Extracting accurate acoustic phonetic data from inaccurate alignment

Christian DiCanio[1], Hosung Nam[1], D. H. Whalen[1,2], Jonathan Amith[3], and Rey Castillo García[4]

1. Haskins Laboratories, 2. City University of New York 3. Gettysburg College

4. Secretaria de Educación Pública (SEP), Guerrero, Mexico.

(((Haskins Laboratories)))

Funded by NSF

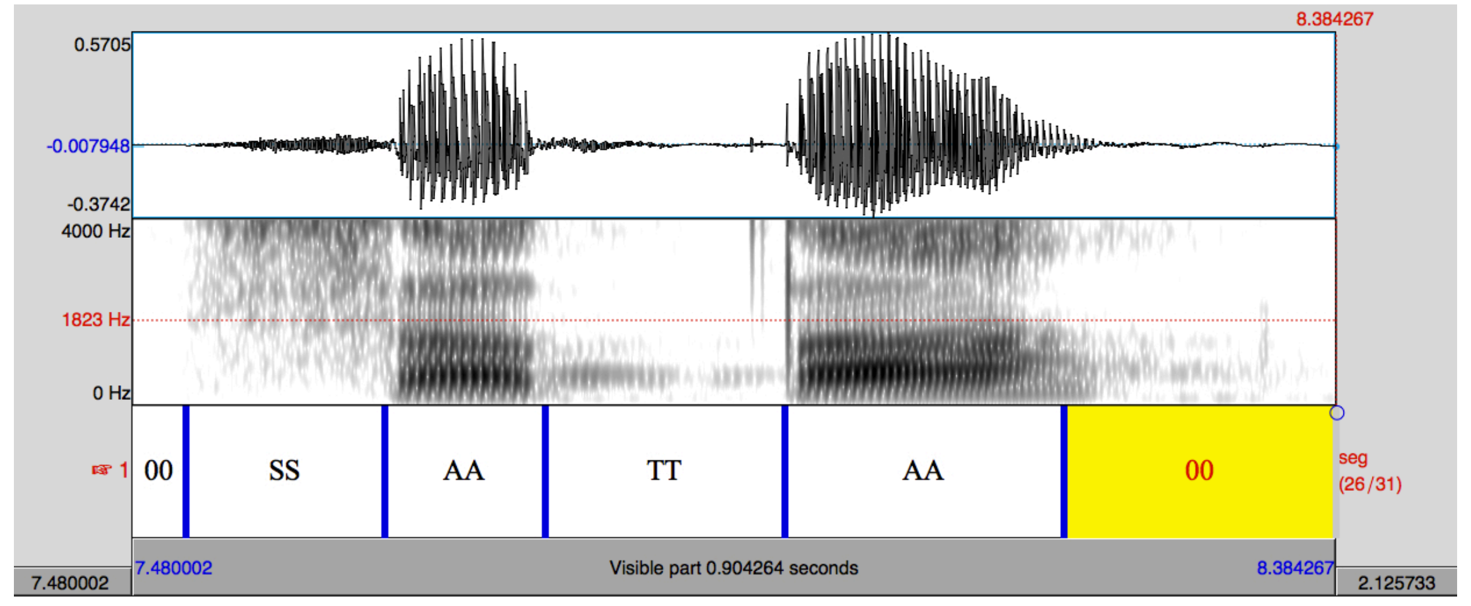
I. Phonetics & endangered languages

1. Endangered language documentation projects involve 30⁺ hours of speech.
2. Vitality of corpora is dependent on accessibility to researchers and the community.
3. Extracting phonetic data from these corpora is arduous. Is there an easier way to segment the data?
4. How much data is enough?

II. Forced alignment (**FA**)

1. Automatically segments speech.
2. Usually language-specific and trained on large corpus, it requires a lexicon of words, a transcription of the speech signal, and the speech signal itself.
3. The training data is used to build HMMs for the acoustic signature of each phone. The FA system then uses its internal model to predict where boundaries between phones occur.

Input: sound file, transcription /sata/, and "lexicon" containing coding of transcription, e.g. /sata/ = SSAATTAA.



