Towards the Mo Piu Tonal System: First Results on an Undocumented South-Asian Language

Geneviève Caelen-Haumont

MICA Institute, HUST - CNRS/UMI 2954 - Grenoble INP
Hanoi University of Science and Technology, 1 Dai Co Viet St., Hanoi, Vietnam

genevieve.caelen@mica.edu.vn

Abstract

This paper presents the first results on the Mo Piu tonal system. This language is undocumented, unwritten and moreover in great danger. As the tasks of labeling phonetics and tones is hard to carry out when references on the language are lacking, this paper aims at presenting our method to try to build reliable data in order to understand the tonal system, and the main findings concerning the Mo Piu tonal system and finally at leading a comparison between these first results and the White Hmong tonal system.

Index Terms: undocumented language, tonal system, isolated words, automatic tones labeling.

1. Introduction

The present study belongs to the "Au Co" Project which started in 2008 in MICA Institute, which aims to help saving endangered languages and cultures in Vietnam. This project is based on the collaboration of a wide variety of experts like French and Vietnamese linguists, specialists of ethnic languages, and computer scientists.

First the project Au Co focused on an ethnic minority called "Mo Piu". According to our preliminary studies [1], the "Mo Piu" language is really an endangered language because it is uncharted, undocumented, unwritten and spoken only by 237 people in 2011. Moreover, the 7 or 8 ethnic groups in their surrounding do not understand this language. The Mo Piu language being not documented at all, it is urgent to study it before the fusion with the dominant Viet Nam culture.

The purpose is to start linguistic study and develop tools to save the Mo Piu culture and language, in the framework of two international CNRS-ANR-Blanc Projects "PI Language" and "AppSy". The study is undertaken according to two main objectives: developing at the same time the linguistic studies and the technologies for under-resourced languages [2].

Our first studies gave evidence that this ethnic minority belongs to the family of Hmong-Mien, but as a little and totally unknown branch. Among the different Hmong languages, the White Hmong is likely the most studied. According to Niederer [3], this Hmong branch presents 7 tones. If we consider now the Hmong language on the whole according to its different branches, the tones may vary from 3 to 11. So the Mo Piu tonal system cannot benefit from previous stable knowledge or cues from the Hmong family and we have to start work by the very first steps. The aim of this work is to get a more precise idea about the Mo Piu tonal system using a tool for an automatic tonal labeling (MISTRAL+ under Praat [4]), and comparing the Mo Piu tones to the White Hmong ones according to Niederer's findings [3], while specifying the methodology used for the analysis if this undocumented language.

2. Mo Piu ethnic minority and language

The Mo Piu ethnic minority which is of extremely small size lives in the remote North Mountains of Vietnam, with no roads reaching their village, and in a district where the foreigners are not free to access.

This minority is located in the mountains of North Vietnam along the Chinese border. The Mo Piu village situated in a sort of circus on the side of a hill, is named *Nam Tu Thuong*, meaning the "the stream river spring up" in Tày language.

Though the Mo Piu language is still unknown, the Hmong language(s) has been studied since a long time, and especially during the French period in Vietnam. As the Hmong people is a great community around the world settled not only in China, in Vietnam, but also in Laos, Thailand, and also in France (and the French Guiana), and in the USA, more and more studies are developed since several decades, for instance [3] [5] [6] [7]. The phonetics and tonal system of the White and Green Hmong are now well known.

For instance Niederer [3] states that the White Hmong uses 58 consonants, 11 oral vowels, simple or complex, 2 nasalized ones, and 7 tones, which represents a phonetic system rich and complex. Among the tones, only one is rising. As for the tonal systems among the other Hmong languages, it is spreading from 3 to 11 tones, but the Green and White Hmong use 7 tones.

For writing the Hmong language, several systems have been supplied, and among them the system by Smalley, Barney and Bertais [8], the "Romanized Popular Alphabet". This alphabet presents the great advantage of not using diacritics, neither for the phonetics nor for the tonal transcription.

3. Method and Mo Piu corpus

3.1. The Mo Piu corpus in 2011

Three field trips were undertaken in 2009, 2010 and 2011. On the whole, our sound and video corpus is composed of 36h for films, 35h for speech (French / Vietnamese / Mo Piu), fluent speech and lists of words included, 1h for songs, 2350 images, 335 video-clips, and more than 2000 sound files and more than 2000 video ones. All the speech recordings are filmed.

