LibriSpeech

ASR corpus based on public domain audio books

V. Panayotov, G. Chen[1], D. Povey[1], S. Khudanpur[1]

[1]Center for Language and Speech Processing
Human Language Technology Center of Excellence
The Johns Hopkins University, Baltimore, MD 21218, USA

ICASSP, 2015

Outline

1. Introduction
2. Audio alignment
3. Corpus preparation
4. Experiments

Introduction to LibriSpeech

- Deep learning is the state-of-the-art technique for training acoustic models for ASR

Introduction to LibriSpeech

- Deep learning is the state-of-the-art technique for training acoustic models for ASR
- Performance of deep neural networks improve faster than GMMs with the amount of training data available

Introduction to LibriSpeech

- Deep learning is the state-of-the-art technique for training acoustic models for ASR
- Performance of deep neural networks improve faster than GMMs with the amount of training data available
- LibriSpeech is a large read-speech ASR corpus
  - Based on carefully aligned public domain audio books
  - Roughly 1000 hours in total
  - Available under a permissive license (CC-BY 4.0) at `http://www.openslr.org/12/`

Kaldi's approach to ASR system building

- Ready to use recipes for building ASR systems on various data sets
- Good, because:
  - Gives examples of the use of the tools available in Kaldi
  - Enables reproducible research
  - Everyone is invited to try and see in practice, how state-of-the-art ASR performs

Kaldi's approach to ASR system building

- Ready to use recipes for building ASR systems on various data sets
- Good, because:
  - Gives examples of the use of the tools available in Kaldi
  - Enables reproducible research
  - Everyone is invited to try and see in practice, how state-of-the-art ASR performs (provided they have data)

Availability of speech databases

- Most of the ASR data sets remain inaccessible to many due to high cost
- Unlike Computer Vision, where many standard benchmarks are freely available, at least for research purposes.

Availability of speech databases

- Most of the ASR data sets remain inaccessible to many due to high cost
- Unlike Computer Vision, where many standard benchmarks are freely available, at least for research purposes.
- Some recipes with free data are already available in Kaldi:
  - AMI (meeting recordings)
  - TED-LIUM (TED talks)
  - Språkbanken (Danish, Norwegian and Swedish)
  - VoxForge (community speech gathering effort)
  - Vystadial (Czech and English telephone speech)

Availability of speech databases

- Most of the ASR data sets remain inaccessible to many due to high cost
- Unlike Computer Vision, where many standard benchmarks are freely available, at least for research purposes.
- Some recipes with free data are already available in Kaldi:
  - AMI (meeting recordings)
  - TED-LIUM (TED talks)
  - Språkbanken (Danish, Norwegian and Swedish)
  - VoxForge (community speech gathering effort)
  - Vystadial (Czech and English telephone speech)
- LibriSpeech is about 2 times larger than all freely available English corpora combined, and has less licensing restrictions than any of them with the exception of Vystadial.

LibriVox project

- LibriSpeech is based on English audio from the LibriVox project
  - LibriVox started in 2005 by Hugh McGuire
  - volunteer readers
  - based on public domain books, mostly from Project Gutenberg
  - audio data hosted on the Internet Archive (archive.org)
  - approximately 7300 completed English audio book projects to date
  - approximately 90 new projects completed per month, most of them in English
  - different genres and duration- short poetry to long epic novels
  - provides API for access to metadata (readers, chapters, language etc)
  - some data were previously used for TTS (Blizzard Challenge 2012)

LibriSpeech Goals

The development of LibriSpeech was guided by the following goals:

- provide high quality data freely usable by anyone for every purpose
- balanced corpus in terms of genders and per-speaker duration
- facilitate reproducible evaluation by defining development and test sets of significant size
- Kaldi example of building state-of-the-art acoustic models
- efficient alignment procedure, suitable for processing large amounts of data

Outline of the alignment procedure

- Objectives
  - establish correspondence between the audio and the reference text
  - flag all places in audio where text-audio discrepancy is hypothesized
  - split the "good" audio into chunks suitable for ASR AM training

