Altangerel Ayush
Munkhnasan Choinzon

Report on Comparative Evaluation of Two Toolkits for Mongolian Speech

1. Introduction

Automatic speech recognition(ASR) is one branch of the field of speech processing and related with a number of different fields of knowledge such as acoustic, linguistics, pattern recognition, and artificial intelligence. The complexity of an ASR system depends on its limitations, such as speaker dependence or independence, continuous or isolated speech, large, medium or small vocabulary.

There are many speech recognition engines for speech recognition.
All Speech Recognition Engines (SRE)are made up of the following components:

- Language Model or Grammar

  Language Models contain a very large list of words and their probability of occurrence in a given sequence. Grammars are a much smaller file containing sets of predefined combinations of words. Grammars are used in **interactive voice response** or desktop command and control applications. Each word in a language model or grammar has an associated list of phonemes.
- **Acoustic Model** is a file that contains a statistical representation of each distinct sound that makes up a spoken word. It must contain the sounds for each word used in your grammar. The words in the grammar give the SRE the sequence of sounds it must listen for. The SRE then listens for the sequence of sounds that make up a particular word, and when it finds a particular sequence, returns the textual representation of the word The Grammar and the Acoustic Model work together.

- **Decoder** - Software program that takes the sounds spoken by a user and searches the Acoustic Model for the equivalent sounds. When a match is made, the Decoder determines the phoneme corresponding to the sound. It keeps track of the matching phonemes until it reaches a pause in the users speech. It then searches the Language Model or Grammar file for the equivalent series of phonemes.

The purpose of our work is to develop Mongolian speech recognition system using Hidden Markov Model. We selected two toolkits, HTK and CMU Sphinx 4. These toolkits are both HMM based and support Windows OS and linux, and have modular design that is easily adaptable.

In this report, we present the results of comparative evaluation of two toolkits for Mongolian Speech.

Following section introduces brief phonetics of Mongolian language and data preparations on each toolkits. In the section 3, comparative evaluation of two toolkits and its selection procedure are described. Finally, conclusions and future works are given in the last section.

2 Data Preparation

The first stage of any recognizer development project is data preparation. Speech data is needed both for training and for testing. In the system to be built here, all of this speech will be recorded from scratch.

Before the data can be recorded, we have to define a phoneme set, and have to construct a dictionary to cover both training and testing, and have to create a task grammar.

2.1 Mongolian Phoneme

To train HMM, phone labeling is necessary. In labeling, we use the Mongolian phoneme set[2,3,4,8,9,10] as can be seen in Table 1.

Phonetic category	Phoneme	Word	phoneme sequence
Vowels			
Short vowels	/a/	ah	**a** h
	/o/	nom	n **o** m
	/u/	uls	**u** l s
	/e/	ter	t **e** r
	/oe/	oed	**oe** d
	/ue/	nued	n **ue** d
	/i/	shig	sh **i** g
Long vowels	/aa/	taar	t **aa** r
	/oo/	oosor	**oo** s o r
	/uu/	uul	**uu** l
	/ee/	eemeg	**ee** m e g
	/oeoe/	hoeoer	h **oeoe** r
	/ueue/	ueuel	**ueue** l
	/ii/	tiim	t **ii** m
Diphtongs	/ai/	ail	**ai** l
	/oi/	oims	**oi** ms
	/ui/	shugui	s u g **ui**
	/uei/	ueguei	ue g **uei**
	/ei/	suertei	s ue r t **ei**
y-vowels	/ya/	yavah	**ya** v a h
	/yo/	yostoi	**yo** s t oi
	/ye/	yeven	**ye** v e n
	/yu/	yum	**yu** m
Consonants			
Plosives	/b/	bagsh	**b** a g sh

	/p/	puujin	**p** uu j i n
	/d/	devter	**d** e v t e r
	/t/	tueuenii	**t** ueue n ii
	/k/	kino	**k** i n o
	/g/	arag	a r **g**
	/G/	arga	a r **G**
Fricatives	/f/	faz	**f** a z
	/v/	ve	**v** e
	/s/	sandal	**s** a n d a l
	/sh/	shuurga	**sh** uu r G
	/ch/	changa	**ch** a n G
	/z/	zasag	**z** a s a g
Trill	/r/	radio	**r** a d i o
Lateral approximant	/l/	lavlah	**l** a v l a h
Nasals	/m/	mongol	**m** o n g o l
	/n/	hana	h a **n**
	/ng/	tung	t u **ng**
	/N/	han	h a **N**

Table 1. Mongolian phoneme set

2.2 Data Preparation on HTK

2.2.1 Task Grammar

We created our task grammar using grammar definition language which specifies simple task grammars. The step task grammar is to create a regular grammar and convert it to an intermediate form of decoding network.

For our recognizer, a suitable grammar might be

$words=am|ami|amid|han|hana|haan|haana|hani|arag|arga|sal|saali|ul|uul|ueuel|aw|aaw|taria|shuurga|surguuli|bar|bari|darga|arga|zarlal|zarlah|holtos|ishig|eeme g|odoo|olon|ah|ahiin|honog|songo|yostoi|shine|said|oeoer|ard|mongol|moengoe|h oeroengoe|hawar|huwi|huw|sonin|sono|halh|halah|dawaa|myagmar|lhagwa|pure w|baasan|byamba|nyam|on|jil|sar|saya|oendoer|gishueuen|eh|tom|tiim|medeelel|sain|toer|baga|ueg|bodno|edlel|tagla|udwal|awbal|onts|sansar|duers|manai|sanal|shueueh|dund|zam|gishueued|gol|ueildwer|ilueue|bi|bid|ted|neg|hoyor|guraw|doe roew|taw|zurgaa|doloo|naim|yes;

(SENT-START <$words> SENT-END)

2.2.2 The Dictionary

The next step of the data preparation is to create a pronunciation dictionary covering all words in the grammar.