The continuous speech is composed of a large set of cultural inquiries, the domains of the Mo Piu life being split up in about 50 inquiries (questions / responses), tales, life stories, and free comments on drawings or video. We use several lists of words and especially the Calmsea list [9], [10], with 200 main entries. Just a part of them (plants, parts of body, animals, directions, natural phenomena, numbers…)

was until now registered but with several repetitions (3 at least) and by 20 male and female speakers.

3.2. The Mo Piu corpus for the present study

For this experiment, the phonetic and tonal analyses have been carried out at once. We focused on one speaker's speech. A part of the corpus composed of 43 isolated words (x 3 repetitions) borrowed from the Calmsea list and concerning the body parts (humans and animals) were manually segmented in words, phonetic units and pauses, labeled in French and in Mo Piu, using IPA symbols. These 43 isolated words may be constituted of 2 or 3 lexical units as well, leading finally to 219 lexical items. Among some of these compound words, we identified some other regular words of the list. For instance *nail* and *claw* in Mo Piu are corresponding respectively to 2 lexical units */nail*, t͡ɛ˧ pɐ/ and */claw*, t͡ɛ˧ t͡sa/, where /t͡ɛ˧/ means *finger*.

Though the phonetic part is not the aim of this present paper, it had nevertheless to be carefully analyzed, segmented and labeled in order to help correctly structuring the tones. It is indeed a very difficult task because there are no direct references for the Mo Piu phonetics or tones. Of course the Hmong phonetic and tonal system (from several Hmong minorities) is known, especially for the White and Green Hmong, but as said above in paragraph 1, the tonal system varies a lot among Hmong branches.

Figure 1: *Example of the final step of MISTRAL procedure providing manual and automatic tiers. Here the word "langue/tongue"(tier 1) corresponding to the Mo Piu /ᵐbla/ in the tier 2, then the suite of the phonetic classes (tier 3), all manually labeled. Now from bottom, the automatic tiers: respectively tier 8, the semi-tone value of the boundary level, tier 7, the same expressed in Hz, tier 5, the same expressed in a 5 levels scale, tier 6, the corresponding tones, same scale.*

At the beginning of the Mo Piu study, two choices were possible: either starting with the Hmong phonetic and tonal system and thus comparing the variations with the Mo Piu, or starting directly from the Mo Piu data. It is this last method we adopted in order to not becoming influenced by the Hmong system and getting more confidence on the specific Mo Piu system. Thus the comparison will be more firmly grounded.

As for most of the Hmong languages, the syllabic structure is C (C) V (C, V). The first consonant (plosives) may be preceded with a prenasal unit. For the tonal study, we focused on the vocalic part.

Before any analysis, first of all, the F0 minimum, mean maximum, (respectively 80, 131, 185Hz) of the speaker (here named VAP01) have been automatically computed on a large speech sample issuing both from continuous speech and isolated words.

After the tasks of segmentation and labeling (see Figure 1 above) have been manually completed under Praat, the automatic procedure MISTRAL+ [11] is launched. This procedure was just rewritten to update its functionalities concerning the tonal languages, especially in the automatic annotation and the use of IPA symbols.

When the step of Praat Manipulation consisting in stylizing the F0 line and adjusting the F0 targets (under the control of the sound) for giving a correct shape and tune to the tone is achieved, a semi-automatic tonal labeling is got due to the MISTRAL+ procedure (see Figure 1 above) via MOMEL [12]. As the overall range can be automatically shared from 2 to 9 levels according the user's needs, 5 levels were chosen in concordance with the standard method for studying the tonal systems. The scale is numeric.

Apart from the lexical, phonetic and comments which are manual tiers, the procedure automatically provides the melodic values of each target (boundaries of each phonetic and/or tonal segment) and the intra-vocalic values (such as /32/ or /43-32/ when the tone is twofold). For instance /32/ means that the tone (falling) starts at level 3 (over 5) and then reaches the level 2 (same scale).They are expressed in Hz and semi-tones, and the temporal address of each boundary in ms. Once the MISTRAL procedure completed an excel file (see the Table 1 below) is automatically filled up with the contents of all the Praat tiers (manual and automatic).