Outline of the alignment procedure

- Objectives
  - establish correspondence between the audio and the reference text
  - flag all places in audio where text-audio discrepancy is hypothesized
  - split the "good" audio into chunks suitable for ASR AM training
- based on a well-known alignment procedure (Hazen, 2006) with small modifications
  - first decoding pass to find "islands of confidence"
  - second decoding pass seeking for text-audio discrepancies between any two islands of confidence

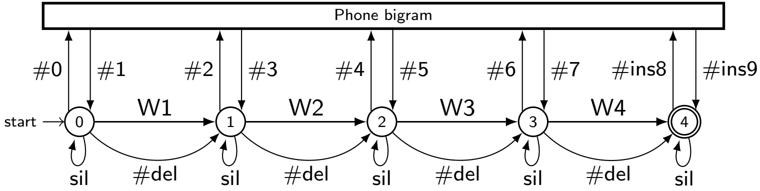
First alignment stage

- Preprocessing of the reference text (we learn text/audio pairing from the metadata)
  - text normalization: expansions of abbreviations, numbers etc, into words
  - automatic generation of pronunciations for the OOV words
- regular decoding with Kaldi's standard 1-best decoder (*gmm-decode-faster*) using a bigram LM, built on book's text
  - long audio chapters are split in slices of up to 20 minutes in length
- localize the part of the text (typically a chapter) corresponding to the imperfect ASR transcript, using Smith-Waterman local alignment
- find islands of confidence- exact matches between the reference and decoding results of total length 12 phones or more
- use dynamic programming to split the audio into segments of approximately 35 seconds in length
  - only silences of length 0.5 seconds or more, inside an island of confidence are used as split points

Second alignment stage

- Tries to detect discrepancies between the audio and the text in each of the 35 second segments produced in the first stage.
- Uses an acoustic model adapted on the results of the first stage decoding
- Some typical sources of mismatch:
  - reader-introduced insertions/deletions/substitutions
  - wrong text normalization
  - grapheme-to-phoneme errors in the OOV words
  - inaccuracies in the reference text

Second alignment stage(implementation)



- deletions(from audio) are modeled as skip arcs(#del) between the words
- insertions: arbitrary string of phonemes between any two words (e.g. phone bigram between arc #6 and #7 allow insertions b/w words "W3" and "W4").
- substitutions: allow any reference word to be substituted by a string of phones (e.g. input arc #4 and output arc #7 allow for substitution of "W3")
- the discrepancies are penalized by empirically determined costs
- phone bigram is shared for space and time efficiency.
- modifications in the decoder: token entering the phone bigram through arc *i* can leave only through arcs *i+1*(insertions) or *i+3*(substitutions).

Final segmentation into utterances

- concatenate the alignment results of the second stage
- split into utterances suitable for training acoustic models(length in the range 3-35 seconds)
- split only on silence intervals of length 0.3 seconds or more, that were recognised as silence in the first alignment pass too.
- all potential utterances for which one or more discrepancies from the transcript were hypothesized are discarded
- two ways of splitting:
  - on arbitrary silence intervals- maximizes the utilization of the audio. Suitable for the AM training material.
  - on silence intervals that coincide with sentence breaks. More suitable for evaluation data. The utterances are more meaningful and amenable to language modeling

Selection and cleaning of corpus data

- single-speaker chapters only. We want clean train/test set separation, so discard e.g. LibriVox' "dramatic reading" works
- to assess the quality and to provide some annotation of the data, a custom GUI application was built:
  - male/female labels
  - allows to flag audio that is of distinctively low quality (noisy etc)
  - allows to inspect recording that are suspected to be read by more than one speaker. Speaker diarization was used to flag audio that has higher probability of such problems.
- the data was prepared, so that there is a good balance in terms of per-speaker audio.
- LibriVox has roughly equal percentage of male and female readers, so the corpus is balanced in that regard too.

