Evaluating the Efficacy of Prosody-lab Aligner for a Study of Vowel Variation in Cantonese

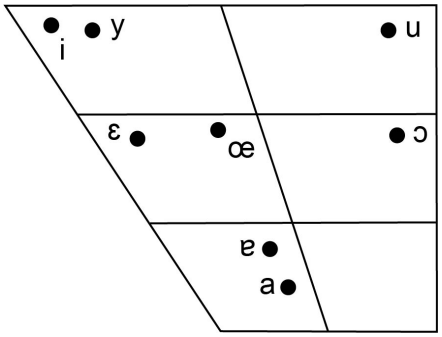
Andrew Peters (彭浩軒) & Holman Tse (謝浩明)
abpeters@yorku.ca hbt3@pitt.edu
YORK UNIVERSITÉ UNIVERSITY
University of Pittsburgh

Workshop on Innovations in Cantonese Linguistics (WICL-3)

THE OHIO STATE UNIVERSITY

Columbus, OH

March 12, 2016



HTTP://PROJECTS.CHASS.UTORONTO.CA/NGN/HLVC

UNIVERSITY OF TORONTO

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Presentation Goals

- To demonstrate the use of Prosodylab Aligner as a tool for a large scale project examining vowel variation and change in Toronto Heritage Cantonese
- To address the effectiveness of Prosody-lab aligner for this purpose
- To assess the best source for training new Models
  - Data from all speakers together (ALL)?
  - Data from each generational group separately (GEN)?
  - Data from each speaker individually (SOLO)?

What is the HLVC Project?

- Large-scale project investigating variation and change in Toronto's heritage languages.
- Includes sociolinguistic interview data from 7+ heritage languages spoken by immigrants and 2 or 3 generations of their descendants
- The corpus makes it possible to investigate contact effects on a wide variety of variables across all languages using the same methodology

Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada
UNIVERSITY OF TORONTO

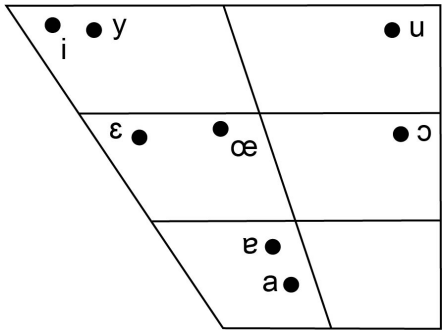
A Sample of Linguistic Variables

	Cantonese	Faetar	Italian	Korean	Polish	Russian	Ukrainian
VOT	✓		✓	✓		✓	✓
Ø-subject	✓	✓	✓		✓	✓	
Borrowing	✓	✓					
Classifiers	WICL-1/3			WICL-3			
Vowels	WICL-3						

GEN 1	GEN 2
Born and raised in HK, Immigrated to Canada as adults	Grew up in Toronto
L1 Cantonese, Some L2 English	Simultaneous (Early) Bilingual in Cantonese and Toronto English

Methodological Issues

- Hour-long interviews (spontaneous speech) from each of ~ 40 speakers
  - 40 speakers X 8 vowels X 6 tones X 10+ tokens/each = 19200!!!



- Forced Alignment Tools
  - FAVE (Rosenfelder et al 2011)
    - Now widely used for sociolinguistic studies of English dialects
    - But only works on English
  - Prosodylab-Aligner (Gorman et al 2011)
    - Can train new models from raw data making it customizable for any language
    - However, its efficacy for Cantonese unknown

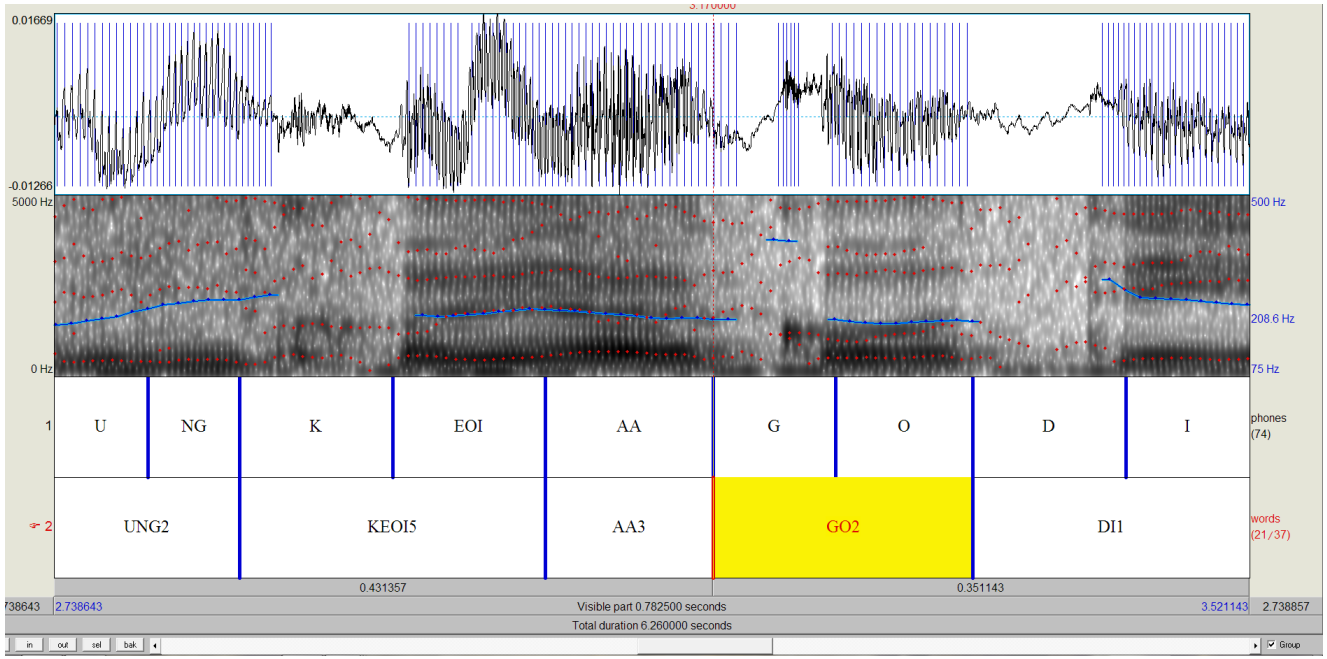
Forced Alignment & Vowel Extraction (FAVE)
An online suite for automatic vowel analysis

