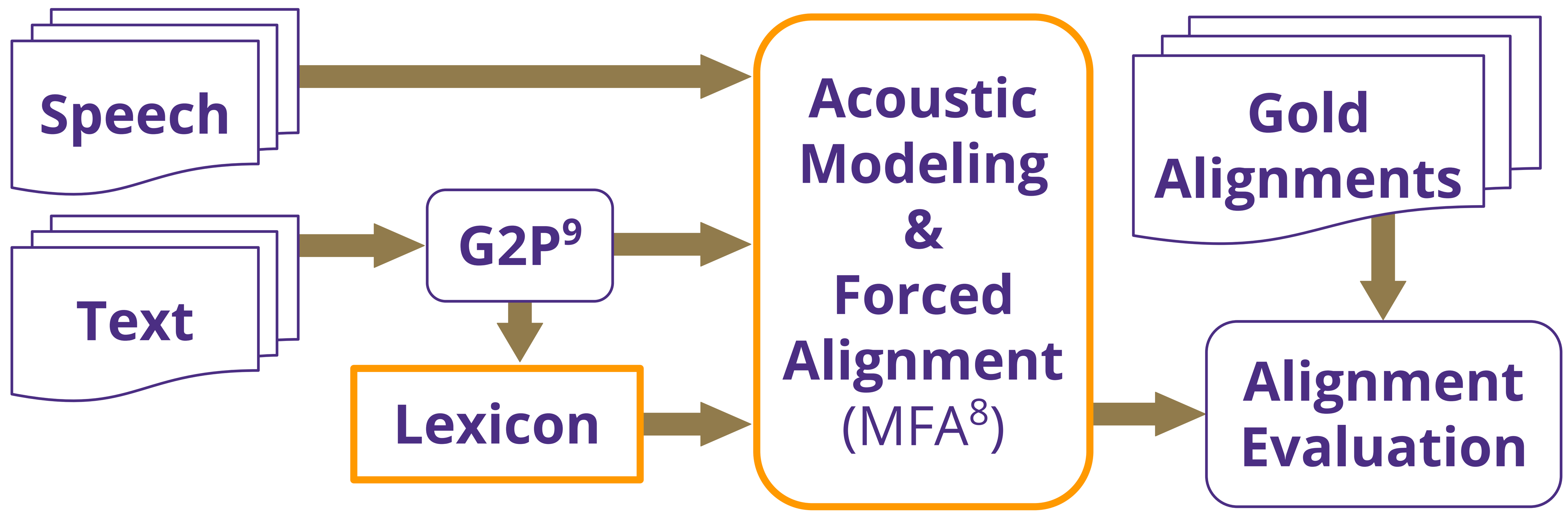
Long oral		
iː	ɯː	uː
eː	ɤː	oː
ɛː	aː	ɔː

Long nasal		
ĩː		
ẽː	ɤ̃ː	õː

Dataset

- 35 min speech, narrative style
- 4 speakers (2 male, 2 female)
- Orthographically transcribed by Myriam Lapierre, corrected with native speakers

3. Pipeline & Methods



Lexicon Manipulation

Broaden phone categories via 2 strategies:

1. No Diacritics
   Ex:

ɔ	297
ɔː	37
ɔ̃	5

→ ɔ 339

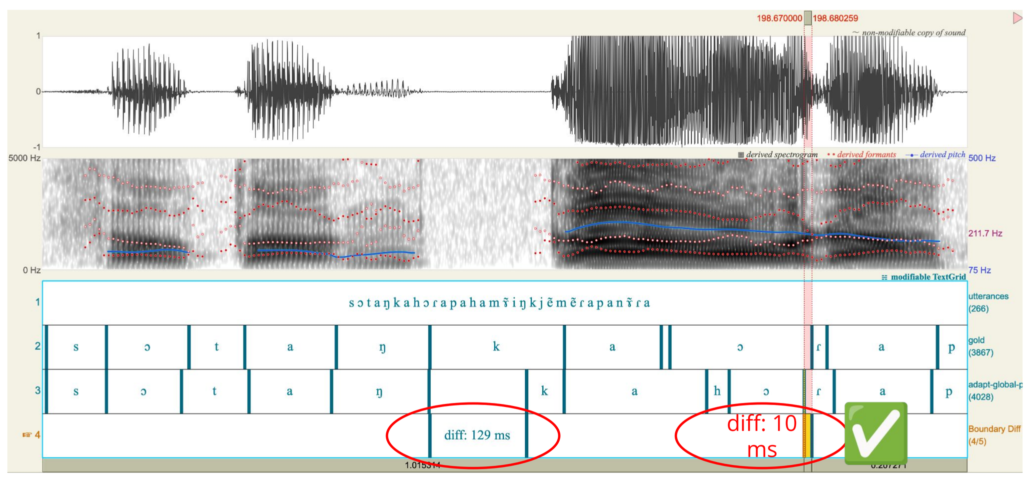
2. Natural classes from SCA[7] (Sound-Class-Based Phonetic Alignment)

