1.0 Enhancements to MALACH distribution.

1.1 Overview

The existing MALACH distribution is being augmented with an additional set of files to enable licensees to conduct speech recognition experiments. The files and file formats are specifically tailored to be consistent with Kaldi [1] but can be easily adapted for other purposes.

The original MALACH distribution consisted of a set of audio interviews and speaker-turn time-marked transcripts in transcriber (.trs) format [2]. However, the audio per interview was not segmented in any fashion, supplied as mp2 files, and the interview text itself also comes in one flat file with XML markup. No lexicon mapping words to phonemes was provided. Therefore, a lot of work is involved for an interested party to take this data in its original format and put it in a form suitable for speech recognition experiments.

This supplement distribution attempts to correct that by providing a set of derived files and a lexicon to enable speech recognition research to proceed. The following files are supplied in the associated tar archive:

File	**Brief Description**
malach.kaldi_lm.v2.stm	Text to build language model
malach.minitest.try3.v2.stm	Text for scoring test data
malach.training.information	Maps training data back to original mp2 interviews
malach.dev.information	Maps dev data back to original mp2 interviews
lexicon.txt	Mapping from orthography to pronunciations.
glm	For NIST-based scoring, normalizes text to avoid irrelevant homonym errors (e.g., “B” vs “B.”)
silence_phones.txt	Non-speech phones
nonsilence_phones.txt	Speech phones
optional_silence.txt	Interword pauses
extra_questions.txt	Augments automatically derived questions with non-speech phones.
train/dev subdirectories:	Kaldi-specific for AM training and testing:
text	Utterance texts
wav.scp	Reformats audio to wav for Kaldi
flac	Sub-directory with actual audio (in flac format) for training/testing

segments	Begin/end times per utterance into audio files
reco2file_and_channel	Maps utterances to audio files
utt2spk	Maps utterances to speakers

As mentioned above, insofar as there are no common standards for input to speech recognition systems, rather than inventing yet another format, Kaldi format was used as it is a popular open-source toolkit that many sites employ. However, the information in the above files should permit any basic speech recognition system to be utilized with only minimal reformatting required.

The rest of this document describes the above files in more detail.

1.2 Naming Conventions and Relationship to Original Malach Distribution

The original Malach distribution was provided as a set of files in the following format:

xxxxx-yyy.[mp2 | trs]

xxxxx - Five digit interview code, numbers between 00009 and 36112
yyy - Three digit tape code [001 | 002 | 003…]
mp2 - mp2 format for audio
trs - transcriber format for transcriptions

The original interviews were recorded on a set of videotapes (hence the term “tape code”). Not all tapes were released for all interviews and not all interviews could be released (so there are gaps in the interview numbering).

Each interview contained multiple speakers, though usually there were only two – interviewee, and the interviewer. Sometimes there was more than one interviewer.

As mentioned above, the interviews themselves were not segmented in any fashion – they were distributed “as-is”. However, this is not an easy format to use for speech recognition purposes, especially for training. When training a speech recognition system, the most convenient format is to have a set of relatively short (under 10 second) utterances, labelled by speaker.

Luckily, the interview transcriptions (the “.trs” files) themselves had time markings and indications of speaker turns. For example:

<Turn startTime="0" endTime="33.383" speaker="spk1">
<Sync time="0"/>
<Sync time="9.929"/>
we were speaking before about
<Sync time="13.995"/>
your forty eight hours in Theresienstadt that you were allowed to
<Sync time="17.71"/>
do whatever you wanted
<Sync time="19.322"/>
and eventually after these forty eight hours
<Sync time="22.297"/>
you couldn't do anything else
<Sync time="24.27"/>
<breath> and we were speaking about the Russian
<Sync time="26.403"/>
soldier coming and asking from you
<Sync time="28.566"/>
bicycle that you had actually took from the Germans
<Sync time="31.21"/>
so let's continue from here
</Turn>
<Turn speaker="spk2" startTime="33.383" endTime="323.312">
<Sync time="33.383"/>
yes they used to say
<Sync time="34.885"/>
<UH-UH> i- an- <UH-UH>
<Sync time="36.457"/>
<unintelligible> believe like that
<Sync time="38.35"/>
<UH-UH> give give it to me

The speaker turns are noted by the XML tag “Turn” and pauses were indicated at times by the tag “Sync”. By convention, the first interviewer was typically assigned to be speaker “1” and the interviewee speaker number “2” (but consistency was not always observed). Other speakers (rarely more than a second interviewer) were given sequentially increasing numbers.

