THE DESCRIPTION OF PHD'S CORPUS

Thesis topic: ***Glottalization, tonal contrast, and intonation: an experimental study of Kim Thuong dialect of Muong***

1. Data collection:

- Speech materials: Minimal sets/ pairs of real words: ***12 minimal sets* of 5 smooth tones** (8 complete minimal sets + 4 near minimal sets) + ***3 checked minimal pairs***. See details in the tables below (section 4).

- Method of collection: each target word is required to speak four times: two times in isolation and two times in a carrier sentence.

- The carrier sentence is a question including 4 words:

/ɟa^{2}	măt^{6}	_____	căŋ^{3}/
2sg	to_know	<target_item>	INTERROGT

"Do you know _____?

Not only the target words but also all 3 frame words of the carrier sentence is annotated and processed with Peakdet. Hence, for each speaker, we have a total maximum of 660 tokens.

The total corpus (per speaker):

((5 x 12) + (2 x 3) + (5 x 12 x 4) + (2 x 3 x 4)) x 2 = 660 (items)

Elements	Meaning
5	5 tones in smooth syllables
12	12 minimal sets (of 5 smooth tones)
2	2 tones in checked syllables
3	3 minimal pairs (of 2 checked tones)
4	4 syllables of the carrier sentence (1 target word + 3 frame words)
2	2 repetitions

<table>
<tr><td colspan="3">Total corpus:
660 items</td></tr>
<tr><td colspan="2">Target syllables:
264 items</td><td>Frame syllables:
396 items</td></tr>
<tr><td>In isolation: 132 items

- 24 tokens each smooth tone (x 5 tones)
- 6 tokens each checked tone (x 2 tones)</td><td>In carrier sentence: 132 items

- 24 tokens each smooth tone (x 5 tones)
- 6 tokens each checked tone (x 2 tones)</td><td>
- /ɟa^{2}/ : 132 tokens
- /măt^{6}/ : 132 tokens
- /căŋ^{3}/ : 132 tokens</td></tr>
</table>

2. Annotation

Annotation scheme: four digits. Structure: AABC, detailed below.

Position	Meaning	Number code
AA	The UID of the target syllable, in two figures. Numbers from 1 to 9 are padded up with a zero. These numbers corresponding to the UID numbers (second column) in table of minimal sets underneath.	01 – 40: words of 8 minimal sets (8*5) 41 – 60: words of 4 near – minimal sets (4*5) 61 – 66: words of 3 checked minimal pairs (3*2)
B	The nature of the annotated rhyme: B = 0: target syllable said in isolation 1 ≤ B ≤ 4: token from the carrier sentence. B indicates the order of the token inside carrier sentence, 3 means target syllable.	0: target syllable (in isolation) 1: 1st syllable of frame - /ja2/ 2: 2nd syllable of frame - /măt6/ 3: 3rd syllable of frame - target syllable (in carried sentence) 4: 4th syllable of frame - /căŋ3/
C	repetition	1: first time 2: second time

For example:

- Item with UID 6801 means: the check syllable /kap7/ in isolation, first time.
- Item with UID 3632 means: the target syllable /ma5/ in carrier sentence, second time.
- Item with UID 3611 means: the first syllable /ja2/ of the carrier sentence in case /ma5/ is target syllable, first time.

As a consequence, based on two first digits (AA), we can classify the tokens into 7 sorts of tone:

T1	T2	T3	T4	T5	T6	T7
04	03	02	05	01	61	62
09	08	07	10	06	63	64
14	13	12	15	11	65	66
19	18	17	20	16		
24	23	22	25	21		
29	28	27	30	26		
34	33	32	35	31		
39	38	37	40	36		
44	43	42	45	41		
49	48	47	50	46		
54	53	52	55	51		
59	58	57	60	56		

3. Data processing using Peakdet: input and output

3.1. Input:

After annotation, two input files are prepared for data processing with Peaket:

- EGG file (.WAV): the EGG signal was recorded simultaneously with acoustic signals, now are extracted from the annotated audio file.

- REGIONS file (.TXT): this is a text file containing the list of regions created at annotation (i.e. the time codes of the target items annotated in SoundForge).

3.2. Output:

As final output, Peakdet produces a ".mat" file. The result is placed in a matrix, containing one sheet per token analyzed. It means the maximum data of each speaker will contain 660 sheets corresponding to 660 rhymes. However, this absolute quantity is not met in many cases. Because the carrier sentence are repeated many time, the speakers tend to reduce the first syllable. The specific status of each data will be provided in the table of speakers in section 5.

