Chapter 3

Methodology

3.1 Overall methodology

3.1.1 Baseline

1. Train an acoustic model(AM-v0) on 200 hours data
2. Build a language model(LM s1) from 200 hour transcriptions
3. By utilizing the acoustic model(AM-v0), the language model(LM s1), and lexicon(provided by MGB), recognize the short development set(8 hours) to know the overall PER and WER.
4. By using the same AM-v0, LM s1, and lexicon, recognize the 200 hour training set. From the recognition process, calculate PMER, WMER, and AWD of each segment.
5. Sort segments according to its PMER. Choose 100 hour segments with top PMER value if the segments has AWD with range between 0.165 and 0.66.
6. Train a new acoustic model(AM-v1) with the 100 hour data(100hr v-1).
7. Utilizing AM-v1, LM s1, and lexicon, recognize 200 hours data to get 100 hour data(100hr-v2). Retrain an acoustic model(AM-v2) by using 100hr-v2 and re-recognize the 200 hour data.

3.2 Language model

3.2.1 Language model s1(Baseline)

- Generated from the 200 hour transcriptions.

- A lexicon is from words of the 200 hour transcription
- The lm is 4 gram language model
- pruned by 10^{-9}

3.2.2 Language model s2

- from 7 weeks transcription and big subtitles(2 LMs generated for each)
- interpolated with ratio 0.9/0.1
- The lexicon is 160k top words collected from the 7 weeks and big subtitles
- pruned by 10^{-9}

3.2.3 Language model s3

- 7 LMs generated from each genre
- Interpolated with ratio 0.9/0.1 with the big subtitle LM
- Vocab is limited by the 160k lexicon
- pruned by 10^{-9}

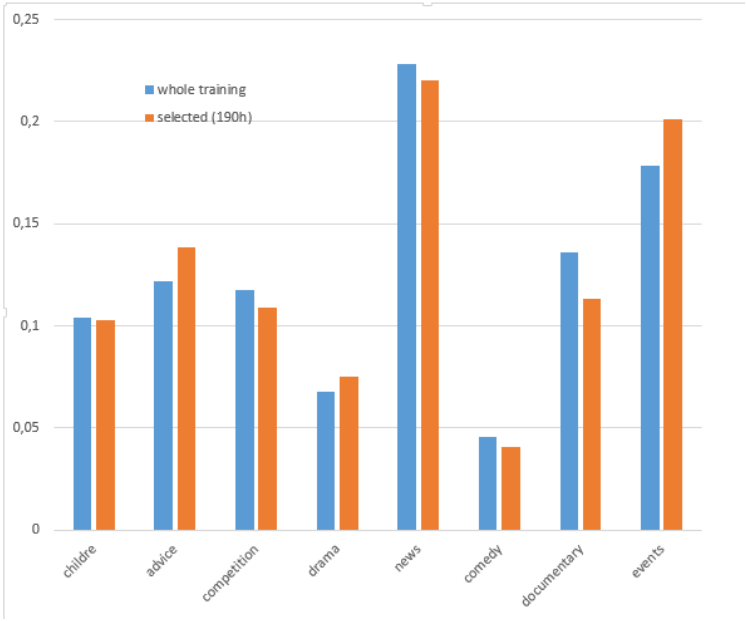
Genre	total time	resource info
documentary	01:54:44	12 cores
news	02:31:24	
events	01:29:03	
drama	01:08:43	
competition	01:32:28	
comedy	01:10:01	
children	01:14:33	
advice	02:56:33	

3.3 Baseline model

3.3.1 The First Acoustic model(AM-V0) and Language Model s1

How to select 200 hours subset of data:
The files in train.full were sorted by their size. After sorting, select one file from 8

Figure 3.1: Bar chart of each genre from all training data and 200 hours training data



files(the biggest file from 8 files). Because the duration of train.full is more or less 1600 hours, 200 hours of data was obtained by using this random selection.

Calculating PMER and WMER for each segment by utilizing AM-v0 and LM s1:

1. GMM: using the 200 hours data

2. TDNN(TDNN-v0): the result of GMM training is utilized to build TDNN.

3. TDNN graph creation
   The language model used is the LM s1; the acoustic model is TDNN-v0; furthermore, the dictionary is provided by MGB(lexique_226604).

4. TDNN development set recognition
   The development set is files from dev.short, which contains 8 hours of audio. The development set is recognized by leveraging the graph created in the previous step.

5. TDNN 200 hours of training data recognition
   We need to recognize 200 hours data to calculate PMER, WMER, and AWD. This data will be used for data selection of the next iteration.