3.3. Post-processing tones classification

The first task consists then in unifying the codings and especially the levels of the tones. Among the tones values, there are 'tonologic' units and their variants. Let us precise that the variants are not only the variants of the phonological system, but also the variants of our semi-automatic procedure. As said before, the overall range was automatically divided in a semi-tone scale into 5 levels. But the thresholds are sharp while the tone levels are naturally more fuzzy, and this confrontation causes an obvious mismatching. For instance some tones may be automatically labeled as /32/ and in fact may correspond to /33/ because their value may be very close from the level threshold.

Besides what really means a "plateau"? In the right meaning, it means a recto-tono voice with no modulation. To which extent this notion may be applied? If the Mo Piu tones are observed in detail, their shapes are quite amazing: a great amount of them present a "slope" simply flat, without any F0 modulation (see the Figure 1 above for instance). So many of them refer to the concept of plateau because of their F0 monotonous, but if one considers the duration, the F0 range between the vowel beginning and its end may also vary a lot.

Therefore the three repetitions of the same word (and also repetition of phonetically close words) are necessary to get

more confidence to find a prototype. Thus the method used is as follow:

- Sorting the data according to the vowels (oral, nasal, diphthongs), putting apart the consonants and pauses,
- Then sorting the data according both to the tone levels, from the lowest to the highest ones, and at the same time to the chronologic order,
- According to one set of vowels belonging to the same words, observing the population of the same tonal level (for instance /21/), and the deviant ones (ex: /22/),
- Then examining the deviant tone levels, and observe if it is ±2 Hz around the threshold between 2 levels of the range. This range is subjective.
- If the value is comprised in this range, correcting it according to that of the other repetitions of the word,
- If not, some causes may be found: for instance an emphasis is often made on the third repetition of the Mo Piu word, rising the higher value of the vowel and / or lowering the lower one, and making its duration shorter. On the other hand, the list effect tends to lower the last part of the last vowel.
- If the 2 other repetitions are conversely in concordance, then correcting the value of the third vowel.

4. Results

4.1. General overview of the results

In the corpus of the present study, over 260 vowels and thus 520 tones boundaries, 69 of them present a variant (27%). Among that ones, 45% are a left boundary variant, 42% a right one, and 13% both of them (8% of which belonging to the third repetition of the lexical item). Besides, 46% (32/69) of the variants correspond to the threshold range (±2 Hz), and 38% (18/69) to the last word repetition. The 28% (19/69) variants still remaining may correspond to hesitation, focus, method, or speaker mistake (more than ±2 Hz around the threshold).

1	Data Nb	Word Nb	Speaker	Words	phon classes	Mo Piu	Tones	Derivative	Duration	Hz (left bound.)	Phonetic	Beginning	End
2	1	1	VAP01	ventre 1a	C	hn̂a	35	22,51	195 ms	133 Hz	h	11,7 ms	11,89 ms
3	2	1	VAP01	ventre 1a	Cn	hn̂a	54	-22,37	72	171	ɲ	11,89	11,96
4	3	1	VAP01	ventre 1a	V	hn̂a	43	-14,33	286	156	a	11,96	12,25
5	4	1	VAP01	PP		hn̂a	33+	1,26	1114	123		12,25	13,36
6	5	1	VAP01	ventre 2a	C	hn̂a	34	17,69	158	134	h	13,36	13,52
7	6	1	VAP01	ventre 2a	Cn	hn̂a	44-	-18,64	54	157	ɲ	13,52	13,57
8	7	1	VAP01	ventre 2a	V	hn̂a	43	-12,16	313	148	a	13,57	13,89
9	8	1	VAP01	PP		hn̂a	34	2,56	1135	119		13,89	15,02
10	9	1	VAP01	ventre 3a	C	hn̂a	44+	4,19	167	141	h	15,02	15,19
11	10	1	VAP01	ventre 3a	Cn	hn̂a	44+	3,33	60	146	ɲ	15,19	15,25
12	11	1	VAP01	ventre 3a	V	hn̂a	43	-11,65	309	148	a	15,25	15,56
13	12	1	VAP01	PP		hn̂a	33+	0,32	4433	120		15,56	19,99
14	13	2	VAP01	sang 1a	Cn	nt͡ʃa	34	9,81	132	130	n	19,99	20,12
15	14	2	VAP01	sang 1a	C	nt͡ʃa	43	-50,59	77	140	t	20,12	20,20
16	15	2	VAP01	sang 1a	C	nt͡ʃa	34	53,79	87	112	ʃ	20,20	20,29

