A phonetic study of Nanai vowels using automated post-transcriptional processing techniques

YUN Jiwon, KANG Hijo, and KO Seongyeon
Stony Brook Univ. (USA), Chosun Univ. (Korea), and CUNY (USA)

1. Introduction

Recent years have witnessed a remarkable growth in Altaic field linguistics. One representative example of this sort is the research project "Research on Endangered Altaic Languages" (often abbreviated as "REAL") conducted by the Altaic Society of Korea (Kim et al. 2008, 2011), which has produced to date multiple descriptive grammars and a fair amount of online language materials. However, compared to this rather fruitful achievement in descriptive and documentary linguistics, relatively less effort has been made to analyze the collected materials from the general phonetic perspectives (cf. Kang & Ko 2012). A considerable deal of data collected through REAL are yet to be fully post-processed and utilized in linguistic analysis. This situation arouses general concerns about language archives of unannotated—thus not quite useful— recordings, sometimes rather harshly called "linguistic graveyards" (Newman 2013). Our primary goal is to show that the automatic annotation system we describe below is easily applicable to the Altaic language data and, with significantly reduced time and effort in annotation tasks, may help "revive" their usability as linguistic data.

For linguistic purposes, the audio data must first be annotated. It is not realistic, nor desirable, to manually annotate a large speech data set: not only does it take too much time and effort but also it does not guarantee the consistency of annotation due to the continuous nature of speech signal. Focusing on the segmentation and labelling as the particular phase of linguistic annotations of our interest, our study performs an automated annotation for the Nanai materials collected through the fieldwork by the REAL project team in 2005 and 2006.

Using the published IPA-based transcriptions of the lexical items (Ko & Yurn 2011), we ran Prosodylab-aligner (Gorman et al. 2011) for the automatic

segmentation and labelling of the Nanai recordings. After this post-processing, the acoustic characteristics of vowels were automatically measured using scripts written for Praat (Boersma & Weenink 2014). The analysis have a particular focus on a vowel feature called "retracted tongue root", a shared property of the vowel systems in many Altaic languages. Therefore, the result will further allow us to understand the historical-comparative dimension of this vowel feature in Altaic languages.

2. The Nanai language materials

The Nanai language, a sister language to Manchu, is a severely endangered southeastern Tungusic language. It is spoken in Russian Far East and China's Heilongjiang province only by approximately 1,400 speakers among 12,200 ethnic Nanais (Lewis et al. 2013). Nanai has three dialects: Upper, Middle, and Lower Amur. The variety to be investigated in this project is the Najkhin dialect of Middle Amur. The consonant and vowel systems of Nanai are as follows:

	Labio-labial	Lamino-alveolar	Dorso-palatal	Dorso-velar
Plosive	b p	d t	ʄ c	g k
Fricative		s		x
Nasal	m	n		ŋ
Trill		r		
Approximant	w	l	j	

Table 1. The consonant system of Nanai (Ko & Yurn 2011:9)

Features	− Back	+ Back	
		− Rounded	+ Rounded
− RTR	i	ə	u
+ RTR	ɪ	a	o

Table 2. The vowel system of Nanai (Ko & Yurn 2011:20)

As described in detail in Ko & Yurn (2001), the language materials of Najkhin Nanai were collected from one Nanai native speaker (= the late Professor Kile Antonina Sergeyevna) in the fieldwork conducted in Khabarovsk, Russia, in October 19-26, 2005 and February 10, 2006, both by the Altaic Society of Korea (Principal Investigator: Prof. Juwon Kim at Seoul National University).[1] The recorded audio and video files amount up to 18 hours (about 11.80 GB)

[1] We are grateful to Professors Juwon Kim and Dongho Ko for allowing us to use the Nanai data and transcriptions.

and include more than 2,500 vocabulary items, 344 everyday phrases and sentences for colloquial uses, and 380 sentences for grammar analysis. The data were elicited in the following way: Once given a Russian word in the questionnaire, the Nanai speaker said its Nanai equivalent twice, as illustrated below:

Investigator: луна? ('moon')
Nanai speaker: bja, bja.