More About Prosodylab

- ProsodyLab (Gorman et al. 2011) is based on the Hidden Markov Toolkit (HTK), a speech recognition toolkit based on Hidden Markov Models, developed at Cambridge University
- Requires
  - Python 2.6 or above
  - SoX (Sound Exchange)
  - HTK (Hidden Markov Model Toolkit)
- Can be downloaded from
  - https://github.com/kylebgorman/Prosodylab-Aligner
  - More info
    - http://prosodylab.org/tools/aligner/

What is Forced Alignment?

- Forced alignment automates the process of time-aligning transcription with audio signal
- Permits automated measure of variable, e.g. formant values



About Acoustic Models

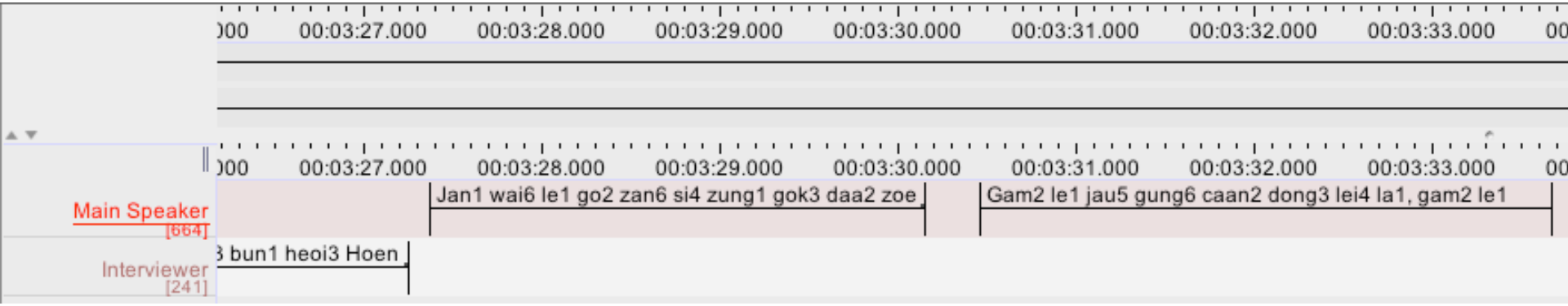
- Uses machine-learning to perform transcript to audio time-alignment
- Speech models map phone lists to audio signal
- Will vary in how well they fit the data, how well they demarcate boundaries etc. Hence our study!

Questions

- Is Prosody-lab aligner effective at producing sufficiently accurate transcript alignment to permit automated measurement of vowel data?
- What is the best data source for training models?
  - All speakers together (ALL)?
    - More robust model, but does it work as well with the variation present in a HL variety
  - Each generational group separately (GEN)?
    - Tse (2015) suggest inter-generational phonological differences
  - Each speaker individually (SOLO)?
    - Requires a large percentage of data, but would it be as accurate?

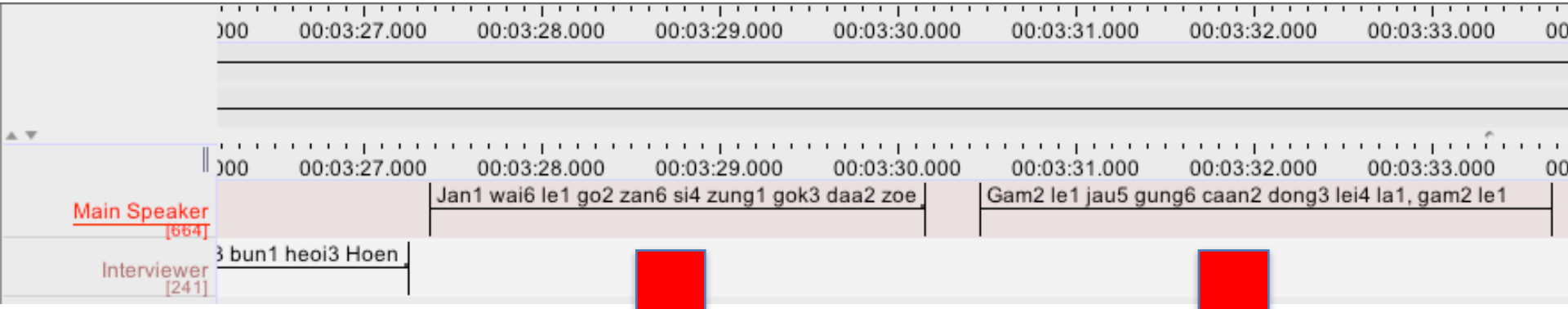
Pre-processing

1. Interviews transcribed by native speakers of Cantonese using Jyutping Romanization in ELAN
   - Manual sentence-level alignment



2. To create input readable by Prosodylab-Aligner, PRAAT script used to create smaller .wav files with matching .txt files for each annotation.

PRAAT Script (Labber)



C1F54A_IV_2074.wav

C1F54A_IV_2074.txt

Jan1 wai6 le1 go2 zan6 si4 zung1 gok3 daa2 zoeng3.

Translation: "Because at that time, China was at war."

C1F54A_IV_2105.txt

Gam2 le1 jau5 gung6 caan2 dong3 lei4 la1, gam2 le1

C1F54A_IV_2105.wav

Translation: "And then the Communist Party came, and then ..."