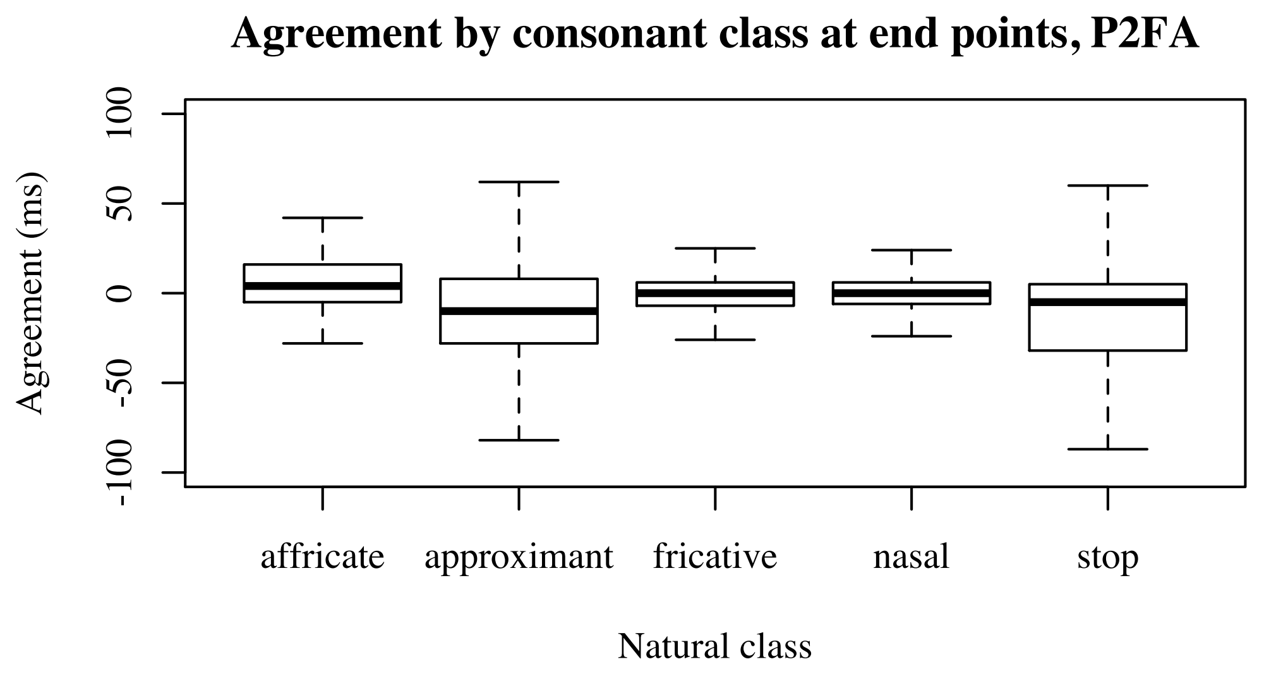
With alignment done, Praat scripts can be used to extract acoustic data.

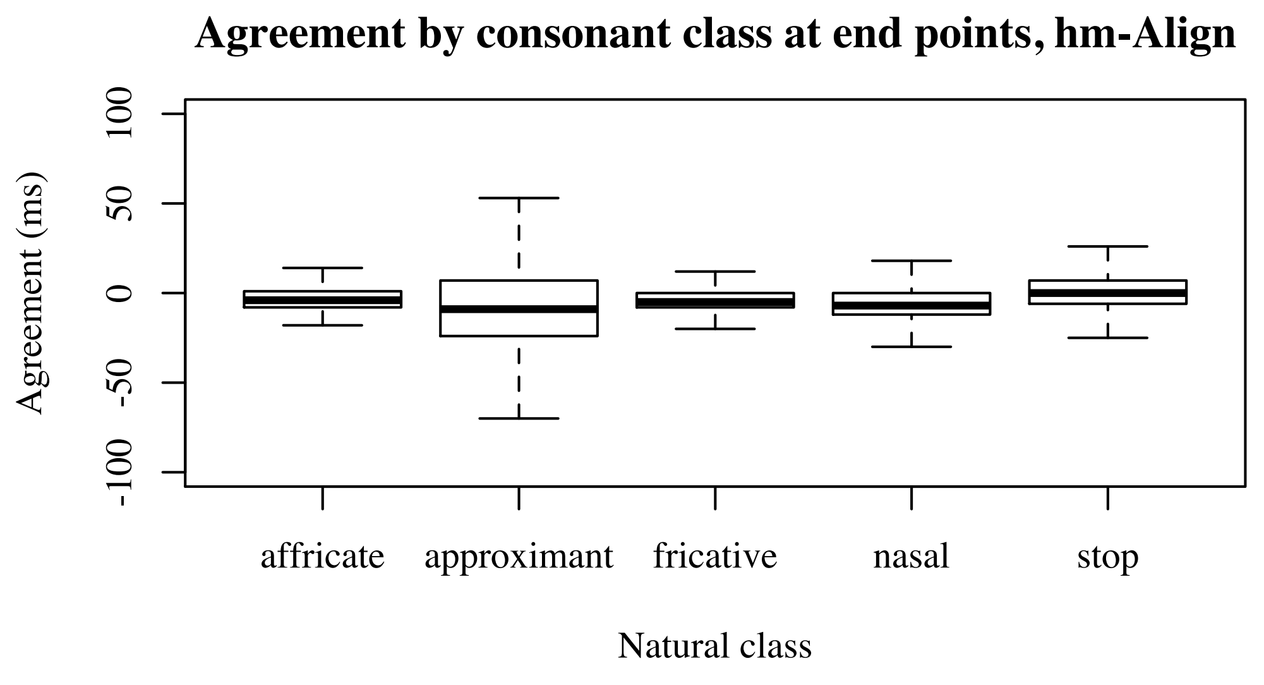
VI. Results: how accurate is FA with Yoloxóchitl Mixtec data?