We created a spreadsheet file for the lexical items (or phrases) on Microsoft Excel, using Ko & Yurn's (2011) broad transcriptions. Then we extracted individual audio files corresponding to the lexical items on the spreadsheet from the original long sound files and assigned each audio file the file name corresponding to the numbering of the lexical item on the sheet. Some of the audio files were trimmed if they contained some mismatching elements such as interjections, infrequent cases of stuttering, or self-corrections. Untrimmable audio files as well as those for Russian loans were excluded from automatic annotation and measurement. The remaining 2,434 files (2 hours and 15 minutes in total) were annotated automatically as described below.

3. Automatic Annotation

3.1. Speech recognition tool

We used the Hidden Markov Model Toolkit (HTK; Young 1994) with help of the Prosodylab-aligner (Gorman et al. 2011) for speech recognition and forced alignment. Prosodylab-aligner provides an end-user interface that encapsulates the low-level technical details of HTK and enables completely automated annotation without requiring any manual annotation as training material to initialize the automatic procedure. Also, the mechanism of the system is independent of language-specific features, thus it is applicable to any language in the world (see Yun et al. 2012 for the application to Korean speech data). Yet, it requires a certain degree of computer literacy such as familiarity with unix-like systems to install and use. The choice of operating system also matters; it is supposed to work on any kind of platforms, but it is much easier to install and use on Mac OS X than Windows.

Once the set of speech recognition programs is installed, automatic annotation can be performed with the following three components: i) audio data to be annotated, ii) transcripts of the data, and iii) a pronunciation dictionary of the target language. The audio data should be in the Waveform Audio File Format (.wav), and at least one hour of data is necessary for model training. Transcripts should be ASCII-only text files containing the word-level transcription of the sound files. Each sound file and its corresponding transcript

file should have the same file name except for their extension. The dictionary should be an ASCII-only text file, which contains all the words in the data and their pronunciation. Alternatively, if the transcripts are already written in the phonetic transcription, as in our data, a pronunciation dictionary is not necessary. In this case, one can construct a dictionary file that contains the list of all the phonemes in the data and its isomorphic mapping, in order to vacuously fulfill the technical requirement to run the speech recognition program.

3.2. Preprocessing

In order to conduct annotation through Prosodylab-aligner, all file names and transcripts should be encoded using ASCII only. Table 3 presents the list of phonemes whose symbol differs between the original transcript and the transcript for automatic alignment.

IPA	ASCII
ə	E
ɪ	I
ʝ	z
ŋ	ng

Table 3. Deviation from the original transcription

The original transcript of the data was provided as a single spreadsheet file in which the transcript of each file was given in separate lines. We wrote and ran Python scripts to convert the original transcript into ASCII characters and create separate label files for each audio file.

3.3. Automatic Alignment

After preprocessing was finished and all materials were ready, we ran the aligner to annotate the audio data automatically. Figure 1 shows an example of automatic annotation for the word /xosıkta/ ‘star’. “Sil (silence)” and “sp (small pause)” are pre-defined labels in the aligner.

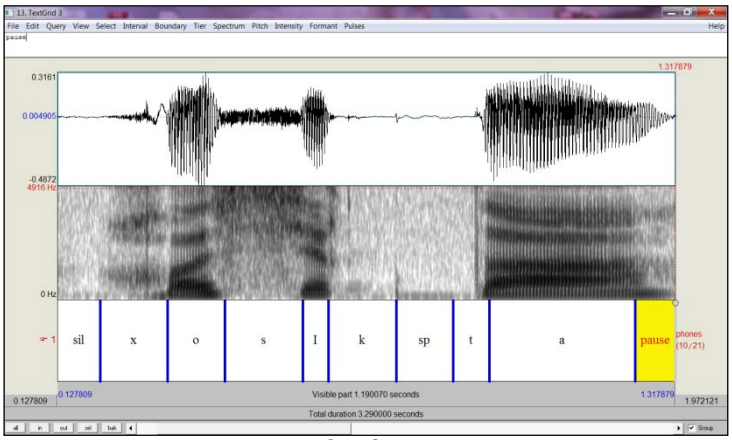


Figure 1. An example of automatic annotation

As shown in Figure 1, the result of automatic annotation was quite successful in most cases. However, for the files that include the nasal sound /n/ and the trill sound /r/, automatic annotation deviated significantly from the expected result. Thus we conducted automatic alignment with several different settings to find the best alignment.