Forced alignment needs a **custom** dictionary

Orthography	Phonemes
GU1	G U
GU2	G U
GU3	G U
GU4	G U
GU5	G U
GU6	G U
TUB	T AH1 B
TUBA	T UW1 B AH0
TUBAL	T UW1 B AH0 L
TUBB	T AH1 B
TUBBS	T AH1 B Z
TUBBY	T AH1 B IY0
TUBE	T UW1 B
TUBE	T Y UW1 B

To train an acoustic model:

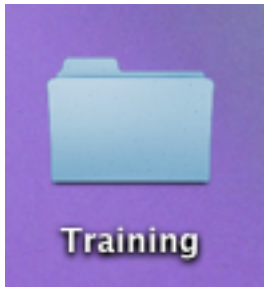
- **pronouncing bilingual dictionary (~ currently 3.6 MB)**
- **important b/c program can't run when there are unrecognized words in the transcript**
- **program needs to convert orthography to phonemic segment as established by custom dictionary**

Training and Evaluation

CAN_dict_4.txt

```
ZUNG3  Z U NG
ZUNG4  Z U NG
ZUNG5  Z U NG
ZUNG6  Z U NG
ZUT1  Z U T
ZUT2  Z U T
ZUT3  Z U T
ZUT4  Z U T
ZUT5  Z U T
ZUT6  Z U T
ZYU1  Z YU
ZYU2  Z YU
ZYU3  Z YU
ZYU4  Z YU
ZYU5  Z YU
ZYU6  Z YU
ZYUN1  Z YU N
ZYUN2  Z YU N
ZYUN3  Z YU N
ZYUN4  Z YU N
ZYUN5  Z YU N
ZYUN6  Z YU N
ZYUT1  Z YU T
ZYUT2  Z YU T
ZYUT3  Z YU T
ZYUT4  Z YU T
ZYUT5  Z YU T
ZYUT6  Z YU T
```

Custom dictionary in the format of The CMU Pronouncing Dictionary



- .wav files and matching .lab files put in a Training directory
- Prosodylab-aligner uses Training directory and **dictionary** to build an acoustic model

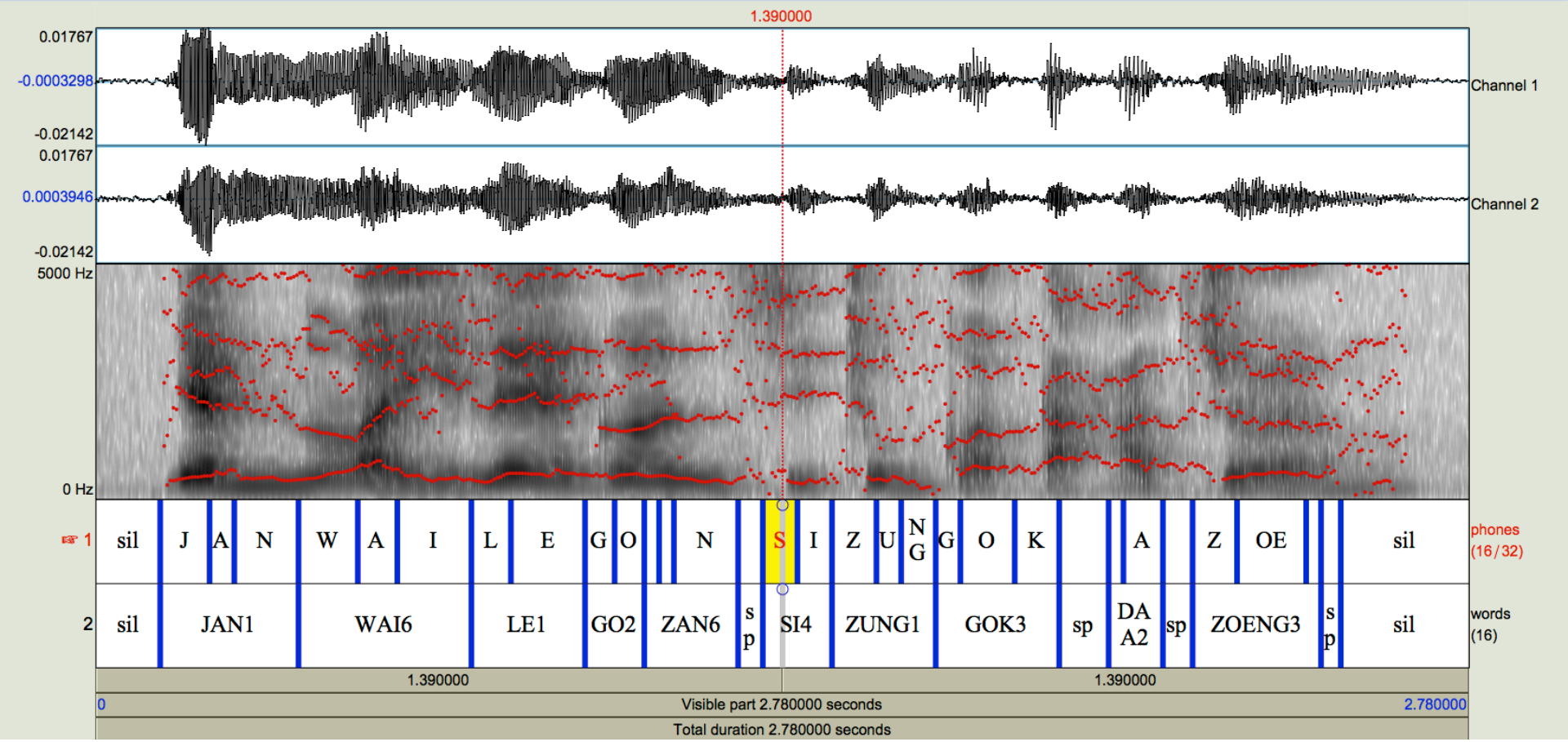
Our 3 Models of Training With 50% of data from each speaker:

1. Solo-trained model: trained only on data for speaker evaluated
2. Generation-trained model: Data from all speakers of each Gen. Combined in Training directory
3. "All"-trained model: Data from all speakers combined in Training directory

More Training Data (Hours of speech) → Better Model

Therefore: More speakers data used in training = Less data lost from each speaker to training

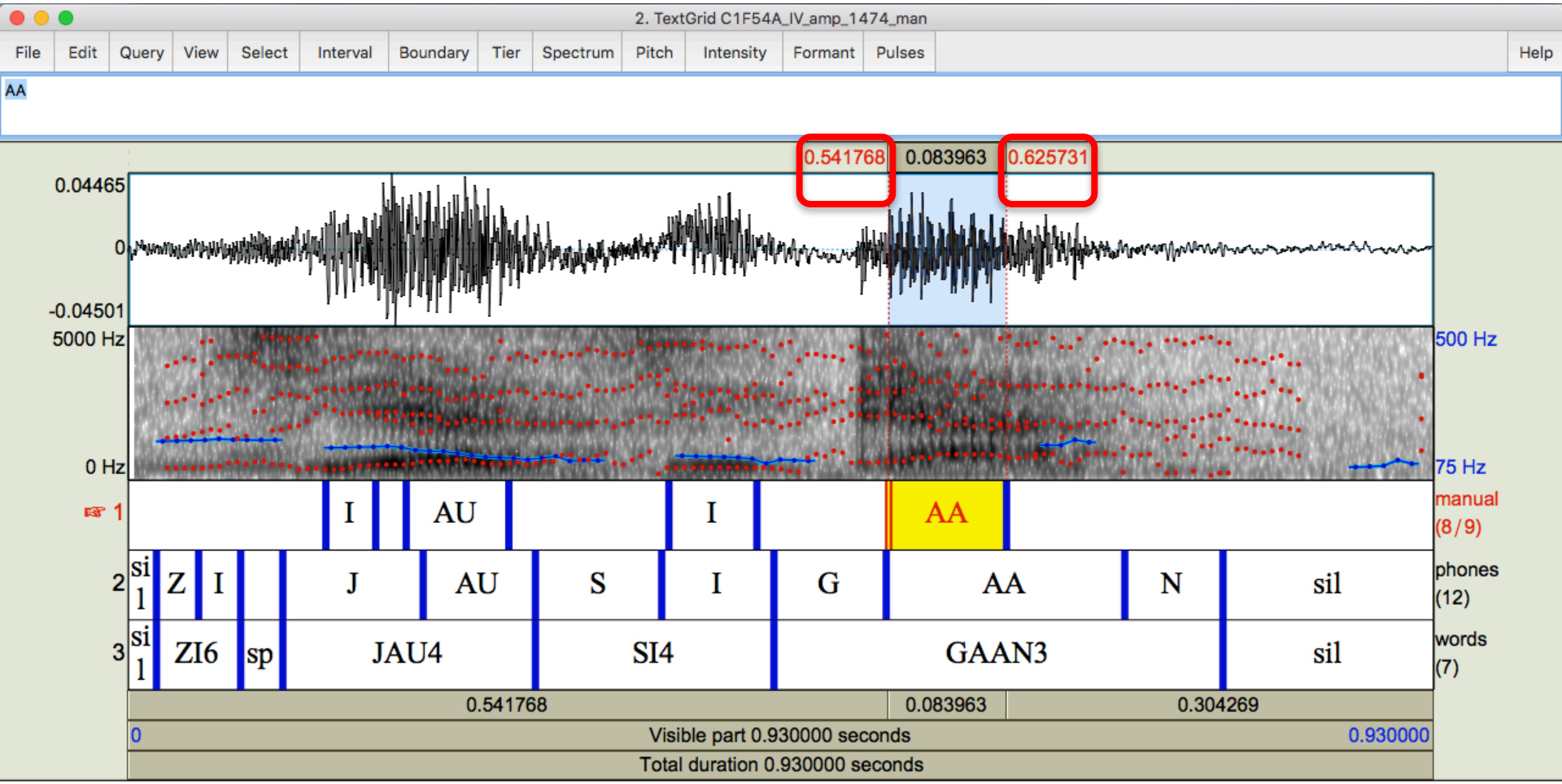
Output of Prosodylab-Aligner: Time-aligned Textgrid



Assessing Accuracy

- Assessment based on 10 speakers (four GEN 1 and six GEN 2)
- Examined first 10 usable textgrids for each speaker

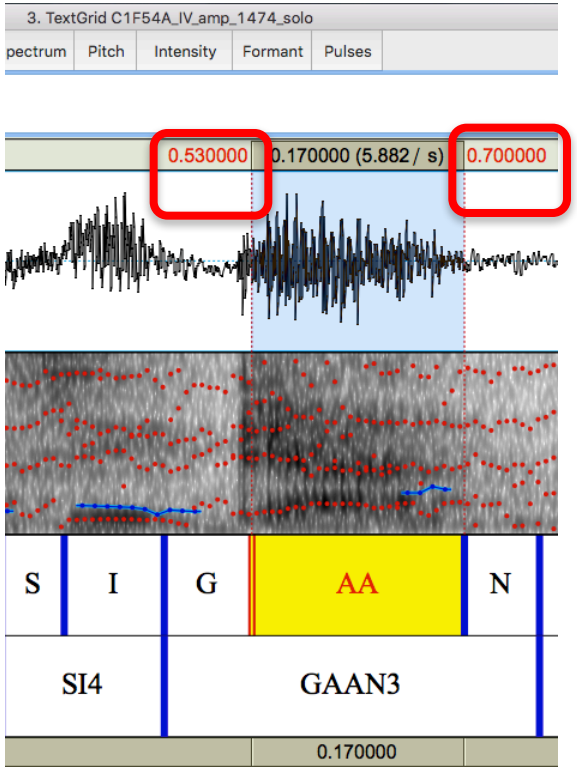
Gold Standard: Manually identify vowel boundaries for all CAN monophthongs