Better agreement between hand-labelling and hm-Align than between hand-labelling and P2FA.

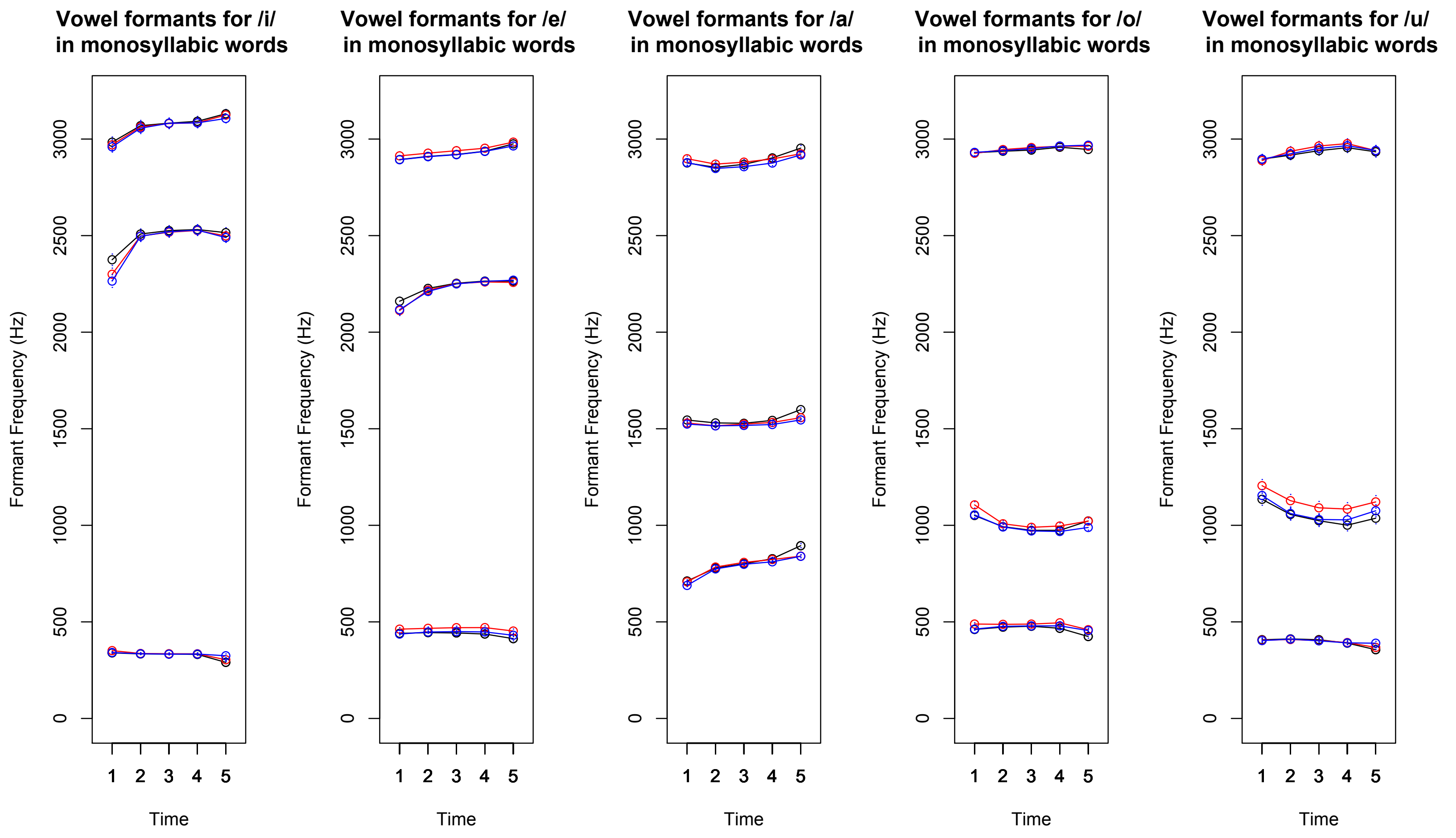
Threshold	P2FA	hm-Align
10 ms	32.3%	40.6%
20 ms	52.3%	61.4%
30 ms	65.7%	70.9%
40 ms	74.8%	81.2%
50 ms	79.6%	86.7%