The result of automatic annotation at the first attempt was particularly poor after the pause between repetitions. For example, as shown in Figure 2a, the second repetition of the word /namo/ 'sea' after the pause was incorrectly labeled as silence. It seems that a long pause in the middle of the sound file could cause a problem for the aligner. To avoid this problem, we added the label "pause" to the transcripts in order to explicitly indicate the location of a long pause between repetitions. Figure 2b indicates that the target word was correctly labeled in the second trial. However, the alignment of each label was not precise, especially in the case of the nasal sound /n/. It seems because the boundaries around the nasal sound were often not quite clear due to the /n/-deletion that deletes the underlying /n/ after it triggers vowel nasalization of the preceding vowel.[2]

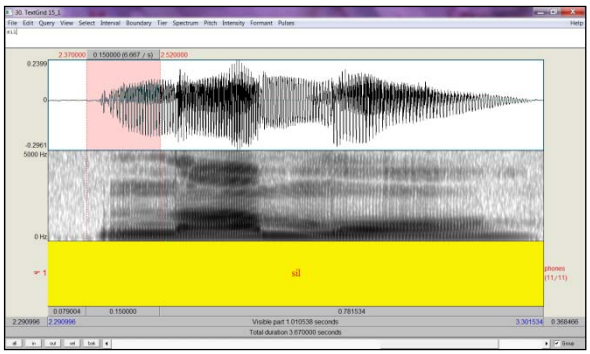
The unclear boundaries between /n/ and vowels cause problems for the aligner, yielding the blurred alignment of nasals and vowels in general, even in the cases where boundaries are rather clear as in Figure 2. Thus, we created additional entries to the pronunciation dictionary in order to indicate nasalized vowels, and converted what was transcribed in the original data as a combination of a vowel and the nasal sound /n/ at the word-final position into nasalized vowels. Figure 2c shows the result of the third trial after vowel na-

[2] 475 lexical item/phrase entries were considered to end with /n/ in their underlying representations by Ko & Yurn (2011). Among these, /n/-deletion took place in both repetitions of 321 entries and only one repetition of 25 entries (667 out of 950 tokens = 70.21%). In the other 129 entries, /n/ was fully realized as [n].

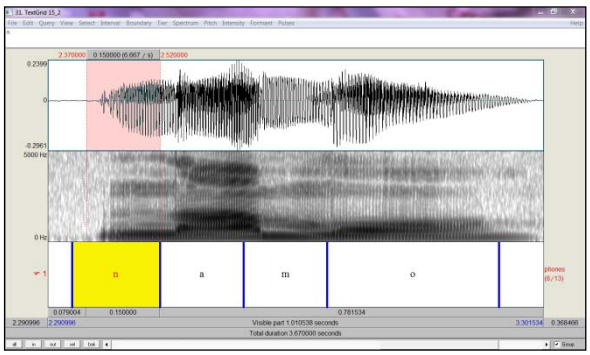
salization was reflected on the transcript. The alignment became better than the second trial in that the boundaries of the interval for the sound /n/ became closer to the actual boundaries. The alignment was still not perfect, however, as the starting point of each interval was estimated earlier than the actual case. In fact, 'perfect alignment' seems impossible because of the inconsistent pattern of vowel nasalization in the data. We leave it to future research to find a way for more precise alignment for the files that include nasal sounds.

The automatic annotation task was performed on Mac OS X 10.10 (Yosemite). The execution time was measured using the *time* command in Linux. Each trial took less than three minutes to annotate the whole data set. Note that the repeated trials did not affect the execution time because each trial was executed independently. For the final trial, it took 2m 40s to annotate the entire audio data of 2 hours and 15 minutes.

a.



b.