No.	Cl.	Description	Examples
1	A	unrounded back vowels	a, ɑ
2	B	labial fricatives	f, β
3	C	dental / alveolar affricates	ts, dz, tʃ, dʒ
4	D	dental fricatives	θ , ð
5	E	unrounded mid vowels	e, ɛ
6	G	velar and uvual fricatives	ɣ , x
7	H	laryngeals	h, ʔ
8	I	unrounded close vowels	i, ɪ
9	J	palatal approxoimant	j
10	K	velare and uvular plosives	k, g
11	L	lateral approximants	l
12	M	labial nasal	m
13	N	nasals	n, ŋ
14	O	rounded back vowels	œ, ɒ

No.	Cl.	Description	Examples
15	P	labial plosives	p, b
16	R	trills, taps, flaps	r
17	S	sibilant fricatives	s, z, ʃ, ʒ
18	T	dental / alveolar plosives	t, d
19	U	rounded mid vowels	ɔ , o
20	W	labial approx. / fricative	v, w
21	Y	rounded front vowels	u, ʊ, y
22	0	low even tones	11, 22
23	1	rising tones	13, 35
24	2	falling tones	51, 53
25	3	mid even tones	33
26	4	high even tones	44, 55
27	5	short tones	1, 2
28	6	complex tones	214

Evaluation

Phone Onset Boundary Accuracy[8]: % of system onsets within 20 ms of manually annotated gold onsets



Original Text	Panãra Orthography	Haa mämä jynkjân rasu hapôô
After Lexicon Manipulation	Panãra Explicit	h aː m ɤ̃ m ɤ̃ j ɯ ŋ k j ɤ n ɾ a s u h a p oː
	Panãra No Diacritics	h a m ɤ m ɤ j ɯ ŋ k j ɤ n ɾ a s u h a p o
	TIMIT English Explicit	h a m ə m ə j ɪ ŋ k j ə n ɾ a s u h a p oʊ
	Broad (SCA)[7]	H A M E M E J I N K J E N R A S Y H A P U
	MFA Global English Explicit	h aː m o m o j u ŋ k j o n ɾ a s u h a p oː