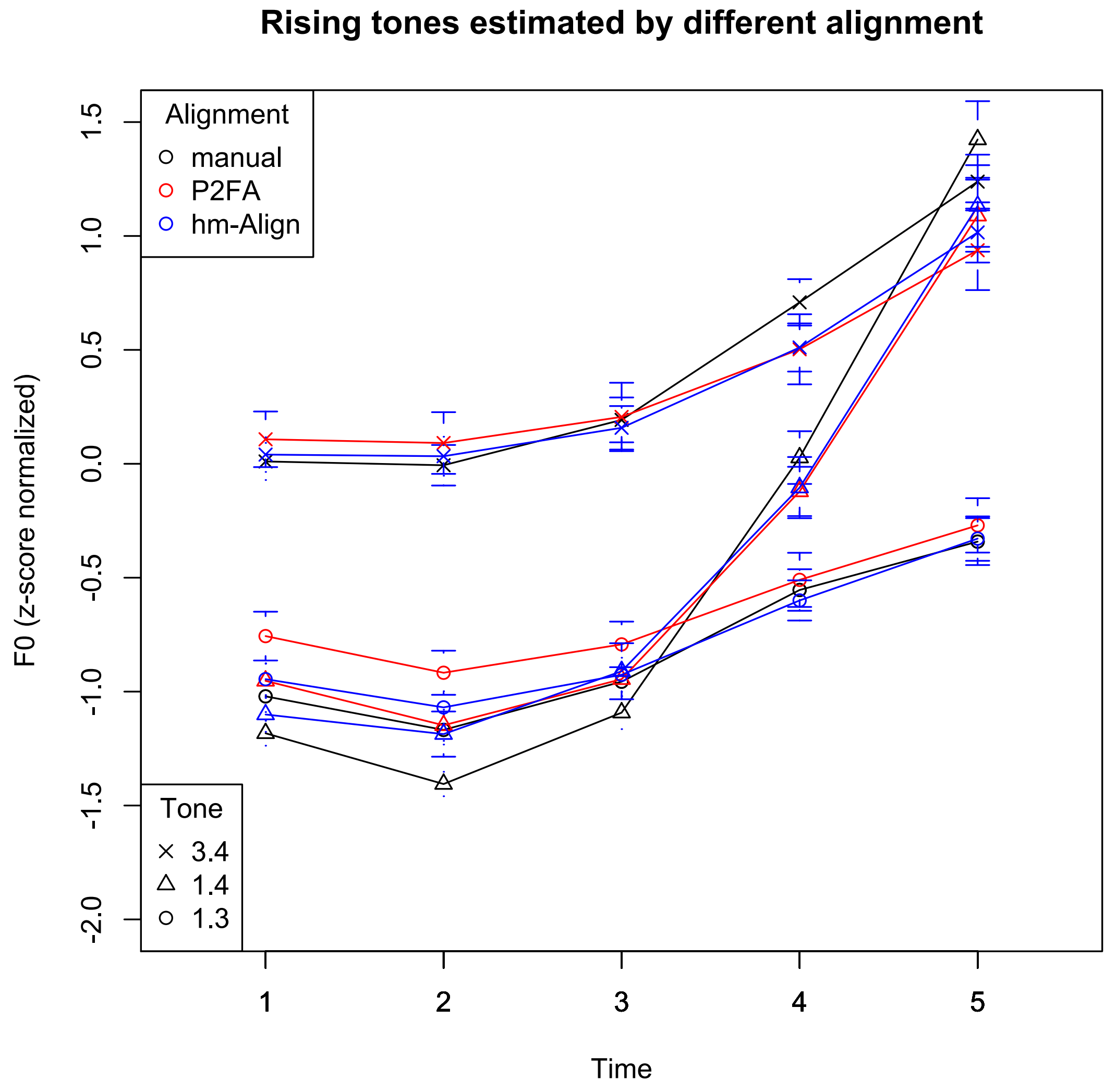
Agreement is typically 70-90% at 20 ms or better for other FA models (Malfrère et al. 2003). Does this matter?