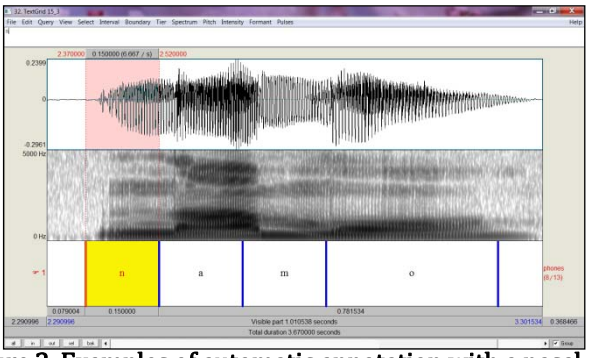
c. 

Figure 2. Examples of automatic annotation with a nasal

4. Phonetic Analysis

The segmented sound files were subject to the measurements of acoustic values, which were conducted by means of a script written for Praat. The measured acoustic cues were fundamental frequency (F0), the first three formants (F1, F2, and F3), and spectral tilts (H1-H2, H1-A1, H1-A2, and H1-A3). Meanwhile, 299 files of 2,434 (12.3%) were excluded because the Praat script failed to measure some of their values due to technical issues. From the 2,135 files, 14,691 vowels went through the measurements. By inspecting F1, 750 vowels were excluded from the data as their F1's were abnormal. At the end, 13,941 vowels were subject to an analysis of variance (ANOVA) to see which acoustic cues play a crucial role in distinguishing the phonemic vowels. The first three formants of short vowels are presented in Table 4, with the results of Tukey's post-hoc tests.

vowel	token	F1		F2		F3	
i	2536	335 (83)	$p<.001$	2352 (360)	$p<.001$	3106 (336)	$p<.001$
ɪ	692	429 (66)		2251 (232)		3017 (310)	
u	1757	390 (46)	$p<.001$	1006 (343)	$p<.001$	2701 (229)	$p<.001$
o	2304	516 (64)		1080 (245)		2630 (234)	
ə	2445	507 (73)	$p<.001$	1400 (298)	$p<.001$	2774 (213)	$p<.001$
a	3179	706 (113)		1490 (220)		2812 (292)	

Table 4. The first three formants of short vowels (means and standard deviations in parentheses) and the results of Tukey's post-hoc tests on each pair of [–RTR] and [+RTR] vowels

As expected, all the first three formants can effectively distinguish a [-RTR] vowel from its [+RTR] counterpart. Among the three formants, F1 seems to be the most reliable acoustic cue since it is consistently lower in [-RTR] vowels than in [+RTR] ones. The other formants do not show this consistency even though they show significant differences. The three formants were compared between short and long vowels. With a few exceptions, short and long vowels were not different from each other in terms of the three formants, which suggests that they are differentiated purely based on the length of vowels.

Spectral tilts were also measured, as they have been mentioned as plausible acoustic cues for the [RTR] feature (see Kang & Ko 2012 for this). With one exception, the four types of spectral tilts show significant differences between [-RTR] and [+RTR] vowels. However, H1-H2 does not show consistency in that [-RTR] vowels have higher values in two pairs but lower value in the other. In contrast, the other values are somewhat reliable. In every pair, all the values are higher in [-RTR] vowels than in [+RTR] ones. These results are in agreement with the previous studies, suggesting that [-RTR] vowels are realized as relatively breathy voice, while [+RTR] vowels are close to creaky voice.

vowel	H1-H2		H1-A1		H1-A2		H1-A3	
i	6.74 (12.96)	$p<.001$	6.33 (10.84)	$p=.05$	29.7 (10.84)	$p<.001$	33.3 (10.87)	$p<.001$
ɪ	1.14 (7.16)		4.78 (12.62)		22.5 (7.27)		27.7 (7.13)	
u	1.31 (6.65)	$p<.001$	8.06 (14.88)	$p<.001$	18.2 (9.33)	$p<.001$	41.5 (7.48)	$p<.001$
o	0.85 (5.09)		1.58 (9.63)		9.33 (7.74)		32.6 (7.14)	
ə	0.30 (6.25)	$p<.001$	2.60 (9.79)	$p<.001$	14.3 (8.01)	$p<.001$	29.2 (7.64)	$p<.001$
a	1.76 (8.22)		0.65 (8.41)		7.54 (8.85)		24.2 (7.82)	