We built dictionary of 100 words from scratch. These words were collected from newspaper “Onoodor” and television news of Mongolian National TV. We attempted to choose words that cover all phonemes in Mongolian language.

The dictionary is shown in Appendix A.

2.2.3 Recording the Data and Preparing Files

In the above steps, we did all necessary preparation related to data before we shift to recording the training data.

When we are recording speech files for training utterances need not to be in the grammar, but need to cover all phonemes. However, having in-grammar utterances gives better performance.

To record training data, 10 native speakers are selected, 5 male and 5 female. Each speaker is asked to read the prepared text 10 times. A speech training set should be large enough i.e., each phoneme should appear at least 10 times for this trial. Recording time for one speaker was 150 minutes. Speech was recorded in a quite room.

We prepared the files; monophn.mlf which is a transcription file to describe phoneme transcriptions of the training data and monophn.list which shows a list of unique phonemes.

2.3 Data Preparation on Sphinx 4

We need the following files to begin the training on Sphinx 4:

1. A set of **feature files** computed from the audio training data, one each for every recording you have in the training corpus. Each recording can be transformed into a sequence of feature vectors using a front-end executable provided with the SPHIN-III training package. Each front-end executable provided performs a different analysis of the speech signals and computes a different type of feature.
2. The **mongol_train.**fileids file contains the list of feature-set filenames with full paths to them. An example of the entries in this file:

   train_words/a10
   train_words/a11
   train_words/a12
   . . . . . .

3. The **mongol_train.transcription** file in which the transcripts corresponding to the feature files are listed in exactly the same order as the feature filenames in the control file.

   <s> am ami amid han hana haan haana hani arag arga </s> (a10)

   <s> am ami amid han hana haan haana hani arag arga </s> (a11)

   <s> am ami amid han hana haan haana hani arag arga </s> (a12)

4. **mongol.dic** file which has all acoustic events and words in the transcripts mapped onto the acoustic units you want to train. Redundancy in the form of extra words is permitted. Here's an example:

am	***a m***
ami	***A m***
amid	***A m d***
han	***h a N***
hana	***h a n***
haan	***h aa N***
haana	***h aa n***
hani	***h A n***
arag	***a r g***
arga	***a r G***

5. The **mongol.filler** file, which usually lists the non-speech events as "words" and maps them to user_defined phones. This dictionary must at least have the entries like following:

<s>	***SIL***
</s>	***SIL***
<sil>	***SIL***

Note that the words <s>, </s> and <sil> are treated as special words and are required to be present in the filler dictionary. At least one of these must be mapped on to a phone called "SIL". The phone SIL is treated in a special manner and is required to be present.

6. The **mongol.phone** file, which is a list of all acoustic units that we want to train models for. The SPHINX does not permit to have units other(different) than those in your dictionaries. All units in your two dictionaries must be listed here. In other words, your phone list must have exactly the same units used in your dictionaries, no more and no less. Each phone must be listed on a separate line in the file, begining from the left, with no extra spaces after the phone. For an example:

   a

   A

   aa

   AA

   ai

   b

   d

   e

   ee

   g

   G

   (etc.)

Here's a quick checklist to verify your data preparation before you train:

1. Are all the transcript words in the dictionary/filler dictionary?
2. Make sure that the size of transcript matches the .ctl file.
3. Check the boundaries defined in the .ctl file to make sure they exist i.e., you have all the frames that are listed in the control file
4. Verify the phone list against the dictionary and fillerdict.

CI MODEL DEFINITION FILE

mongol.ci.mdef

mongol.1000.mdef

CI MODEL PARAMETERS

CI models consist of 4 parameter files :

- **mixture_weights**: the weights given to every Gaussian in the Gaussian mixture corresponding to a state
- **transition_matrices**: the matrix of state transition probabilities
- **means**: means of all Gaussians
- **variances**: variances of all Gaussians

3. Evaluation

The performance of speech recognition systems is usually specified in terms of accuracy and speed. Accuracy may be measured in terms of performance accuracy which is usually rated with word error rate (WER).
We compared HMM based small vocabulary speech recognizers built using HTK and CMU Sphinx 4 toolkits.

3.1 Testing Data

The recognizers were evaluated on three different sentences each from 5 speakers who didn't attend to prepare the training data(in total 155 words in 15 sentences). The results were shown in Table 2 and Table 3.

To select three sentences for recording, we could manually construct sentences which cover the phonetic features of the language, to get as broad a coverage as possible over the language.

For each iteration, the columns in Table 2 and 3 give the percentage of substitutions, insertions, and deletions, as well as the word accuracy,and the percentage of correct sentences.

Speaker №	**Substitution**	**Insertion**	**deletion**	**Word Acc%**	**Corr%**
Sentence 1 (with 5 isolated words)					
Speaker 1	0	0	0	100.0	100.0
Speaker 2	1	0	0	80.0	80.0
Speaker 3	0	0	0	100.0	100.0
Speaker 4	0	0	0	100.0	100.0
Speaker 5	0	0	0	100.0	100.0
Sentence 2 (with 12 isolated words)					

Speaker 1	1	0	0	91.67	91.67
Speaker 2	0	0	0	100.0	100.0
Speaker 3	0	0	0	100.0	100.0
Speaker 4	0	0	0	100.0	100.0
Speaker 5	0	0	0	100.0	100.0
Sentence 3 (with 14 isolated words)					
Speaker 1	1	0	0	92.86	92.86
Speaker 2	1	0	0	92.86	92.86
Speaker 3	0	0	0	100.0	100.0
Speaker 4	0	0	0	100.0	100.0
Speaker 5	0	1	0	92.86	100.0

Table 2: Recognition accuracy of HTK on unknown speakers.