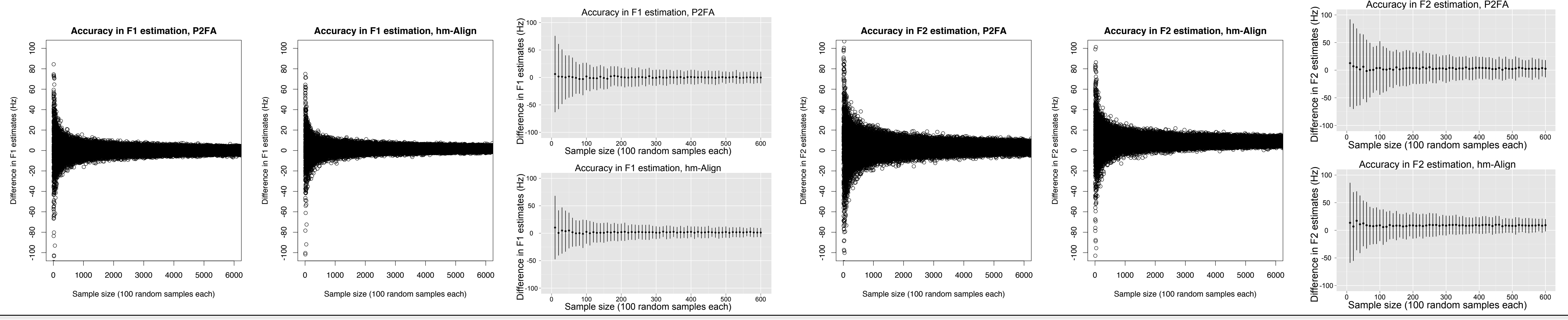


VII. Results: how accurate are the extracted acoustic measures?





VIII. Results: how much data is needed to approximate the mean?



III. Yoloxóchitl Mixtec

1. Endangered Mixtec variant spoken in Guerrero, Mexico.
2. Large-scale documentation project with 100⁺ hours of transcribed texts; phonetic/phonological studies (Castillo-García 2007, DiCanio et al. 2012).
3. Corpus = 261 words in isolation x 6 reps x 10 speakers = 15,660 tokens (monosyllables & disyllables).

IV. Phonological system

1. Small consonant inventory but large tonal inventory. 4 levels or 5 contours possible on a single mora.

	Front	Central	Back
Close	i, ĩ		u, ũ
Close-mid	e, ẽ		o, õ
Open		a, ã	

	Bilabial	Dental	Post-alveolar	Palatal	Velar	Labialized Velar	Glottal
Plosive	(p)	t			k	kʷ	ʔ
Pre-nasalized plosive	(mb)	nd					
Affricate			tʃ				
Nasal	m	n					
Tap		(ɾ)					
Fricative	β	s	ʃ				
Approximant		l		j			

2. Open, simple syllables. Words are maximally trimoraic (CVCVCV, CVCVV) and minimally bimoraic (CVCV, CVV).

V. Forced alignment

1. Compared accuracy of FA to hand-labelled data using P2FA and hm-Align
2. P2FA (Yuan & Liberman 2008, 2009) uses GMM-based monophone-HMMs trained using the SCOTUS corpus; phonemic.
3. hm-Align (Bunnell et al. 2005) uses a set of discrete monophone HMMs trained on data from the TIMIT corpus (Garofolo et al. 1993); allophonic.