Table 5. The spectral tilts of short vowels (means and standard deviations in parentheses) and the results of Tukey's post-hoc tests on each pair of [-RTR] and [+RTR] vowels

The last question is about the neutral vowel /i/, which appears both in [-RTR] and [+RTR] words. Benus and Gafos (2007) show that the neutral vowel /i/ in Hungarian is not actually neutral at the phonetic or articulatory level. Hungarian has a palatal vowel harmony, by which a word is required to have only front or back vowels. However, the high front vowel /i/ can appear in both contexts. By an acoustic analysis, Benus and Gafos show that the high front vowel /i/ is phonetically different depending on the contexts it belongs to. It is fronter when it is among front vowels than when among back vowels. It is

noteworthy that the difference is bigger than that by normal vowel-to-vowel coarticulations. Similarly, Gick et al. (2006) analyze the low vowel /a/ in Kinande, which has tong root harmony. They show the same results that the vowel, which is known to be neutral in the vowel harmony process, actually participates in the harmony at the phonetic level.

On the basis of the previous studies, we classified the high front vowel /i/ in Nanai into three types: /i/ in [-RTR] context, /i/ in [+RTR] context, and /i/ by itself (e.g., *ənduri* 'god' vs. *gaori* 'to buy' vs. *mi* 'I'), whose formants and spectral tilts were compared. If /i/ is not neutral to but various by the harmony, we would expect that the vowels in different contexts will give birth to different formants and spectral tilts as the phonemic vowels do.

context	token	F1		F2		F3	
[−RTR]	1253	330 (76)	$p<.05$	2314 (369)	$p<.001$	3039 (327)	$p<.001$
[+RTR]	1132	340 (93)	$p=.799$	2380 (341)	$p=.217$	3159 (336)	$p<.01$
By itself	151	347 (65)	$p=.082$	2441 (388)	$p<.001$	3268 (280)	$p<.001$
[−RTR]	1253	330 (76)		2314 (369)		3039 (327)	

Table 6. The first three formants of /i/ in the contexts of [-RTR], [+RTR], and by itself (means and standard deviations in parentheses) and the results of Tukey's post-hoc tests on each pair

Table 6 above shows that the vowels in [-RTR] and [+RTR] contexts are significantly differentiated in terms of all the three formants, while the vowel by itself is not very different from the other two. However, these differences are not observed in the comparison of spectral tilts. As shown in Table 7, spectral tilt values cannot distinguish the vowel in different contexts.

vowel	H1-H2		H1-A1		H1-A2		H1-A3	
[−RTR]	7.49 (12.58)	$p=.074$	6.83 (14.25)	$p=.389$	29.2 (10.38)	$p=.068$	32.9 (10.41)	$p=.396$
[+RTR]	6.23 (13.34)	$p=.337$	5.84 (13.18)	$p=1$	30.3 (11.35)	$p=.789$	33.7 (11.39)	$p=1$
By itself	8.40 (12.13)	$p<.05$	5.93 (14.93)	$p=.939$	29.26 (10.47)	$p=1$	33.5 (10.55)	$p=.959$
[−RTR]	7.49 (12.58)		6.83 (14.25)		29.2 (10.38)		32.9 (10.41)	

Table 7. The spectral tilts of /i/ in the contexts of [-RTR], [+RTR], and by itself (means and standard deviations in parentheses) and the results of Tukey's post-hoc tests on each pair

All the results imply that in Nanai the high front vowel /i/ does not vary according to the harmony context. Rather, its variation just results from vowel-to-vowel coarticulation, which means that it is realize as lower (and fronter) in the [+RTR] contexts because [+RTR] back vowels (/o/ and /a/) are lower (and fronter) than their [-RTR] counterparts (/u/ and /ə/).