Best result: 100% word accuracy.

The Sphinx 4 recognizer's results are illustrated as in Table 3.

Speaker №	**Substitution**	**Insertion**	**deletion**	**Word Acc%**	**Corr%**
Sentence 1 (with 5 words)					
Speaker 1	1	0	0	80.00	80.00
Speaker 2	0	1	0	80.00	100.00
Speaker 3	1	1	0	60.00	80.00
Speaker 4	1	0	0	80.00	80.00
Speaker 5	0	1	0	80.00	100.00
Sentence 2 (with 12 words)					
Speaker 1	1	1	0	83.33	91.67
Speaker 2	2	1	0	75.00	83.33

Speaker 3	3	1	0	66.67	75.00
Speaker 4	1	1	0	83.33	91.67
Speaker 5	0	1	0	91.67	100.00
Sentence 3 (with 14 words)					
Speaker 1	1	0	0	92.86	92.86
Speaker 2	0	1	0	92.86	100.00
Speaker 3	1	0	1	85.71	85.71
Speaker 4	1	0	0	92.86	92.86
Speaker 5	1	1	0	85.71	92.86

Table 3: Recognition accuracy of Sphinx 4 on unknown speakers.

3.2 The Results

First, let's see recognition performance in HTK toolkit.

For speakers found in the training data, the worst results obtained were a 95.6% word accuracy**.**
When tested on data from five previously unseen speakers, the recognizer had a 91.*67*% word accuracy(Table 2).

With Sphinx 4, for speakers found in the training data, the worst results obtained were a 87% word accuracy.
When tested on data from five previously unseen speakers, the recognizer had a 83.3% word accuracy(Table 3).

When the same recognizer was tested for speakers who were not included in the training data with three sentences, the recognition rate degraded.

The word accuracy of HTK recognizer for previously unseen speakers was reduced by 3.93%, while for Sphinx 4 recognizer the word accuracy was reduced by 4.7%.

3.3 Selection of toolkit

We compared two HMM-based small vocabulary that speech recognizers built using HTK and CMU Sphinx 4 toolkits. From the compared results, speaker independent recognition performance for two systems comparable with above 90% resp. above 80% word accuracy for HTK resp. CMU Sphinx 4.

The Sphinx 4 recognizer gave a 4.7% decrease in word accuracy, when tested on data from previously unseen speakers, compared to HTK recognizer, which had 3.93% decrease in word accuracy.

Therefore we selected HTK toolkit based on its recognition accuracy.
Another reason why we select HTK is because it can be trained automatically and are simple and computationally feasible to use.

The detail recognition results of HTK on the data from five speakers who did not attend to prepare the training data with three different sentences are shown in Appendix B.

4 Conclusions and Future Works

In this report, we presented experiences with using the two toolkits to build a general isolated-word recognizer for Mongolian language based on the dictionary of 100 words.
The best result of 100% word recognition accuracy was achieved on HTK toolkit.
For the dictionary of 500 words, HTK toolkit was also evaluated.

Fortunately, Mongolian language is not a tonal language. Moreover, there are no words that those have the same pronunciation[2,3,4,8,9,10]. However, the language has no widely accepted computer representation such as speech corpus. Speech corpus is only at the beginning of development for Mongolian.

In this phase, we recorded only isolated words, not sentences. Thus we did not built language model. However, in the next phase to increase the recognition rate we are planning to test on continuous speech. In that case we need to build language model.

A language model is a file containing the probabilities of sequences of words.

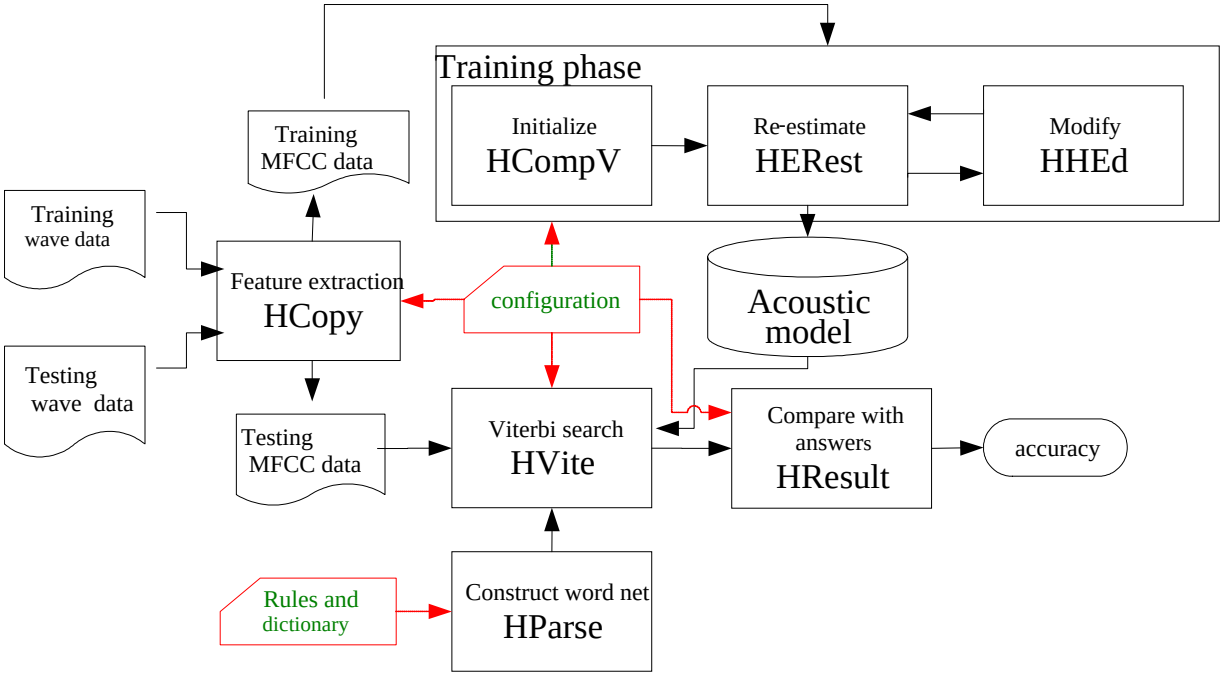
The Mongolian transcription system which we used needs more work in order to generate high quality transcriptions. From inspection of its current output it seems that vowel quantity is one particular area which needs more work.