Experiments

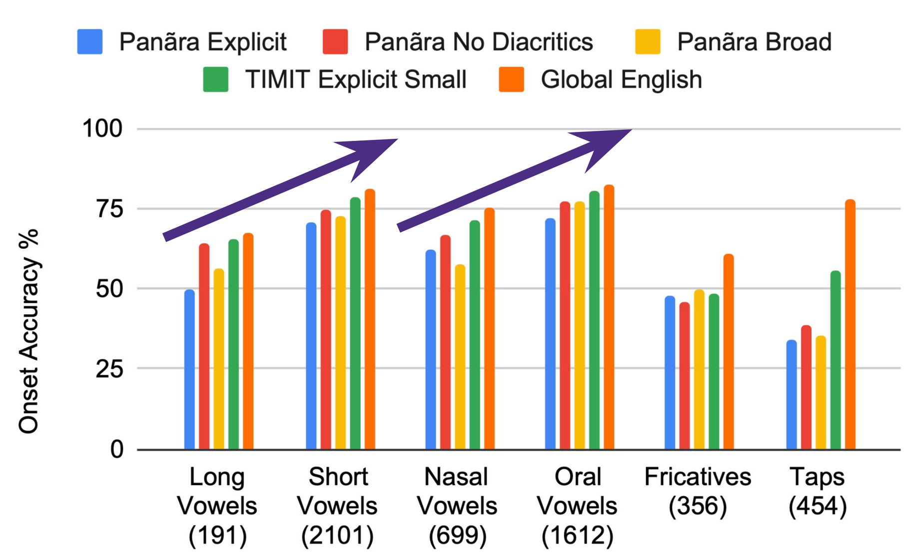
1. ***Language-specific + broaden phones***
   Panãra-only trained models
   a. Explicit: all Panãra phones
   b. No Diacritics: 🚫 length/nasal markers
   c. Broad natural class[7]
2. ***Cross-language + broaden phones***
   English model trained on TIMIT[2] (4 hours, 519 speakers from US)
   –"Full": 224 min, 495 speakers
   –"Small": 26 min, 51 speakers
   a. Explicit: all Panãra phones mapped to TIMIT English phones
   b. Broad natural class[7]
3. ***Large, pretrained English model***
   English MFA[8] acoustic model 2.2.1 pretrained on 3770 hours speech from US, UK, Nigeria, India
   a. Explicit: all Panãra phones mapped to Global English phones

4. Results

1. Broadening phone categories **improved** alignment accuracy in **language-specific (Panãra only)** training
2. Broadening phone categories **did not improve** alignment accuracy in **cross-language (English-Panãra)** training
3. A **large, pretrained English model outperformed previous strategies**.

Trained Dataset	*Trained Settings (# phone categories)*	*Onset Accuracy within 20ms (%)* ⬆️
Panãra	Explicit (63)	60.20
	No Diacritics (29)	**62.35**
	Broad (17)	61.92
English (TIMIT)	Explicit Full (46)	62.65
	Explicit Small (46)	**66.09**
	Broad Full (19)	56.07
	Broad Small (19)	61.14
English (Global MFA)	**Explicit (100)**	**69.82**

5. Analysis



Phone natural class affected onset accuracy

1. Long & Nasal vowel boundaries more difficult than Short & Oral
   — Long & Nasal vowels typologically less common[3]
2. Poor [h] alignment in Fricatives
   — variable [h] insertion in onsetless syllables[5]
3. Poor [ɾ] alignment in Taps
   — vowel insertion within a complex onset[5], e.g. /kɾɤ/ → [kVɾɤ] 'thigh'

6. Conclusion

Summary

1. We tested phonetic granularity effects in acoustic modeling and alignment of Panãra
2. For best alignment performance: use a large acoustic model for cross-language alignment

Future Work

> Apply techniques to other language varieties

> Automatic phone category/granularity discovery

> Multilingual, language-agnostic alignment

References

1. Coto-Solano, R., & Solórzano, S. F. (2017). Comparison of Two Forced Alignment Systems for Aligning Bribri Speech. In *CLEI ELECTRONIC JOURNAL*.
2. Garofolo, J. S. (1993). TIMIT Acoustic Phonetic Continuous Speech Corpus. *Linguistic Data Consortium*, 1993.
3. Gordon, M. K. (2016). *Phonological Typology*. Oxford University Press.
4. Hoffmann, S., & Pfister, B. (2013). Text-to-speech Alignment of Long Recordings Using Universal Phone Models. In *Interspeech*.
5. Lapierre, M. (2023a). The Phonology of Panãra: A Prosodic Analysis. In *International Journal of American Linguistics*.
6. Lapierre, M. (2023b). The Phonology of Panãra: A Segmental Analysis. In *International Journal of American Linguistics*.
7. List, JM. (2012). SCA: Phonetic Alignment Based on Sound Classes. In: Lassiter, D., Slavkovik, M. (eds) *New Directions in Logic, Language and Computation*.
8. McAuliffe et al. (2017). Montreal Forced Aligner: Trainable Text-speech Alignment Using Kaldi. In *Interspeech*.
9. Mortensen et al. (2018). Epitran: Precision G2P for Many Languages. In *LREC*.

Acknowledgments

We thank Hossep Dolatian for idea inspiration, and Bridget Tyree for help with annotations.