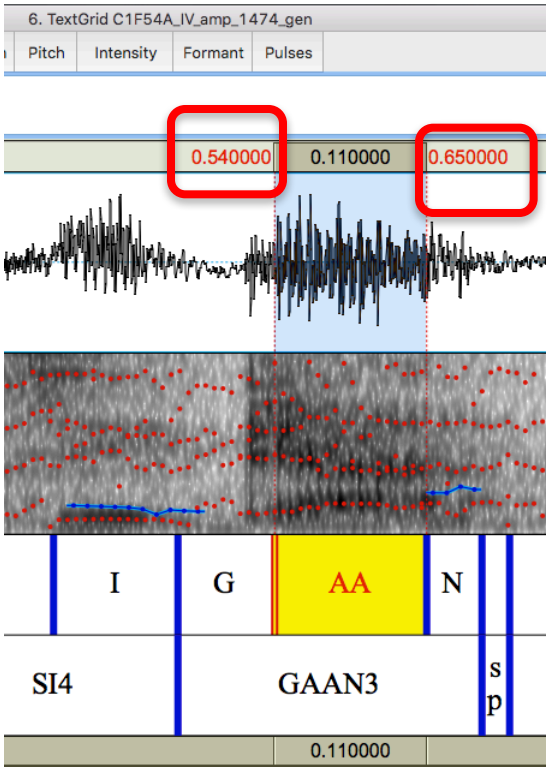
Assessing Accuracy Procedures

- Record "Gold Standard" vowel boundaries
- Record Auto-aligned vowel boundaries

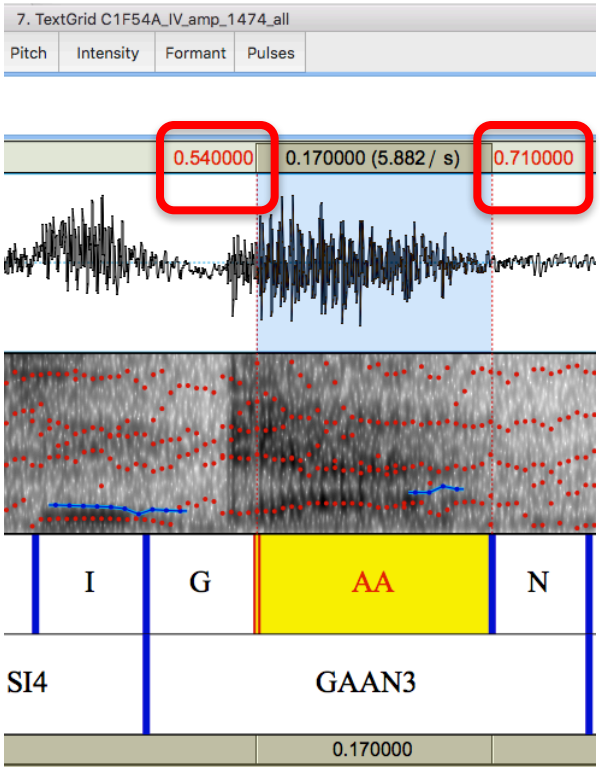
Segment Boundaries: ***Solo-aligned***



Segment Boundaries: ***Gen-aligned***



Segment Boundaries: ***All-aligned***



Assessing Accuracy

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Speaker	Timestam	Vowel	trainin	Man. Left	Man. right	Auto left	Auto right	left diff.	right diff.	ABS Left Diff	ABS right diff	Left diff ^2	Right Diff ^2	V IN TARGET?	V. Length - Manual	V. Length - Auto	V. Length Diff.
92	C1F78A	15854	F(U)1	solo	3.53	3.65	3.54	3.67	-0.0100	-0.0200	0.0100	0.0200	0.0001	0.0004	1	0.120000	0.130000	0.010000
93	C1F78A	15897	L(OE)NG5	solo	0.35	0.42	0.35	0.46	0.0000	-0.0400	0.0000	0.0400	0.0000	0.0016	1	0.070000	0.110000	0.040000
94	C1F78A	15897	G(O)3	solo	0.53	0.62	0.53	0.63	0.0000	-0.0100	0.0000	0.0100	0.0000	0.0001	1	0.090000	0.100000	0.010000
95	C1F78A	15897	NG(U)K1	solo	1.64	1.7	1.64	1.7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1	0.060000	0.060000	0.000000
96	C1F78A	15897	S(A)N1	solo	2.1	2.17	2.11	2.18	-0.0100	-0.0100	0.0100	0.0100	0.0001	0.0001	1	0.070000	0.070000	0.000000
97	C1F78A	15897	G(AA)1	solo	2.3	2.48	2.3	2.54	0.0000	-0.0600	0.0000	0.0600	0.0000	0.0036	1	0.180000	0.240000	0.060000
98	C1M59A	18070	H(AA)6	solo	2.87	2.95	1.9	1.93	0.9700	1.0200	0.9700	1.0200	0.9409	1.0404	0	0.080000	0.030000	-0.050000
99	C1M59A	18070	G(O)2	solo	3.67	3.75	3.67	3.8	0.0000	-0.0500	0.0000	0.0500	0.0000	0.0025	1	0.080000	0.130000	0.050000
100	C1M59A	18070	G(O)3	solo	3.84	3.92	3.84	3.95	0.0000	-0.0300	0.0000	0.0300	0.0000	0.0009	1	0.080000	0.110000	0.030000
101	C1M59A	18070	G(E)3	solo	4.03	4.24	4.04	5.44	-0.0100	-1.2000	0.0100	1.2000	0.0001	1.4400	0	0.210000	1.400000	1.190000
102	C1M59A	18140	D(I)1	solo	0.35	0.43	0.34	0.4	0.0100	0.0300	0.0100	0.0300	0.0001	0.0009	1	0.080000	0.060000	-0.020000
103	C1M59A	18140	L(E)K1	solo	0.66	0.74	0.67	0.85	-0.0100	-0.1100	0.0100	0.1100	0.0001	0.0121	0	0.080000	0.180000	0.100000
104	C1M59A	18140	G(E)3	solo	0.9	0.98	0.92	0.98	-0.0200	0.0000	0.0200	0.0000	0.0004	0.0000	1	0.080000	0.060000	-0.020000
105	C1M59A	18140	J(A)N4	solo	1.22	1.35	1.24	1.32	-0.0200	0.0300	0.0200	0.0300	0.0004	0.0009	1	0.130000	0.080000	-0.050000
106	C1M59A	18140	C(E)NG2	solo	3.44	3.5	3.14	3.25	0.3000	0.2500	0.3000	0.2500	0.0900	0.0625	0	0.060000	0.110000	0.050000
107	C1M59A	18140	J(A)N4	solo	3.6	3.62	3.45	3.49	0.1500	0.1300	0.1500	0.1300	0.0225	0.0169	0	0.020000	0.040000	0.020000