References

1. Lawrence Rabiner, Biing-Hwang Juang, “Fundamental of Speech Recognition”, Pearson Education, 2005.
2. Ц. Дамдинсүрэн, “Монгол үсгийн дүрмийн толь” , 1983”.
3. Ц. Дамдинсүрэн, Я. Цэвэл, “Үсгийн дүрмийн зөв бичих толь” , 1951”.
4. А. Лувсандэндэв, “Монгол орос толь” , 1957”.
5. Mark Aronoff, Kirsten Fudeman, “What is Morphology”, Blackwell Publishing, 2005.
6. Peter Ladefoged, “A Course in Phonetics”, Thomson Learning, 2001
7. Stephen Haag, Maeve Cummings, James Dawkins, “Management Information Systems for the Information Age”, McGraw-Hill, 1998.
8. Я. Цэвэл, “Монгол хэлний товч тайлбар толь” , 1966”.
9. Х. Далхжав, Ц. Цэрэнчимэд, “Зөв бичих зүйн толь бичиг” , 1974”.
10. С. Шагж, “ Монгол үсгийн дүрмийн толь бичиг”
11. Xueoong Huang, Alex Acero, and Hsiao-Wuen Hon, “Spoken Language Processing”, Prentice-Hall, 2001.

Appendix A

HTK recognizer flowchart



Dictionary

am	a m sp
ami	A m sp
amid	A m d sp
han	h a N sp
hana	h a n sp
haan	h aa N sp
haana	h aa n sp
hani	h A n sp
arag	a r g sp
arga	a r G sp
sal	s a l sp
saali	s AA l sp
ul	u l sp

uul uu l sp
aw a w sp
aaw aa w sp
ueuel ueue l sp
taria t a r ia sp
shuurga sh uu r G sp
surguuli s u r g uu l sp
bar b a r sp
bari b A r sp
darga d a r G sp
arga a r G sp
zarlal z a r l a l sp
zarlah z a r l a h sp
holtos h o l t o s sp
ishig i sh i g sp
eemeg ee m e g sp
odoo o d oo sp
olon o l o n sp
ah a h sp
ahiin a h ii n sp
honog h o n o g sp
songo s o n G sp
yostoi yo s t oi sp
shine sh i n sp
said s ai d sp
oeoer oeoe r sp
ard a r d sp
mongol m o n g o l sp
moengoe m oe n G sp
hoeroengoe h oe r oe n G sp
hawar h a w a r sp

huwi	h U w sp
huw	h u w sp
sonin	s o n i n sp
sono	s o n sp
halh	h a l h sp
halah	h a l a h sp
dawaa	d a w aa sp
myagmar	m ya g m a r sp
lhagwa	lh a g w sp
purew	p u r e w sp
baasan	b aa s a n sp
byamba	b ya m b a sp
nyam	n ya m sp
on	o n sp
jil	j i l sp
sar	s a r sp
saya	s a ya sp
oendoer	oe n d oe r sp
gishueuen	g i sh ueue n sp
eh	e h sp
tom	t o m sp
tiim	t ii m sp
medeelel	m e d ee l e l sp
sain	s ai n sp
toer	t oe r sp
baga	b a G sp
ueg	ue g sp
bodno	b o d n o sp
edlel	e d l e l sp
tagla	t a g a l sp
udwal	u d w a l sp

awbal a w b a l sp

onts o n ts sp

sansar s a n s a r sp

duers d ue r s sp

manai m a n ai sp

sanal s a n a l sp

shueueh sh ueue h sp

dund d u n d sp

zam z a m sp

gishueued g i sh ueue d sp

gol g o l sp

ueildwer ue i l d w e r sp

ilueue i l ueue sp

bi b i sp

bid b i d sp

ted t e d sp

neg n e g sp

hoyor h o yo r sp

guraw g u r a w sp

doeroew d oe r oe w sp

taw t a w sp

zurgaa z u r g aa sp

doloo d o l oo sp

naim n ai m sp

yes ye s sp

SENT-START [] sil

SENT-END [] sil

List file of words

#!MLF!#

"*/a10.lab"

am

ami
amid
han
hana
haan
haana
hani
arag
arga
.
"*/a20.lab"
sal
saali
ul
uul
ueuel
aw
aaw
taria
shuurga
surguuli
.
"*/a30.lab"
bar
bari
darga
arga
zarlal
zarlah
holtos
ishig

eemeg
odoo
.
"*/a43.lab"
olon
ah
ahiin
honog
songo
yostoi
shine
said
oeoer
ard
.
*/a50.lab"
mongol
moengoe
hoeroengoe
hawar
huwi
huw
sonin
sono
halh
halah
.
"*/a60.lab"
dawaa
myagmar
lhagwa

purew
baasan
byamba
nyam
on
jil
sar
.
"*/a70.lab"
saya
oendoer
gishueuen
eh
tom
tiim
medeelel
sain
toer
baga
.
"*/a80.lab"
ueg
bodno
edlel
tagla
udwal
awbal
onts
sansar
duers
manai

.
"*/a90.lab"
sanal
shueueh
dund
zam
gishueued
gol
ueildwer
ilueue
bi
bid
.
"*/a100.lab"
ted
neg
hoyor
guraw
doeroew
taw
zurgaa
doloo
naim
yes
.