5. Conclusion

We aimed to provide a plausible way to analyze 'big' data, specifically raw sound files. By adopting automated tools for speech annotation and phonetic analysis, we were able to drastically decrease the amount of time and effort for the analysis of 2,434 sound files. Additionally, the acoustic analysis of vowels showed that, by and large, the automated process can give birth to relatively reliable results. Though the effectiveness was evidenced in terms of time and effort, it was also shown that we still need to improve the accuracy. Possibly, we will have to adjust the automatic annotation system on the basis of the phonetic characteristics of a language, as in the case of nasalized vowels in Nanai. Another thing to note is that about 16.8% of data were lost due to technical issues, which should be inspected thoroughly. Despite these remaining technical issues, we believe that Altaic linguistics will be highly benefited from the automated post-transcriptional processing techniques described in this paper.

References

Benus, Stefan & Adamantios Gafos. 2007. Articulatory characteristics of Hungarian 'transparent' vowels. *Journal of Phonetics* 35. 271-300.

Boersma, Paul & David Weenink. 2014. Praat: doing phonetics by computer [Computer program]. Version 5.4.03, retrieved 18 December 2014 from http://www.praat.org/

Gick, Bryan, Douglas Pulleyblank, Fiona Campbell & Ngessimo Mutaka. 2006. Low vowels and transparency in Kinande vowel harmony. *Phonology* 23. 1-20.

Gorman, Kyle, Jonathan Howell, & Michael Wagner. 2011. Prosodylab-aligner: A tool for forced alignment of laboratory speech. Canadian Acoustics 39.3. 192–193. http://prosodylab.org/tools/aligner/

Kang, Hijo & Seongyeon Ko. 2012. In search of the acoustic correlates of tongue root contrast in three Altaic languages: Western Buriat, Tsongol Buriat, and Ewen. *Altai hakpo* 22. 179-203.

Kim, Juwon, et al. 2008. *Salacye kanun althaienelul chacase* [Documentation of endangered Altaic languages] (2nd ed.). Phacwu, Korea: Thaehaksa.

Kim, Juwon, et al. 2011. *Ene tayangseng poconul wihan altai ene mwunsehwa* [Documentation of Altaic languages for the maintenance of language diversity] (2nd ed.). Phacwu, Korea: Thaehaksa.

Ko, Dongho & Gyudong Yurn. 2011. *A Description of Najkhin Nanai*. Seoul National University Press.

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). 2013. *Ethnologue: Languages of the World*, Seventeenth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Newman, Paul. 2013. The law of unintended consequences: How the endangered languages movement undermines field linguistics as a scientific enterprise. The Linguistics Departmental Speaker Series. Lecture delivered at SOAS, University of London, UK, October 15th, 2013. http://www.hrelp.org/events /seminars/paulnewman/index.html

Young, Steve J. 1994. *The HTK hidden Markov model toolkit: Design and philosophy*: Cambridge University Engineering Department.

Yun, Jiwon, Hyun Kyung Hwang, & Seongyeon Ko. 2012. Automatic Annotation for Korean Speech Corpus Analysis. Poster presented at the International Workshop on Corpus Linguistics and Endangered Dialects. National Institute for Japanese Language and Linguistics, Tokyo, Japan.

ABSTRACT

Recent years have witnessed a remarkable growth in Altaic field linguistics, with recognizable achievements especially in language description and archiving. However, relatively less effort has been made to analyze the collected materials from the general phonetic perspectives. In this paper, the authors demonstrate how the archived language data can be analyzed using automated tools for speech annotation and phonetic analysis and thus with drastically reduced time and effort. Using the published IPA-based transcriptions of the lexical items (Ko & Yurn 2011), we ran Prosodylab-aligner (Gorman et al. 2011) for the automatic segmentation and labelling of the Nanai recordings collected by the Altaic Society of Korea. After this post-processing, the acoustic characteristics of vowels were automatically measured using scripts written for Praat (Boersma & Weenink 2014). The acoustic analysis of Nanai vowels shows that, by and large, the automated process produces relatively reliable results with its effectiveness evidenced in terms of time and effort. We believe that Altaic linguistics will be highly benefited from these practical techniques with the improvement of accuracy.