The structure of a single matrix:

Each sheet presents the main results of a token in a single matrix including 10 columns x 100 line.

Column	Value	Unit
1st	the beginning of cycle	ms
2nd	the end of cycle	ms
3th	f_0	Hz
4th	DECPA: Derivative-EGG Closure Peak Amplitude	
5th	Oq determined from raw maximum without smoothing	%
6th	DEOPA	
7th	Oq determined from maximum after smoothing	%
8th	Oq determined from peak detection without smoothing	%
9th	Oq determined from peak detection after smoothing	%
10th	The open quotient values retained after user's verification	%

4. Speech materials: tables of minimal sets/pairs

12 minimal sets (8 complete minimal sets + 4 near minimal sets)

No	**UID**	**IPA**	**PS**	**Vietnamese**	**English**
1	01	**paj5 (tʰăj1)**	n	Sải tay	Armspan
	02	**paj3**	n	(Cái) vại	A kind of porcelain jar to salt vegetables or tubes
	03	**(tɤ̆p6) paj2**	n	Đập nước	Barrage
	04	**paj1**	n	Vải	Cloth
	05	**paj4**	n	Quả, trái	Fruit
2	06	**rɔ5**	n	(Con) rùa	Tortoise
	07	**rɔ3**	v	Mò (cua, cá)	To find crab or fish (by hand) in a swamp
	08	**rɔ2**	adj	No	To be sated
	09	**(ʔɤ5) rɔ1**	adj	Rảnh rỗi	Idle
	10	**(paj4) rɔ4**	n	Hoa chuối	Banana flower
3	11	**pa5**	n	Bà	Grand-mother/ an old woman
	12	**pa3**	v	Bám, trạm vào vai	To cling on one's shoulder

	13	**pa2**	n	Ba	Three
	14	**pa1 (tʰiən5)**	v	Trả (tiền)	To pay
	15	**pa4**	v	Vá (xăm)	To patch
4	16	**laj5**	n	Lưỡi	Tongue
	17	**(pɤ2) laj3**	v	Trở lại,	Return
	18	**laj2**	v	Lai	To take someone up by bike/ motobike?
	19	**laj1 (tɤ̆m5)**	n	(Cái) rào/ chắn ao	A fence to prevent fish go out from lake
	20	**laj4**	v	Lái	To drive
5	21	**taj5**	v	Đãi gạo	To wash rice
	22	**taj3**	v	Lôi (nứa, bương bằng tay hoặc xe máy)	To drag bamboo by hand, or by motorbike
	23	**taj2 (nan3)**	n	Tai nạn	Accident
	24	**taj1**	n	Thác (nước)	Waterfall
	25	**taj4**	n	Đái	To pee
6	26	**kɔ5**	n	(Con) cò	Stork
	27	**kɔ3**	v	Nói	To speak
	28	**(kiɛw4) kɔ2**	v	Kéo co	Tug of war
	29	**kɔ1**	n	Cỏ	Grass
	30	**kɔ4**	v	Có	To have
7	31	**kiɛŋ5**	n	Cái chum nhỏ	A jar to
	32	**kiɛŋ3**	pre	(Ở) cạnh	Next to
	33	**kiɛŋ2**	n	Canh	Soup
	34	**kiɛŋ1**	n	(Cái) kẻng	A kind of gong to gather people
	35	**kiɛŋ4**	n	Cánh	Wing
8	36	**ma5**	n	Lỗ cua, lươn	Cave hole of eel or crab
	37	**ma3**	n	Mạ	Rice seedling
	38	**ma2**	n	(Con) ma	Ghost
	39	**ma1**	n	(Mồ) mả	Tomb
	40	**ma4**	n	Má	Cheek
9	41	**ŋa5**	v	Ngã	to fall
	42	**ŋa3**	v	Ngứa	To itch
	43	**ŋa2**	n	Chói mắt	Dazzle
	44	**ŋa1**	v	(Nằm) ngả	To recline
	45	**na4**	n	Cung tên	Archery
10	46	**ka5**	n	Cà (tím)	African eggplant
	47	**ta3**	n	Tạ	Dumbbells
	48	**ka2**	n	(Con) gà	Chicken
	49	**ka1**	adj	To	Big
	50	**ka4**	n	Cá	Fish
11	51	**kaj5**	v	Cài (cúc áo)	To fasten buttons of clothing
	52	**paj3**	n	(Cái) vại	A kind of porcelain jar to salt vegetables or tubes
	53	**kaj2**	n	(Cái) gai	Thorn
	54	**kaj1**	n	(Rau) cải	Borecole
	55	**kaj4**	adj	(Con) cái	Female
12	56	**ku5**	adj	Cũ	Old, ancient
	57	**tu3**	v	Tụ (máu)	To converge (blood)