Appendix B

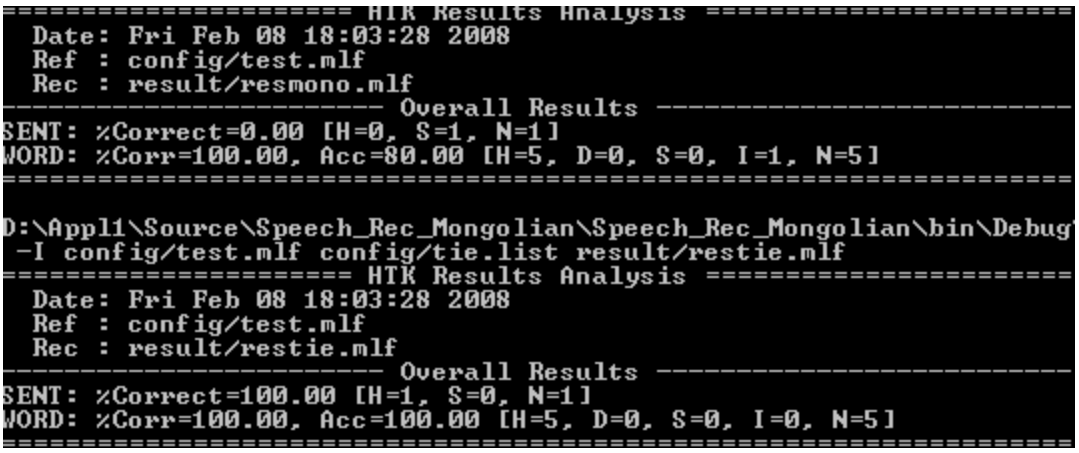


Figure1: Result of speaker 1 on sentence 1.

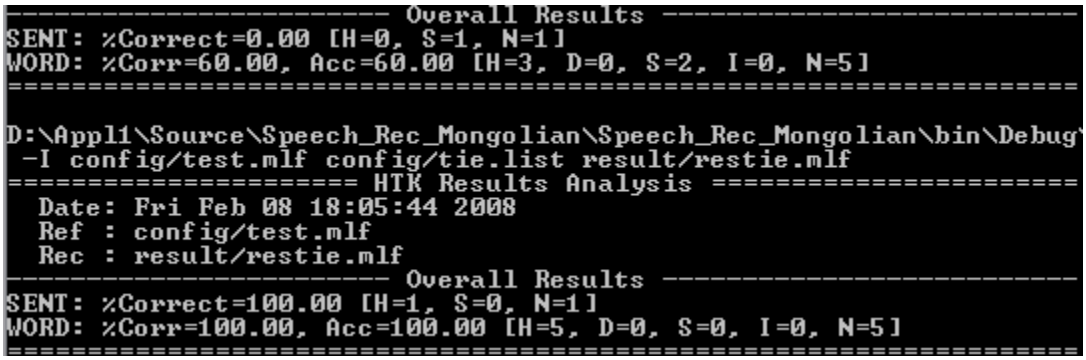


Figure2: Result of speaker 2 on sentence 1.

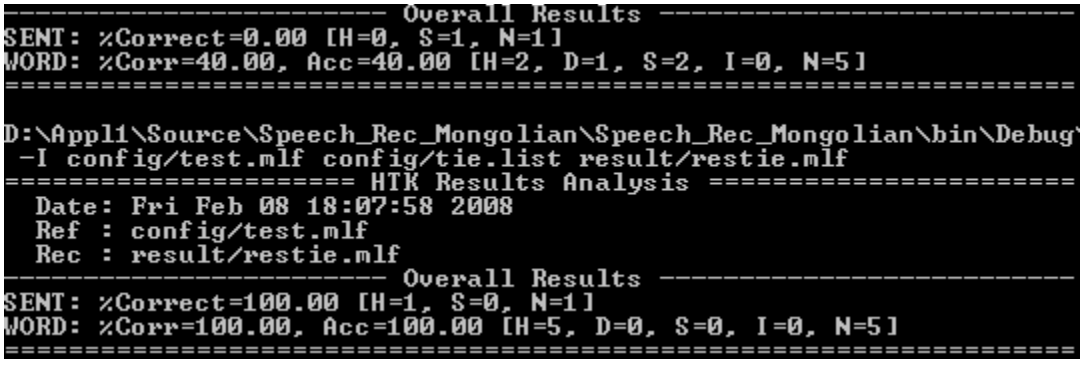


Figure3: Result of speaker 3 on sentence 1.

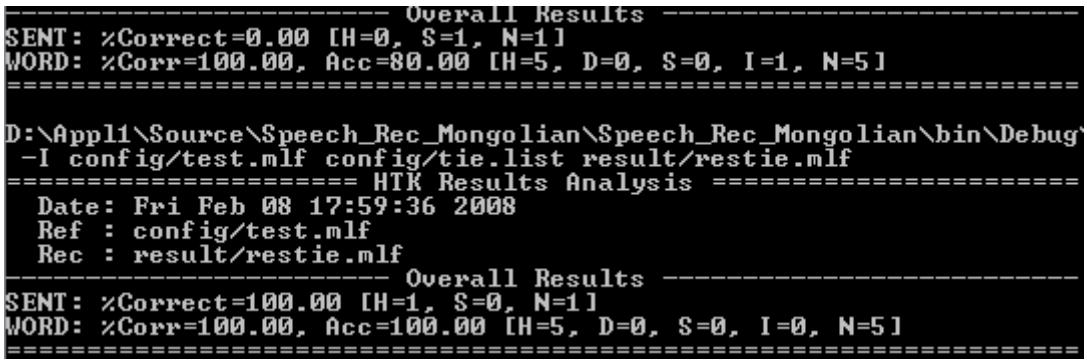


Figure4: Result of speaker 4 on sentence 1.