Each “Sync” pause was mapped onto a sequentially increasing number (0001, 0002, 0003, etc). Numbering spanned speakers so that if the last Sync pause of Speaker “1” was 0010, the first pause for Speaker 2 at the speaker turn was given the number 0011. Therefore, a section of speech bounded by should pauses for a specific speaker (an “utterance”) can be uniquely located by the following “quadruplet”:

xxxxx-yyy-z-nnnn

xxxxx – Interview code
yyy - Tape code
z - Speaker id

nnnn - Utterance number

In order to compact the notation somewhat, the actual id for each utterance was constructed as

xxxxxzyy-nnnn

and used as the *identifier* for the utterance in the files described in the following sections.

Appendix 1 lists the interviews included as training and test/development data. Out of the original 784 interviews, 674 were selected as training data and 8 for test/development data. The additional 102 interviews will be curated at a later point in time.

The files malach.training.information and malach.dev.information contain the mappings from each utterance identifier back to the original interview. The format is:

identifier original-interview-filename channel begin-time end-time

For convenience, we provide the training and test audio data in a format where the channels are separated in advance and (for the training data only) pre-segmented (see documentation on wav.scp in Section 1.6.5, below).

1.3 Language modeling text

The MALACH text data is distributed as a set of .trs files, which are a set of interview transcripts marked up by XML to indicate speaker turns and silence breaks. The format is consistent with the Transcriber program [2]. The interviews were divided into training and test data. A list of the interviews included in each is given in the Appendix.

A typical text input format for building language models for speech recognition consists of lines of text with whitespace-delimited tokens corresponding to separate sentences. Insofar as the MALACH transcriptions correspond to real

interviews, there is only a loose concept of “sentence” as much of the speech consists of stream-of-consciousness remembrances.

Kaldi itself comes with multiple processing pipelines depending upon the nature of the training materials. The processing pipeline that seemed to best match MALACH was the processing pipeline for the AMI corpus. In that corpus, the language model is constructed from speech transcripts corresponding to the acoustic modelling training data. In this processing pipeline, the language model text is supplied as utterance-by-utterance transcripts stored an “.stm” file [3]. The .stm file has lines of the form:

waveform-name channel speakerID start-time end-time [<attr>] transcription

It was originally designed for scoring test utterance accuracies but here is using as input to a language modeling pipeline that extracts the text from the transcriptions and then is input to the language model building process.

The .stm file was produced by taking each .trs file, segmenting it by the time markers, and noting for each interview when there was a speaker turn. The .trs file was taken verbatim, with no additional attempt to process the data further to rectify spelling inconsistencies or outright errors. The production of the original .trs files did undergo some amount of quality control but the large number of non-English words and variations in regional pronunciations of many of these words made it difficult to remove all inconsistencies across the interviews. The resultant audio was also extracted from the original .mp2 files.

Note that for scoring (if using the sclite option of Kaldi), the waveform-name in the .stm file needs to match the waveform name in the reco2file_and_channel (this is not used in training).

1.4 Development transcripts

As mentioned above, the interviews were separated into training and test data. In addition, the test data was divided into two parts: a “minitest” which can be used for development (“dev”), and a “full” test that can be used for broader evaluations.

The minitest is a random subset of the test data. **This distribution only contains verified transcripts (and audio) for the minitest (dev) data.** The complete test data will follow in a later distribution. The MALACH material presents huge

difficulties in terms of rationalizing multiple spellings of the same foreign names, towns, and concepts, making verification a difficult process even for those familiar with these terms. For example, the Yiddish term for the wife of a Rabbi is "Rebbitzen". Since the original spelling comes from Hebrew characters (Yiddish is written using Hebrew characters) the English spelling is somewhat arbitrary. The transcripts have this term alternately spelled as: "Rebbitzen", "Rebbitsin" and "Rebitisin". There are many other similar cases in both the training and test data. In order to produce an initial release and allow for scoring, this type of spelling "homogenization" was performed for the minitest (dev) data.

The format is the .stm file described in Section 1.3 The .stm format is often used with the "sclite" scoring program produced by NIST [4].

1.5 Lexicon

The lexicon maps words onto pronunciations. Specifically, each word is represented as a string of speech sounds, called "phones". The format is simple:

word phone1 phone2 phone3 …. phone n.