Table 1. *Example of an xls file providing the manual and automatic data issuing from the Praat Texgrid. The columns B to K are automatically filled up from the final Praat TextGrid, and the columns H, I and J are automatically computed by the MISTRAL+ procedure.*

4.2. Towards the tonal Mo Piu system

4.2.1. The tonal perspective

Applying the rules described above (see 4.1.), 2 true plateau (varying at the most of 3 Hz), 5 simple tones and 1 twofold one appeared in the data. They are respectively shaped as /22, 33, 21, 32, 42, 43, 54 and 43-32/. In fact 4 simple tones present only a difference of one level, the fifth one, two levels such as a double tone (see Table 2 below). On the other hand, one can observe that there is no rising tone in the present data.

In this Table 2, the examples may correspond to a single lexical unit or to an item of a compound word. Anyway all are lexical units. The third column corresponds to the White Hmong tonal system established by Niederer [3].

The previous and primary study about Mo Piu tones [1] we undertook in continuous speech, showed indeed a great amount (68%) of flat tones (or plateaux) and among them, 42% were plateaux of the middle range. Moreover the previous study (using four levels) stated 4 plateaux: high, high-middle, low-middle and low tones, a falling one from high to middle range, and a double tone analyzed as high-middle/middle. So this present study is in phase with the primary one. One difference nevertheless: the present study does not attest the presence of rising tones while the first one does it (middle to high level), in the same way as the Niederer's findings [3].

If we compare now more thoroughly (Table 2 below) these results with Niederer's ones concerning the White Hmong, one can observe that there is some concordance between the 2 tonal systems. However in Niederer's study, the F0 variation within a plateau is not specified, so we do not know which exact shape they take. The tones in concordance with our results are /33, 22, 21, 42/. More investigation is necessary to check if the tones which do not exist in White Hmong but in Mo Piu are real tones or speaker/method variants. We noticed also in the vowels some features such as creaky voice, breathy voice and it remains to check if these features are phonologic.

Besides, among these Mo Piu tones the population is varying a lot (see Figure 3 below). Though the /21/ and /42/ tones are not numerous at all, their specific shapes incline to think that it is likely to refer respectively, for the first one, to the low tone, and for the second one, to a more contrastive tone. For the /21/ tone, there is no variant at all, which strengthens the presumption of its existence.

Shapes	Mo Piu Tones	White Hmong tones	English	Mo Piu
Plateaux	not found	55	not found	not found
	33	33	arm	/tjã/
	22	22	egg	/taa/
One level slope	21	21 glottalized	nasal mucus	/m͡bja/
	32	not found	skin	/ta͡ɜ/
	43	not found	neck	/nd͡zɜ/
	54	not found	horn	/kɔ̃/
Two levels slope	not found	52	not found	not found
	42	42 murmured	dejection	/fa/
	not found	24	not found	not found
bitonal	43-32	not found	bouche	/m͡bjɐ/

Table 2. *The Mo piu tones found in the present corpus*

As for the twofold /43-32/ tone, one might think that it could be also classified as /42/. However one argument prevailed against that hypothesis: while the mean duration of the 260 vowels (or 260 tones) is 287 ms, that of the double tone is 423 ms (minimum 380 ms, maximum 490 ms).

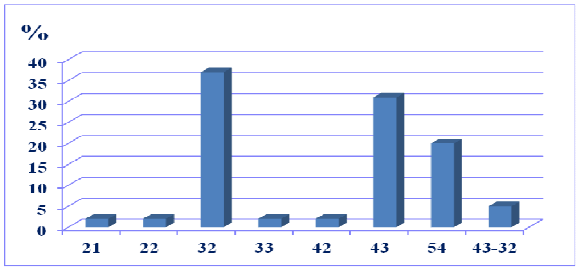

Figure 3: *Percentages of the population of each Mo Piu tone of the data set.*

The existence of the /42/ tone is problematic essentially because of its low population in the present corpus of the study, but it was attested in the previous study [1] concerning continuous speech, and moreover it is the only tone presenting a more contrastive slope. A further investigation could solve this problem. The other tones /32, 43, 54/ are more often attested in this corpus.