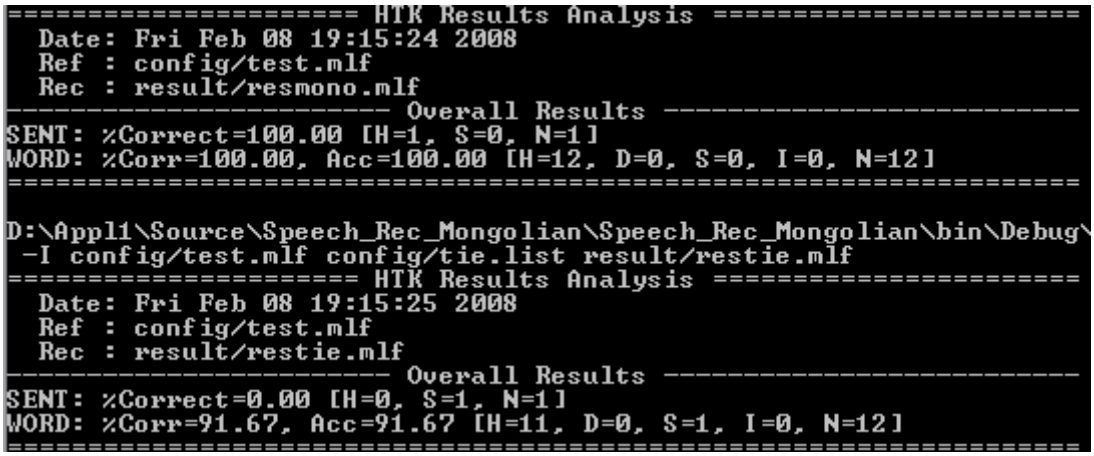


Figure6: Result of speaker 1 on sentence 2.

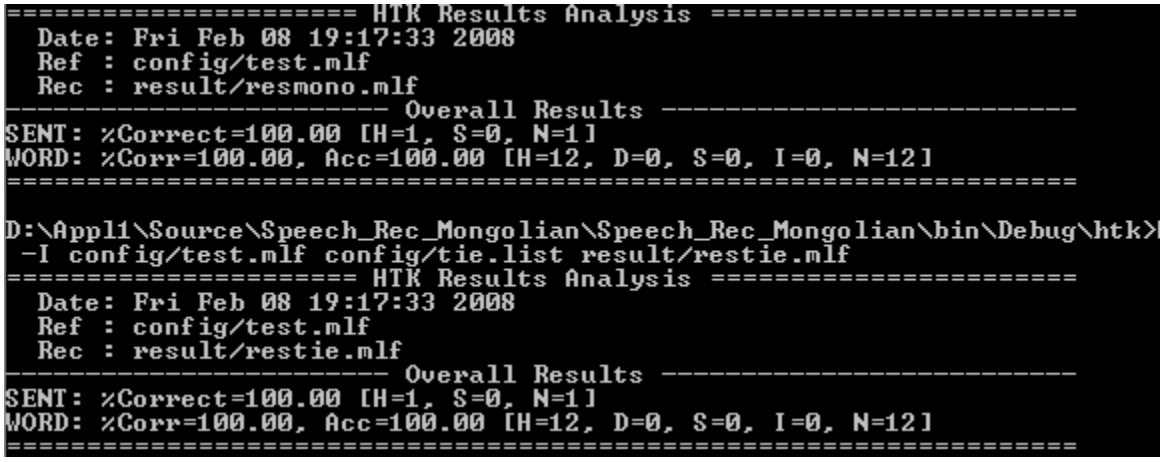


Figure7: Result of speaker 2 on sentence 2.

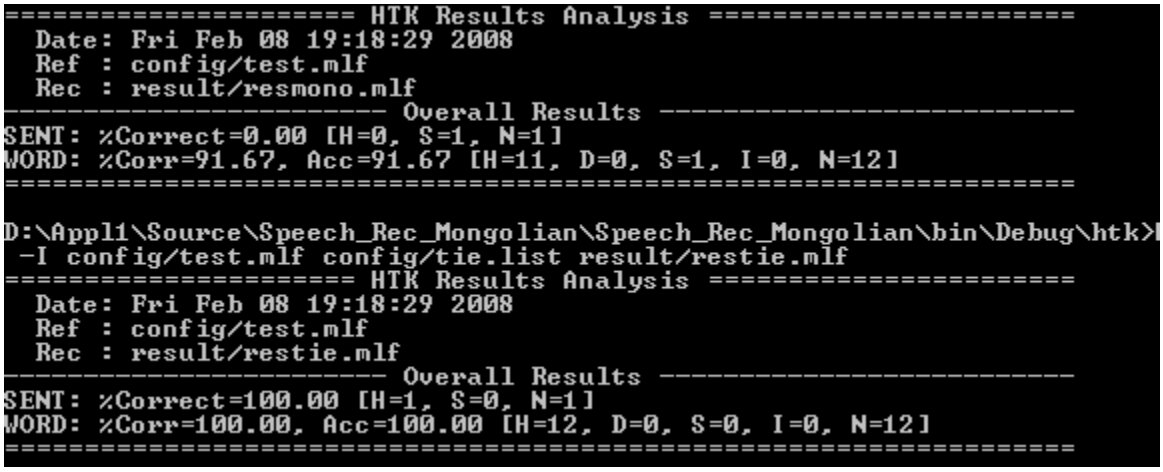


Figure8: Result of speaker 3 on sentence 2.