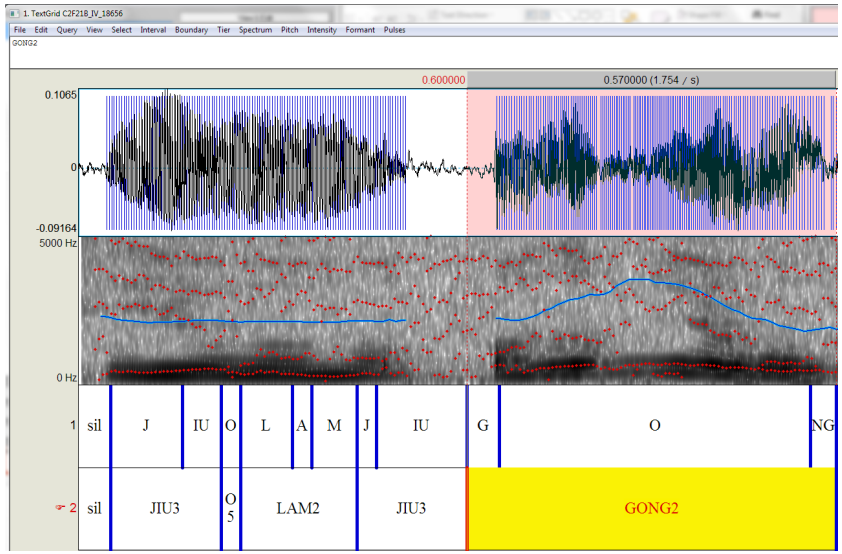
- Manual ("Gold Standard") Measurements taken of left & Right boundaries of Monopthongs
- Compared to Auto boundaries: Differential on left & right, ABS of diff., diff. of total length

T	U	V	W	X	Y
Solo Model Metrics		Gen Model Metrics		All Model Metrics	
RMSD - LEFT Boundary	0.18527152	RMSD - LEFT Boundary	0.193158	RMSD - LEFT Boundary	0.213991
RMSD - RIGHT Boundary	0.18690933	RMSD - RIGHT Boundary	0.197117	RMSD - RIGHT Boundary	0.207087
No. Vowels in Target	383	No. Vowels in Target	368	No. Vowels in Target	382
% Vowels in Target	81.84%	% Vowels in Target	78.63%	% Vowels in Target	81.62%
Avg. Auto V. Length	0.126816	Avg. Auto V. Length	0.123650	Avg. Auto V. Length	0.132073
Avg. V. Length deviation	0.013920	Avg. V. Length deviation	0.010753	Avg. V. Length deviation	0.019176

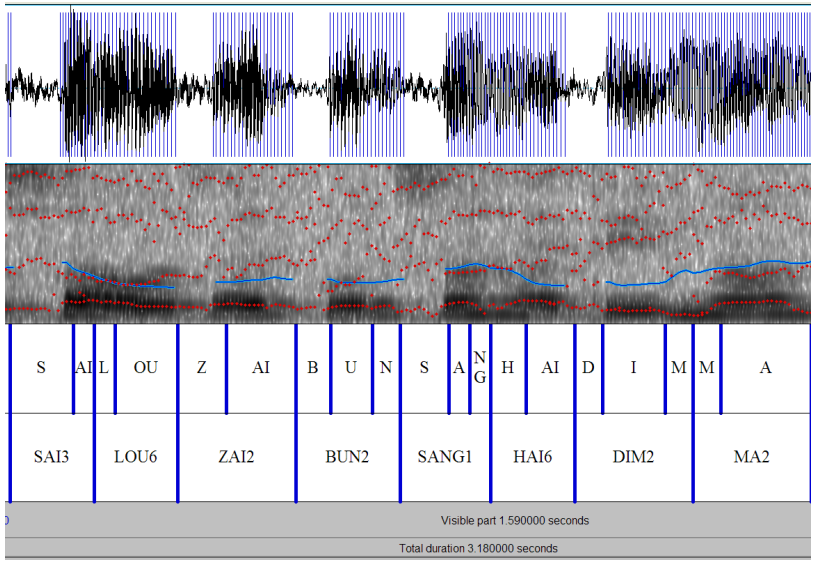
- Root-Mean-Square-Deviation taken of each boundary (Chen et al 2004)
- Average Length of vowels for each model
- % of vowels' centres (by "Gold Standard") which fall within the auto-aligned boundaries