Words with multiple pronunciations are just listed twice in the lexicon; e.g.:

the dh ax
the dh iy

The phone inventory in MALACH is given in Appendix 2 and is a simplified version of the phones in the ARPABET [5]. There is no explicit lexical stress.

The lexicon was created from the word tokens in the utterances corresponding to the training text (below). This was to ensure that each word in the training utterances had an explicit pronunciation (note that Kaldi applies G2P rules on the fly to generate missing pronunciations for training a speech recognition system, but not all recognition systems have this capability). A multi-step process was used to determine the pronunciations.

First, a relatively recent general internal IBM lexicon was searched for each word pronunciation. If the pronunciation was located, it was taken verbatim. Then, a

MALACH lexicon created during the original MALACH program was searched. The problem with the original MALACH lexicon is that a slightly different phone set than the more recent lexicon was used, so such pronunciations had to be mapped onto the phone set in the more recent lexicon. If the word was not found in the original MALACH lexicon, a grapheme-to-phoneme algorithm was utilized [6].

Note that the training text has a variety of non-speech noises explicitly marked in the text, e.g.:

<BGRD_NOISE>
<BKGRD_BREATH>
<BKGRD_COUGH>
<BKGRD_HUM>

They are all marked by angle brackets in the lexicon: "<" followed by the name of the noise followed by ">". In our experiments most of these noises were given the pronunciation of the silence phone, but it is not clear this is optimal and deserves additional experimentation.

1.6 Various Files Closely Associated with building Kaldi Systems

The files in Sections 1.2-1.4 are fairly generic files needed by most speech recognition systems. The sets of files described in this section (except for the training transcripts and training and test audio) are more tightly connected with building systems using Kaldi. Only brief descriptions will be given here; interested parties are directed to the Kaldi documentation for more detailed descriptions of these files and how they are employed in Kaldi.

1.6.1 silence_phones.txt

Lists four silence phones: sil, laughter, noise, oov. The last three phones do not appear in the MALACH training text and are only included here for consistency with other Kaldi builds.

1.6.2 nonsilence_phones.txt

List of speech phones used in lexicon.txt, see Appendix 2.

1.6.3 optional_silence.txt

Phone used to model interword optional silences and ignored in language model context. Here, this is simply the “sil” phone.

1.6.4 extra_questions.txt

Used in the construction of the context-dependent phone decision tree. The question listed here segregates out speech and non-speech phones. The other questions are generated by Kaldi during decision tree building directly from the phone set.

1.6.5 Contents of train/dev subdirectories

The detailed formats of these files are described in [7] and will not be repeated here.

text – utterance by utterance training text using format described in Kaldi documentation

wav.scp – a command line script. Produces a 16 KHz wav file stream from input audio (which is not supplied in .wav format). For the files in dev, there are multiple utterances per wav file; utterance boundaries given by the “segments” file. Note that we list fully qualified names in the wav.scp file so once you untar the distribution, you need to change these absolute references to match where you downloaded the files onto your own system..

flac – audio files. In this distribution provided as flac files (this is a compressed format so it is somewhat smaller to distribute than uncompressed pcm).

segments – begin and end time for each utterance in audio file

reco2file_and_channel – maps utterances to audio files (needed when there are multiple utterances with different begin and end times in a single audio file). Also used by sclite scoring (and needs to match info in the .stm file – a “feature” of sclite).

utt2spk – maps utterances to speaker ids

Again, for more details, please see the Kaldi documentation.

References

[1] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Silovsky, J. (2011). The Kaldi speech recognition toolkit (No. CONF). IEEE Signal Processing Society.

[2] Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production. Speech Communication, 33(1-2), 5-22.

[3] http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/infmts.htm#stm_fmt_name_0

[4] http://www.openslr.org/4/

[5] https://en.wikipedia.org/wiki/ARPABET

[6] Stanley F. Chen. Conditional and Joint Models for Grapheme-to-Phoneme Conversion, In *Proceedings of Eurospeech*, 2003.