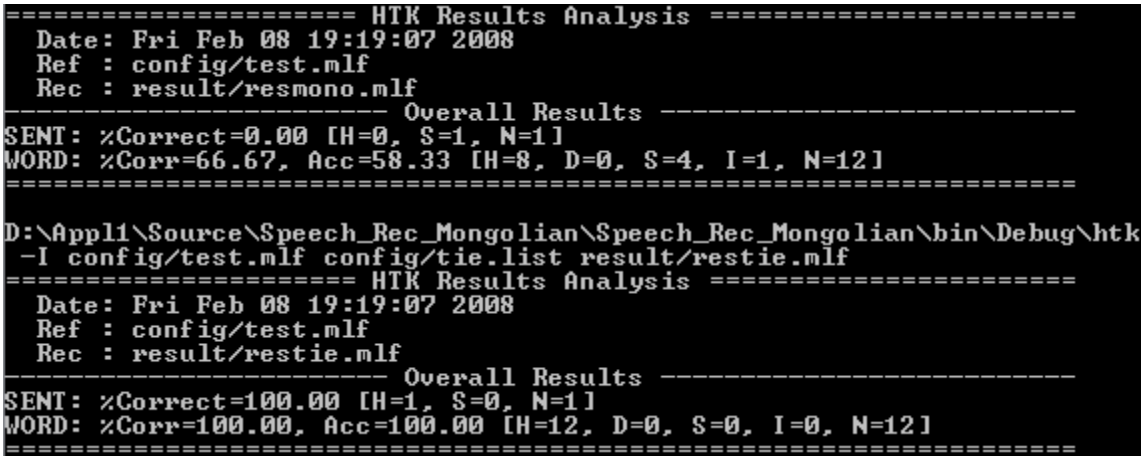


Figure9: Result of speaker 4 on sentence 2.

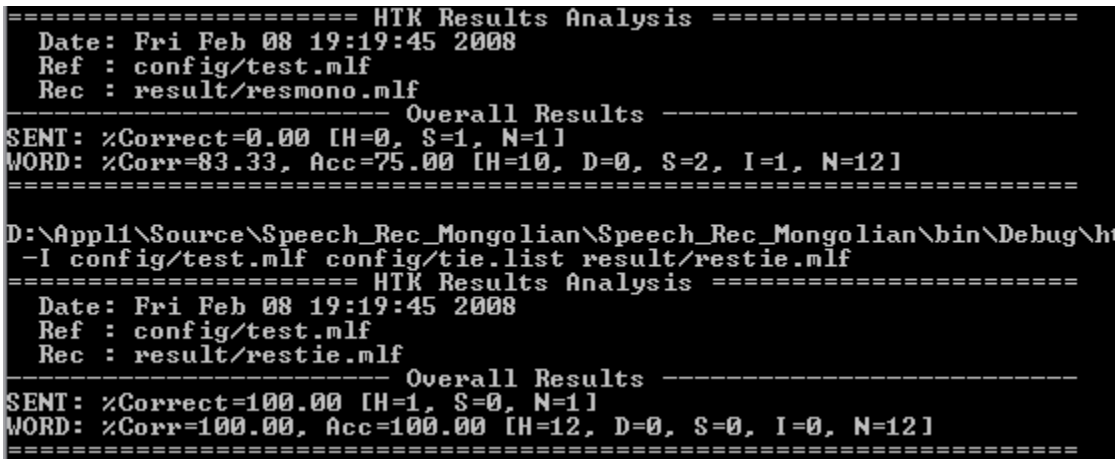


Figure10: Result of speaker 5 on sentence 2.

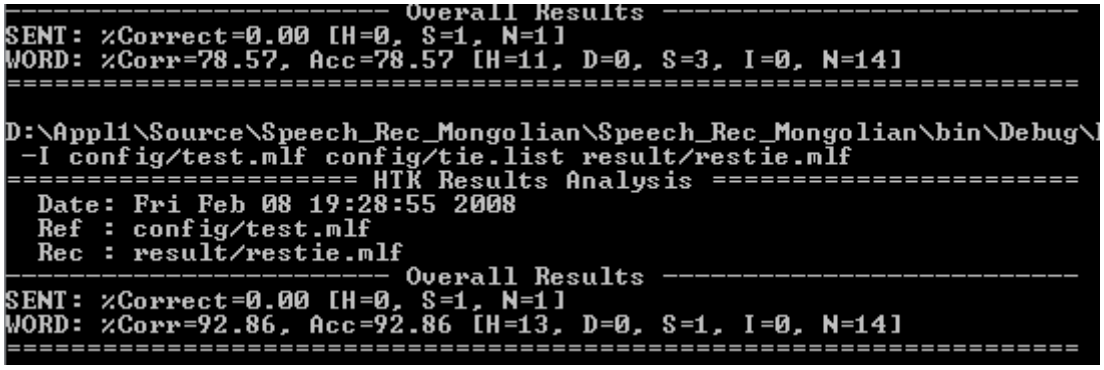


Figure11: Result of speaker 1 on sentence 3.

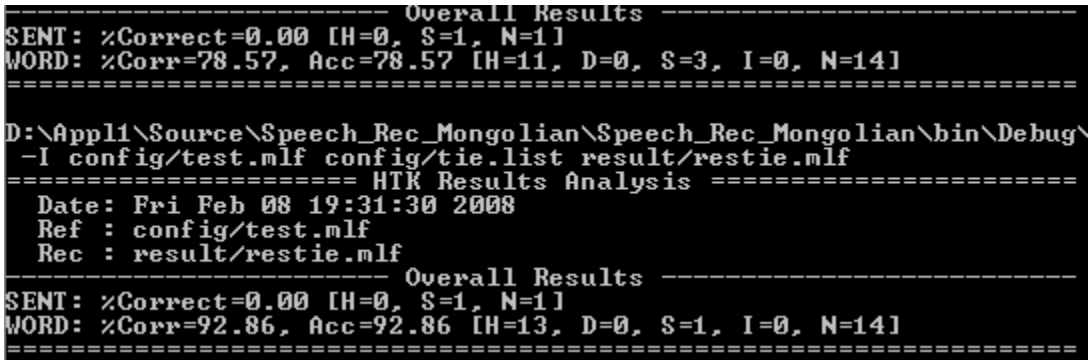


Figure12: Result of speaker 2 on sentence 3.