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y)^2}{n}}.$$

Transcription Issues

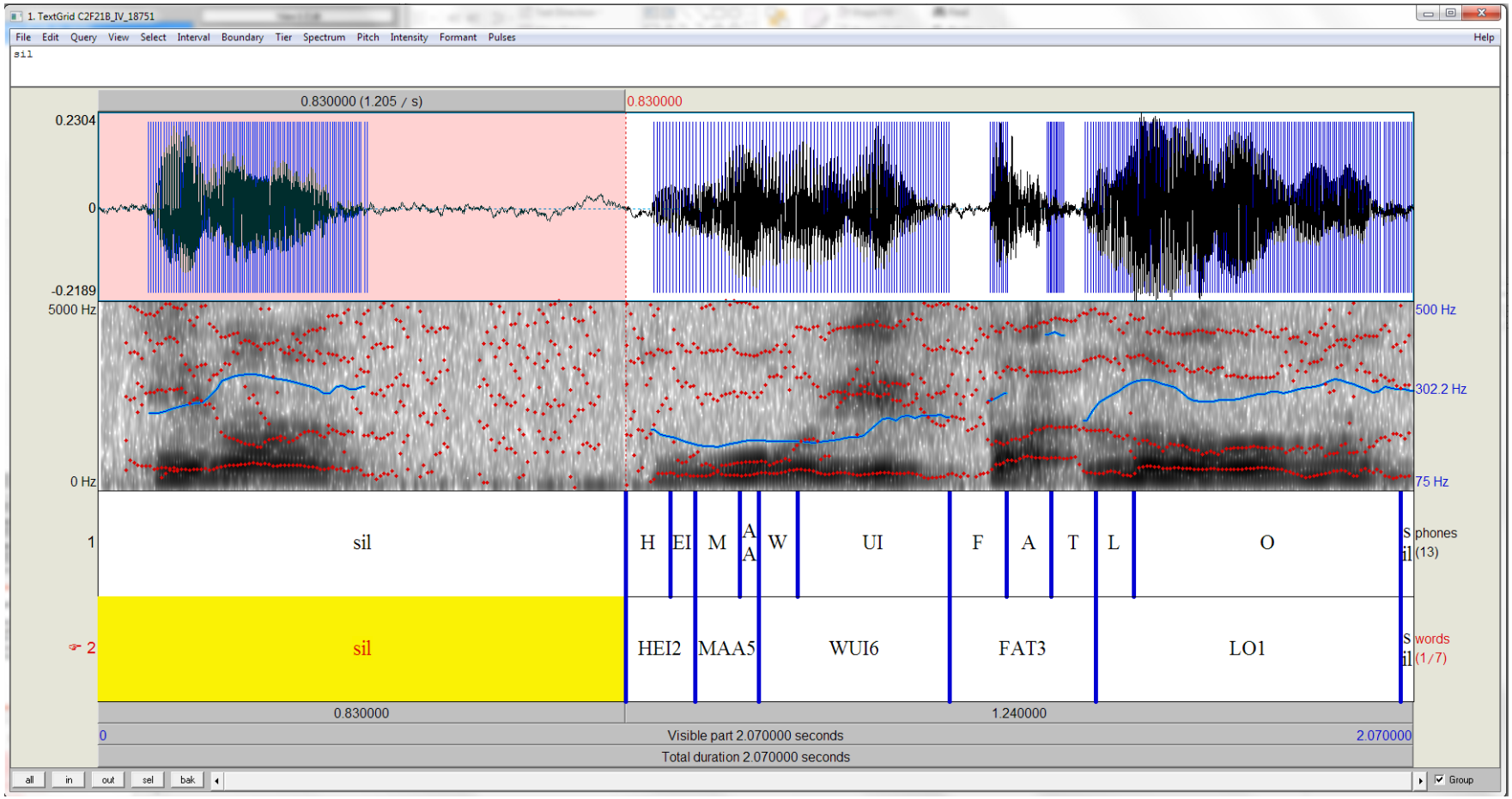


Entirety of “O5 Lam2 Jiu3” within “Gong2” boundaries



Same file: The aligner “Catches up” and aligns later sections with excellent accuracy

Modeling Silence

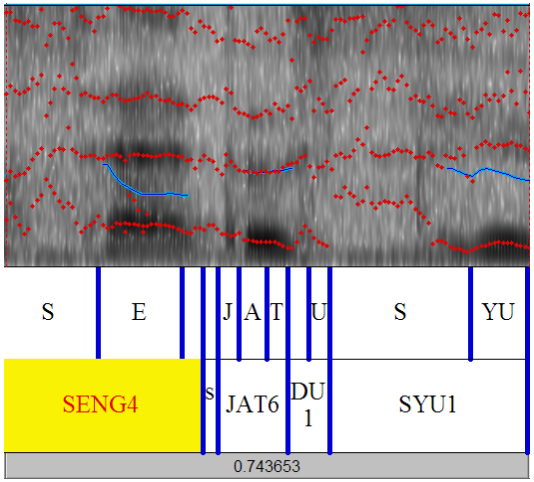


Aligner places "Hei2 Maa5" audio signal within silence

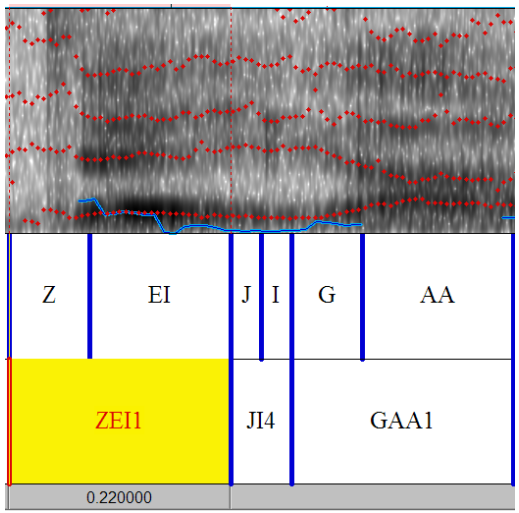
- The effect is more common in Solo-aligned textgrids
- Hypothesis: Silence modelling is better with more data for model training