[7] http://kaldi-asr.org/doc/data_prep.html#data_prep_data_yourself

Appendix 1 – List of Training and Test Interviews

Training Interviews

01580-003	15223-003	18403-002	19223-002	19769-003
01753-004	15369-003	18405-003	19224-003	19770-003
02023-003	15433-003	18478-004	19267-002	19771-002
03435-004	15453-002	18481-003	19269-002	19779-002
03647-002	15462-002	18511-004	19273-001	19782-002
03658-002	15467-003	18513-002	19275-003	19788-003
03688-002	15596-002	18533-003	19285-003	19790-002
04024-003	15634-002	18548-006	19288-002	19795-002
04044-003	15773-004	18551-002	19293-002	19797-003
04153-004	15980-002	18564-002	19314-002	19799-003
04311-002	16476-002	18633-002	19317-001	19800-002
04556-005	16492-002	18634-002	19327-003	19802-002
04557-002	16543-002	18644-002	19328-002	19803-003
04963-003	16619-002	18648-003	19540-002	19806-003
05357-003	17092-003	18651-003	19564-003	19808-002
05831-005	17175-002	18676-003	19574-004	19813-003
08123-005	17375-003	18682-003	19586-002	19814-002
09584-002	17749-003	18696-003	19610-002	19815-002
12488-002	17826-002	18737-002	19612-002	19829-003
13078-001	17866-002	18760-002	19613-003	19850-002
13096-003	17951-003	18795-003	19615-002	19863-004
14155-003	18032-002	18811-005	19626-002	19864-003
14171-003	18040-003	18970-002	19641-003	19884-002
14175-002	18057-003	19046-002	19692-002	19886-002
14235-001	18061-006	19048-002	19702-002	19887-002
14242-002	18084-003	19064-002	19705-002	19894-002
14251-001	18087-002	19145-002	19707-003	19894-003
14629-002	18090-002	19154-003	19734-002	19894-010
15153-002	18294-003	19171-002	19738-002	19894-013
15211-002	18373-004	19200-003	19753-002	19895-003

19897-002	20029-003	20278-004	20990-001	24542-001
19897-003	20030-003	20279-003	21016-002	24543-003
19898-002	20038-002	20283-002	21326-001	24617-001
19899-002	20040-002	20284-002	21416-001	24651-003
19902-002	20045-003	20295-001	21496-001	24669-002
19903-002	20047-003	20297-003	21668-003	24707-004
19904-002	20050-002	20299-002	21755-001	24756-003
19907-002	20056-003	20332-002	21769-001	24771-001
19908-001	20067-002	20333-003	21857-002	24809-002
19914-003	20071-002	20413-004	21862-002	24812-003
19915-003	20075-003	20416-004	21963-002	24813-003
19917-002	20078-003	20422-005	22576-002	24822-001
19928-003	20080-002	20434-001	22696-002	24880-001
19937-003	20081-003	20451-002	22750-002	24889-001
19939-002	20082-002	20454-002	22750-003	24935-002
19942-002	20083-003	20479-001	22843-004	25004-001
19945-002	20086-002	20526-002	22868-002	25042-002
19947-003	20102-002	20551-004	22877-001	25043-002
19949-002	20107-002	20624-003	22920-002	25078-001
19952-002	20119-002	20635-002	22984-002	25098-001
19954-003	20129-002	20768-001	22984-004	25171-001
19960-005	20131-002	20771-002	22995-001	25246-003
19961-002	20136-002	20778-002	23656-001	25254-001
19977-002	20139-003	20792-002	24105-001	25294-002
19983-002	20140-001	20806-002	24131-005	25296-001
19988-003	20166-002	20815-001	24161-001	25299-001
19989-001	20167-004	20848-003	24177-001	25377-001
19990-002	20174-002	20862-005	24188-003	25419-002
19995-001	20175-002	20863-003	24243-001	25450-002
20001-002	20179-003	20923-003	24265-001	25453-002
20003-003	20182-002	20933-001	24267-002	25460-002
20004-002	20196-001	20936-001	24294-001	25541-003
20008-002	20203-003	20955-002	24311-002	25616-002
20009-003	20207-004	20960-001	24360-001	25639-001
20014-002	20211-003	20963-002	24400-001	25646-002
20015-003	20213-002	20971-002	24431-003	25661-002
20020-003	20214-002	20975-002	24454-001	25817-001
20022-002	20230-002	20976-002	24467-001	25822-002
20023-003	20253-003	20977-002	24517-002	25895-003
20027-002	20277-002	20986-002	24528-003	25929-001