4.2.2. Discussion

In the previous study, the first results induced that duration could have a role to play in the tonal system. Let us consider this hypothesis further. As written above, the mean duration of all the vowels of this study is 287 ms while the standard-deviation is 105 ms, which means a great range between the shortest and largest duration (i.e. 80 ms / 540 ms). We made several sorting according to each kind of tones, to words, to the vowel nature (simple vowel or diphtong), but no classification was clearly put to light. The use of isolated words might be the cause of this lack of clearness. Just one hypothesis might be worth exploring further: the tone /21/ seems to be 70% shorter than the other tones, but the data are not enough numerous to validate this finding.

For this undocumented language, we cannot carry out all the domains at once, we proceed step by step, it is unfortunately the specific problem of this kind of corpus that we have to allow. In fact we are facing to two problems: if the analysis is done on continuous speech, we cannot identify at the present time the word boundaries, and if we use isolated words, the tonal system misses the duration contrast.

Conclusions

This study presents the first results on the tonal study concerning an undocumented and unwritten South-Asian language (Mo Piu). Though the present study concerns a different corpus (isolated words and manual tonal labeling) than the previous one (continuous speech and semi-automatic tone labeling [1]), the results are in concordance. Concerning the comparison with the Niederer's results on the White Hmong tonal system, some tones seem to be shared (plateaux: /22, 33/; falling one slope /21, 42/). The next step of the study will consists in extending the experiment (same word list, new speaker) in order first to assess the actual variants of the speaker of this present study and secondly to consider the role of some features such as creaky voice, breathy voice, in the phonological tonal system of the Mo Piu language.

Acknowledgements

This research is granted thanks to 2 Projects ANR-CNRS Blanc "*Langues PI*" and *"AppSy"*.

References

[1] Caelen-Haumont, G., Cortial, B., Culas, C., Hong, T. D., Lê, T. X., Nguyen, T.N., Pannier, E., Salmon, J.-P., Vittrant A., Hoang, T.V., "Mo Piu minority language: data base, first steps and first experiments", *Proceedings of the Second International Workshop on Spoken Languages Technologies for under-resourced Languages,* SLTU, Penang, Malaysia, 2010, 42-50.

[2] Caelen-Haumont, G., Sam, S., Castelli, E., "Automatic Labeling and Phonetic Assessment for an Unknown Asian Language: the Case of the Mo Piu North Vietnamese Minority (early results)", *Proceedings of the International Conference on Asian Language Processing* (IALP), 2011.

[3] Niederer, B., "La langue Hmong", *Amerindia*, 26/27, 345-381, 2001-2002.

[4] Boersma, P., Weenink D., "Praat: doing phonetics by computer", http://www.praat.org/, 2011.

[5] Golston, C., & Phong Y., 2001, "Hmong loanword phonology", *Proceedings of HILP 5*, ed. C. Féry, A. D. Green, & R. van de Vijver, 40-57, Linguistics in Potsdam 12, Potsdam, University of Potsdam.

[6] Ratliff, M., "Meaningful Tone: A Study of Tonal Morphology in Compounds, Form Classes, and Expressive Phrases in White Hmong", Dekalb, Illinois: Center for Southeast Asian Studies, Northern Illinois University, 1992.

[7] Ratliff M., "*Hmong-Mien language history*", Canberra, Australia: Pacific Linguistics, 2010.

[8] Smalley, W., Vang, C.K, Gnia Y., "*Mother of Writing*", Chicago: University of Chicago Press, 1990.

[9] Swadesh, Morris, 1971, "*The origin and diversification of Language*", J. Sherzer (ed). Chicago : Adline.

[10] Matisoff, James A., 1978, "*Variational semantics in Tibeto-burman: the 'organic' approach to linguistic comparison*", Philadelphia: ISHI Publications.

[11] Weber, B., Caelen-Haumont, G., Tran, D.-D., Pham, B. H., 2012, "MISTRAL+: A dedicated tool for under-resourced languages analysis", *Proceedings of SLTU 2012*, (to be published).

[12] Hirst, D., Espesser, R., 1993, "Automatic modelling of fundamental frequency using a quadratic spline function, *Travaux de l'Institut de Phonétique d'Aix,* 15, 71-85.