6. Sclite score PMER and WMER
   Use the lattice and graph to calculate overall PMER and WMER. However, the

script does not provide PMER, WMER, and AWD for each segment. Fortunately, the script generates prf files which contains number of correction, insertion, deletion, and substitution for each segment. Two prf files exist:(ctm_words.filt.filt.prf contains word match error rate and ctm_phones.filt.filt.prf contains phone match error rate). From there, we are able to calculate PMER, WMER, and AWD of each segments. For the development set, we only need to know the overall PER and WER. In contrast, training set recognition needs PMER, WMER, and AWD for each segment.

No.	Description	Result	execution time	extra info
1	GMM		6h 22m 45s	20 hosts, 4 cores
1.1	GMM decoding	57.2% WER	3h 36m 47s	20 hosts, 4 cores
2	TDNN AM-V0 LM s1		38h 30m 31s	3 nodes
3	TDNN graph creation		7m 22s	1 host/ 2 cores
4	TDNN dev set recognition		2h 58m 58s	40 cores
5	TDNN training set recognition		9h 25m 38s	20 hosts
6	TDNN AM-V0 dev short PMER, WMER calculation	32.7 % PER 49.2% WER	1h 10m 57s	1 host
6.1	TDNN AM-V0 train 200 PMER, WMER calculation	24.8 % PER 32.4 % WER	5h 5m 51s	1 host
6.2	Calculate PMER and WMER per segment		20m 22s	1 host

3.3.2 The First Acoustic model(AM-V0) and Language Model s2

No.	Description	Result	execution time	extra info
3	TDNN graph creation		3h 12m 25s	2 hosts
4	TDNN AM-V0 dev set recognition		1h 10m 59s	20 hosts
5	TDNN AM-V0 training set recognition		10h 9m 35s	20 hosts
6	TDNN AM-V0 dev short PMER, WMER calculation	31.9% PER 50.2% WER	4h 58m 26s	1 hosts
6.1	TDNN AM-V0 train 200 PMER, WMER calculation	26.4% PER 39.7% WER	4h 58m 26s	1 hosts

3.3.3 The First Acoustic model(AM-V0) and Language Model s3

3.3.4 The First Acoustic model(AM-V1) and Language Model s1

How to select 100 hours subset of data for baseline model v1:
Section 3.3.1 calculates PMER, WMER, and AWD for each segments. Consequently, we can extract segments(of 200 hours transcriptions) which satisfy $0.165 \leq AWD \leq 0.66$. In the other word, the segments which does not satisfy are removed. Then, segments are sorted based on their PMER in ascending. We can deduce PMER threshold and its total duration by sorting PMER as presented in the following table:

PMER	Total duration	Percentage of total segment duration
0.0	14.83 h	10%
0.05	29.66 h	20%
0.1	44.49 h	30%
0.16	59.32 h	40%
0.24	74.15 h	50%
0.33	88.98 h	60%
0.47	103.81 h	70%
0.82	118.64 h	80%

The Table 3.3.4 shows that 60% of data has 0.33 PMER with total duration of 88.98 hours of audio. From the table, we conclude that threshold 0.33 is selected as the threshold to select data for the next iteration.

WMER	Total duration	Percentage of total segment duration
0.0	14.83 h	10%
0.08	29.66 h	20%
0.15	44.49 h	30%
0.23	59.32 h	40%
0.32	74.15 h	50%
0.43	88.98 h	60%
0.58	103.81 h	70%
0.98	118.64 h	80%

No.	Description	Result	execution time	extra info
1	GMM-V1		3h 28m 43s	20 hosts
1	TDNN-V1		on training	3 hosts

Figure 3.2: Graph of PMER and WMER produced transcription AM-V0 and LM s1. The red line represents PMER, while the blue line represents WMER.

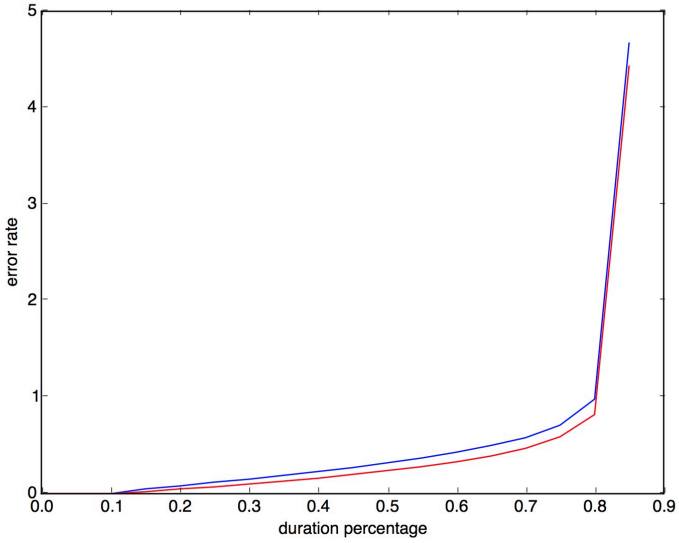


Figure 3.3: Graph of PMER and WMER from transcription align, provided by MGB.