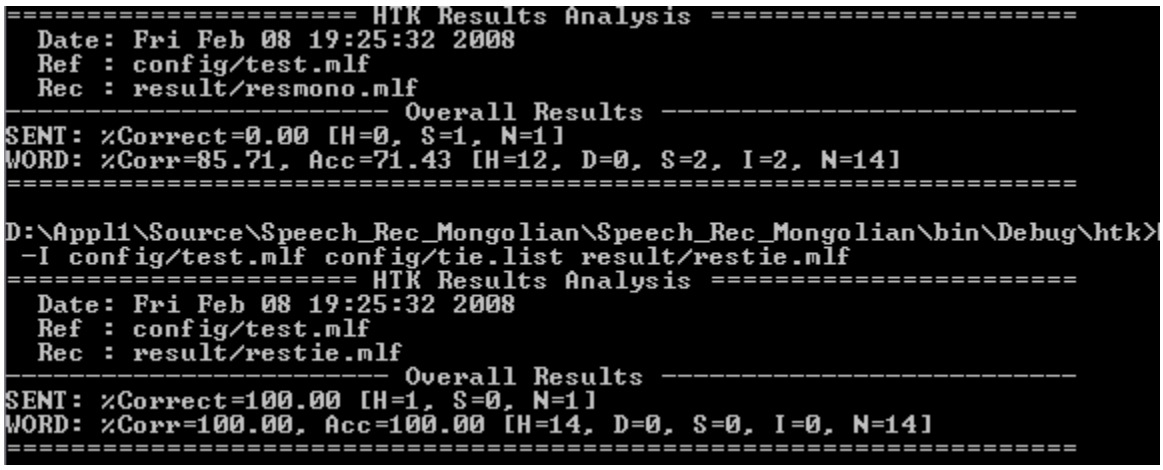


Figure13: Result of speaker 3 on sentence 3.

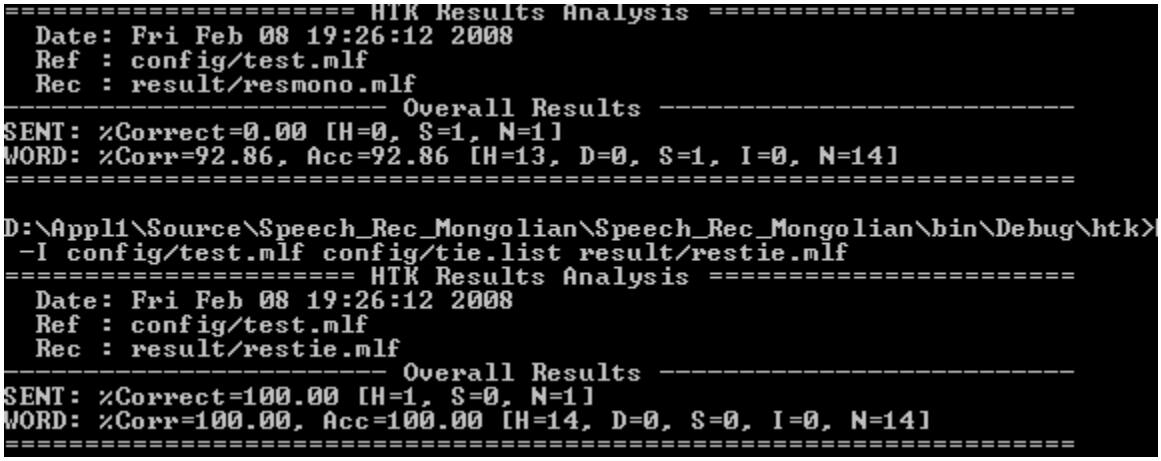


Figure14: Result of speaker 4 on sentence 3.

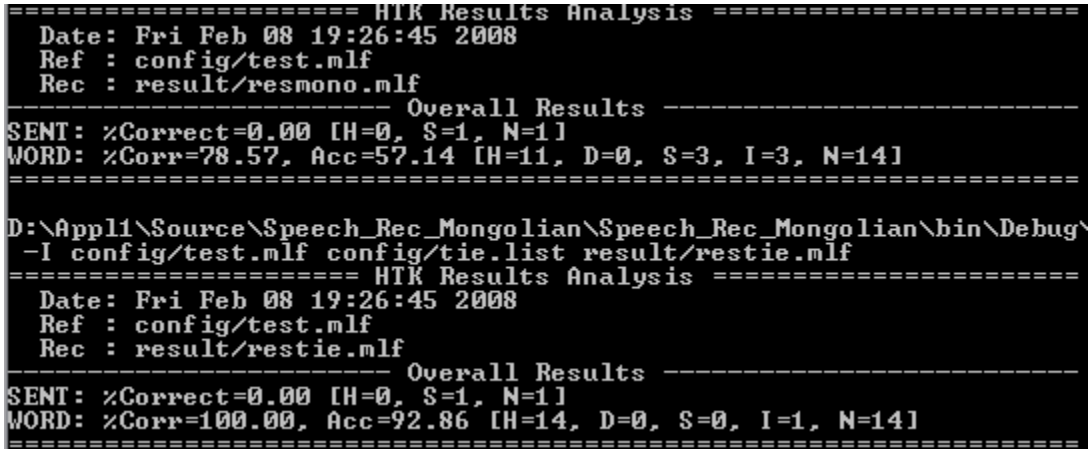


Figure15: Result of speaker 5 on sentence 3.