25933-002	26910-001	27938-004	28649-001	33039-004
25988-002	26914-001	27947-002	28670-001	33058-001
25997-003	26923-002	27950-001	28713-005	33099-002
26021-002	27046-002	27987-002	28734-001	33123-003
26059-002	27075-002	27997-001	28738-002	33137-002
26071-003	27113-004	28008-003	28744-003	33137-007
26077-001	27127-001	28039-003	28859-001	33137-010
26106-002	27135-003	28040-002	28866-002	33165-001
26116-001	27137-004	28062-002	28895-001	33169-002
26267-002	27150-001	28070-002	28904-001	33176-003
26271-003	27153-005	28071-002	32531-001	33177-004
26279-002	27155-003	28099-001	32539-002	33201-001
26365-001	27191-001	28101-002	32551-002	33215-002
26367-001	27269-001	28131-002	32554-001	33234-003
26393-002	27279-003	28134-002	32565-003	33241-003
26408-002	27336-003	28176-002	32579-001	33246-003
26409-001	27347-002	28183-001	32584-004	33258-006
26419-002	27369-001	28187-002	32599-003	33258-008
26419-004	27401-001	28197-002	32636-002	33260-002
26439-002	27409-007	28255-002	32638-002	33265-002
26475-001	27422-002	28285-002	32699-002	33266-004
26510-001	27430-002	28313-002	32716-003	33274-002
26524-002	27463-001	28316-003	32722-001	33275-002
26530-001	27546-002	28318-003	32744-003	33279-003
26575-001	27552-002	28320-002	32745-003	33308-001
26582-002	27612-002	28325-002	32755-008	33310-003
26598-004	27624-002	28353-001	32757-003	33315-002
26603-001	27625-001	28372-003	32780-005	33333-001
26604-002	27663-001	28375-001	32791-002	33346-001
26615-002	27683-002	28408-002	32849-002	33363-003
26653-002	27684-003	28413-001	32860-004	33369-002
26734-003	27726-004	28430-002	32907-001	33375-001
26747-003	27728-002	28430-004	32907-004	33375-002
26765-001	27744-003	28431-002	32910-002	33375-003
26769-004	27776-001	28437-003	32914-003	33385-004
26804-002	27793-001	28520-001	32925-002	33393-002
26821-001	27798-002	28521-002	32932-001	33408-003
26823-001	27861-002	28539-002	32992-002	33414-001
26834-003	27882-004	28545-002	32994-001	33414-002
26901-001	27885-003	28628-002	32996-003	33421-002

33422-003	33853-003	34157-002	35099-002	35558-002
33432-002	33855-003	34203-001	35109-001	35579-001
33433-001	33864-002	34235-002	35117-002	35597-001
33448-002	33867-001	34319-003	35120-001	35621-002
33473-002	33871-002	34337-002	35120-003	35634-003
33480-001	33885-002	34359-004	35146-003	35640-002
33497-002	33892-003	34364-002	35147-002	35643-002
33528-003	33894-002	34412-002	35178-002	35715-001
33544-003	33934-003	34416-003	35179-003	35720-002
33579-002	33935-001	34420-002	35183-002	35742-001
33586-002	33937-003	34485-002	35205-002	35747-002
33606-003	33957-002	34603-002	35257-004	35750-002
33689-003	33960-002	34668-004	35259-001	35763-001
33691-001	33987-002	34683-003	35267-002	35774-002
33692-003	34002-002	34707-001	35287-002	35867-002
33725-004	34008-006	34714-002	35302-002	35869-001
33745-002	34015-002	34793-002	35314-001	35944-002
33747-004	34024-002	34976-001	35379-001	35950-001
33758-002	34059-008	34986-002	35424-001	35971-003
33787-006	34075-002	35001-002	35459-002	35995-002
33822-004	34100-003	35057-002	35459-003	36019-002
33833-002	34102-002	35076-001	35461-001	36032-002
33836-003	34122-001	35076-002	35507-002	36035-001
33839-002	34125-002	35077-002	35509-002	36112-003
33845-002	34145-005	35083-001	35519-002	

Test (dev) Interviews

00018-002
00036-002
00041-001
00042-001
00045-002
00055-004
00095-003
00103-002

Appendix 2 – List of Phones in the MALACH Lexicon

AA
AE
AH
AO
AW
AX
AY
B
CH
D
DH
EH
ER
EY
F
G
HH
IH
IY
JH
K
L
M
N
NG
OW
OY
P
R
S
SH
T
TH
UH
UW

V
W
Y
Z
ZH
sil