Syllable Fusion Issues

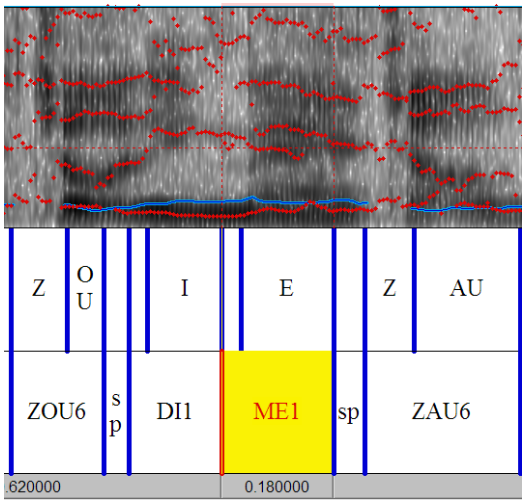
Wong (2006)



Fusion of Seng-Jat--> Set



Fusion of Za-Hai --> Zei



Fusion of Mei-Je --> Me

- Some rare examples cause problems: Seng Jat
- However, when we use a closer transcription, the aligner does well

Results Table

In spite of problems, quite accurate:

	SOLO	GEN	ALL
Root Mean Square Deviation – Left Boundary	0.185	0.193	0.214
Root Mean Square Deviation – Right Boundary	0.187	0.197	0.207
# of Vowels in Target	383	368	382
% Vowels in Target	81.84%	78.63%	81.62%
Avg. Auto V. Length	0.127s	0.124s	0.132s
Avg. V. Length Deviation	0.014s	0.011s	0.019s

- Solo-trained model has the lowest deviation from gold-standard boundaries
- All-trained model predicts longer vowels: hence higher % of vowel centres within boundaries, despite high deviation
- Overly-long segment prediction would be bad for studies of length, VOT, etc.

Summary

- Is Prosody-lab aligner effective at producing sufficiently accurate transcript alignment to permit automated measurement of vowel data? YES
- What is the best baseline to start with
  - All speakers together (ALL)?
  - Each generational group separately (GEN)?
  - Each speaker individually (SOLO)?

Discussion

- Is Prosody-lab aligner effective at producing sufficiently accurate transcript alignment to permit automated measurement of vowel data?
  - Yes, Overall, 80% accuracy for all three models
  - Can still be a useful tool in facilitating the vowel measurement process with a preliminary estimate of where the vowel boundaries are
  - Boundaries can be manually adjusted later.

Discussion

- What is the best baseline to start with
  - ALL
    - More data used, but model overgenerates → resulted in high RMSD
  - SOLO
    - Slightly more accurate and smaller RMSD than ALL and GEN models, but not much data / too much data lost to training
  - GEN
    - A reasonable compromise between amount of data used in training vs. general accuracy

Conclusion

- The GEN model works better than ALL (contrary to expectations) possibly because of significant inter-generational differences (cf. Tse 2015)
- Yet, even with as much variation as present, it is still generally accurate, and can be a useful tool for Cantonese corpus-based studies.
- Useful for any study that requires segmental boundary information
  - Ex: VOT, vowel length, vowel formant measures, tone, consonants, etc

감사합니다 Дякую Grazie molto Спасибо 多謝 gratsiə namuor:ə

HLVC RAs:

Cameron Abma
Vanessa Bertone
Ulyana Bila
Rosanna Calla
Minji Cha
Abigail Chan
Karen Chan
Joanna Chociej
Sheila Chung
Tiffany Chung
Courtney Clinton
Rachel Coulter
Radu Craioveanu
Marco Covi
Zahid Daudjee
Derek Denis
Tonia Djogovic
Joyce Fok
Paolo Frasca
Matt Gardner
Rick Grimm
Dongkeun Han
Natalia Harhaj
Taisa Hewka
Melania Hrycyna
Michael Iannozzi
Diana Kim
Janyce Kim
Iryna Kulyk
Mariana Kuzela
Ann Kwon
Alex La Gamba
Carmela La Rosa
Natalia Lapinskaya
Kris Lee
Nikki Lee
Olga Levitski
Arash Lotfi
Samuel Lo
Paulina Lyskawa
Rosa Mastri
Timea Molnár
Jamie Oh
Maria Parascandolo
Rita Pang
Tiina Rebane
Hoyeon Rim
Will Sawkiw
Maksym Shkvorets
Vera Richetti Smith
Anna Shalaginova
Konstantin Shapoval
Yi Qing Sim
Mario So Gao
Awet Tekeste
Josephine Tong
Sarah Truong
Dylan Uscher
Elaine Wang
Ka-man Wong
Junrui Wu
Olivia Yu
Minyi Zhu

Collaborators:

Yoonjung Kang
Alexei Kochetov
Naomi Nagy
James Walker

Funding:

SSHRC, University of Toronto, Shevchenko Foundation

References

Chen, L., Liu, Y., Harper, M. P., Maia, E., & McRoy, S. (2004). Evaluating Factors Impacting the Accuracy of Forced Alignments in a Multimodal Corpus. In LREC. Retrieved from https://www-new.comp.nus.edu.sg/~rpnlpir/proceedings/lrec-2004/pdf/307.pdf

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. Canadian Acoustics, 39(3), 192–193.

Nagy, N. (2011). A Multilingual Corpus to Explore Variation in Language Contact Situations. Rassegna Italiana Di Linguistica Applicata, 43(1/2), 65–84.

Rosenfelder, I., Fruehwald, J., Evanini, K., & Yuan, J. (2011). FAVE (Forced Alignment and Vowel Extraction) Program Suite. Retrieved from http://fave.ling.upenn.edu

Wong, Wai Yi Peggy. 2006 "Syllable Fusion in Hong Kong Cantonese Connected Speech." Ph.D. Dissertation. The Ohio State University.

- Slides will be available at http://www.pitt.edu/~hbt3/presentations.html
- Thank you!
- 多謝晒!

HERITAGE LANGUAGE VARIATION AND CHANGE IN TORONTO

HTTP://PROJECTS.CHASS.UTORONTO.CA/NGN/HLVC