	58	**ku2**	n	Trâu	Buffalo
	59	**ku1**	n	Củ (khoai)	Tubes
	60	**ku4**	n	Cú mèo	Owl

3 Checked minimal pairs

N₀	**UID**	**IPA**	**PS**	**Vietnamese**	**English**
1	61	**pat6**	n	Sàn (nhà)	Floor made by a kind of bamboo
	62	**pat7**	n	Bát	Bowl
2	63	**rɔc6**	n	Ruột	Intestine
	64	**rɔc7**	v	Rót (nước)	To pour (water)
3	65	**lak6**	n	(Củ) lạc	Peanut
	66	**lak7**	adj	Lác	Squint eye

	Phonemes	Appearance	Minimal set(s)
Smooth syllables			
Initial consonants	p	2	1, 3
	t	1	5
	k	5	6, 7, 10, 11, 12
	m	1	8
	ŋ	1	9
	l	1	4
	r	1	2
Vowels	a	8	1, 3, 4, 5, 8, 9, 10, 11
	ɔ	2	2, 6
	u	1	12
	iɛ	1	7
Final consonants			
	j	4	1, 4, 5, 11
	ŋ	1	7
Checked syllables			
Initial consonants	p	1	1
	l	1	3
	r	1	2
Vowels	a	2	1, 3
	ɔ	1	2
Final consonants	t	1	1
	c	1	2
	k	1	3

5. Information of speakers and their data

- M = male, F = female

- Number of data that can be analyzed: 10 man, 11 woman

- Number of data has been processed: 10 man, 10 woman

- Total: 26 participant, 28 files, 20 files have been process

ID	Birth year	Rec. time	EGG quality	Status of data	Size of .mat file
F1	1983	2018	Bad Crackling noise	No analysis	No analysis
F3	1984	2018	Good	First time: 421/460 tokens - Missing 35 first syllables (position AA1C): vowels reduced - Missing 4 tokens of a carrier sentence. One target syllable (in carrier sentence) is problematic. Second time: 660/660 tokens	100x10x 660
F7	1988	2018	OK	660/660 tokens	119x10x660
F8	1973	2018	Weak EGG	No analysis	No analysis
F9	1991	2018	Good	660/660 tokens	119x10x660
F10	1973	2018	Good	585/660 tokens (Missing 75 tokens) Missing minimal set 11 in carrier sentence, minimal set 12 and 3 minimal pairs both in isolation and carrier sentence (in first repetition) - 11 AA01 - 16 AA31 - 16 AA11, 16 AA21, 16 AA41	178x10x660
F11	1982	2018	Very weak EGG	No analysis	No analysis
F12	1972	2018	Good	660/660 tokens	111x10x660
F13	1967	2018	Good	646/660 tokens Missing 14 tokens (AA1C)	133x10x646
F14	1965	2018	OK But there are a few flat segments	No analysis	No analysis
F16	1956	2019	Weak EGG	No analysis	No analysis
F17	1955	2019	Good	660/660 tokens	100x10x660
F18	1988	2019	Weak EGG	No analysis	No analysis

F19	1975	2019	Good	660/660 tokens	100x10x660
F20	1989	2019	OK	660/660 tokens	100x10x660
F21	1988	2019	OK	660/660 tokens	111x10x660
M1	1979	2018	Good	660/660 tokens	100x10x660
M5	1955	2018	OK	660/660 tokens	100x10x660
M7	1984	2018	Good	660/660 tokens	100x10x660
M8	1975	2018	OK	660/660 tokens	100x10x660
M9	1990	2018	Good	660/660 tokens	100x10x660
M10	1988	2018	Good	660/660 tokens	100x10x660
M11	1987	2018	Good	660/660 tokens	100x10x660
M12	1937	2018	Signal out of range	No analysis	No analysis
		2019	Good	660/660 tokens	100x10x660
M13	1962	2018	Good	660/660 tokens	100x10x660
M14	1959	2018	OK	656/660 tokens Missing 4 tokens: 3 AA1C, 1 AA4C	100x10x656