Partitioning into subsets

- data is partitioned, to make easy for the user to download and use as much data as needed
- data is first decoded with an AM built on WSJ. Readers are then ranked according to the average WER of their recordings, and are partitioned into 2 pools:
  - "clean" pool- readers with lower than the median WER. We expect the readers in this group represent speakers with native or near-native US English pronunciation.
  - "other" pool- readers with higher than the median WER. We expect that this group contains most of the speakers with non-native pronunciations and accents different from US English.

Partitioning into subsets(cont.)

- evaluation data is then drawn randomly from both pools, in such a way as to provide approximate lower and upper bound on the performance of an acoustic model on clean read speech:
  - "clean" test and development sets- drawn at random from the "clean" pool
  - "other" test and development sets- drawn randomly from the readers having very high WER with the WSJ AM (third quartile)
  - all evaluation sets are roughly 5 hours in length and include around 20 men and 20 women.
- training data is split into three disjoint subsets:
  - 100 hour "clean" subset
  - 360 hour "clean" subset
  - 500 hour "other" subset

Partitioning into subsets(cont.)

subset	hours	per-spk minutes	female spkrs	male spkrs	total spkrs
dev-clean	5.4	8	20	20	40
test-clean	5.4	8	20	20	40
dev-other	5.3	10	16	17	33
test-other	5.1	10	17	16	33
train-clean-100	100.6	25	125	126	251
train-clean-360	363.6	25	439	482	921
train-other-500	496.7	30	564	602	1166

Table: Data subsets in LibriSpeech

Language models

- allows for reliable and reproducible way of evaluating acoustic models on the LibriSpeech data set
- based on 14500 texts Project Gutenberg, containing around 803 million tokens and 900 000 unique words
- the LM data were carefully filtered to ensure it doesn't include in whole or in part some of the texts on which the test data is based
- the pronunciations for the OOV words autogenerated with G2P
- OOV rate of the evaluation sets around 0.5%
- LM perplexity for the 3-gram models is around 170 and for the 4-gram it is 150
- the language models along with the original data are available for download from http://www.openslr.org/11/

Experiments

- LibriVox audio is .mp3-compressed and can be cleaned-up(volume normalization and de-noising)
- evaluate the performance of models built on LibriSpeech on non-compressed data(WSJ):
  - WSJ's LM models
  - WSJ test sets
- then reverse the situation- evaluate WSJ acoustic models on LibriSpeech test sets
- evaluation using different training data sizes/subsets
- evaluate two kind of acoustic models:
  - SAT GMM
  - DNN
- efficient rescoring with large LMs was implemented in Kaldi to facilitate these experiments

Experiments(cont.)

Acoustic model		eval'92	dev'93	eval'93
LS	SAT 100h	5.72	10.10	9.14
	SAT 460h	5.49	8.96	7.69
	SAT 960h	5.33	8.87	8.32
	DNN 100h	4.08	7.31	6.73
	DNN 460h	3.90	6.75	5.95
	DNN 960h	**3.63**	**6.52**	**5.66**
WSJ	SAT si-284	6.26	9.39	9.19
	DNN si-284	**3.92**	**6.97**	**5.74**

Table: WERs on WSJ's test sets under the "open vocabulary" (60K) test condition

Experiments(cont.)

Acoustic model		dev-clean	test-clean	dev-other	test-other
LS	SAT 100h	8.19	9.32	29.31	31.52
	SAT 460h	7.26	8.34	26.27	28.11
	SAT 960h	7.08	8.04	21.14	22.65
	DNN 100h	5.93	6.59	20.42	22.52
	DNN 460h	5.27	5.78	17.67	19.12
	DNN 960h	**4.90**	**5.51**	**12.98**	**13.97**
WSJ	SAT si-284	10.87	12.44	39.44	41.26
	DNN si-284	**7.80**	**8.49**	**27.39**	**30.01**

Table: WERs on LibriSpeech's test sets; all results are obtained by rescoring with a 4-gram language model.

Experiments(cont.)

Language model	dev-clean	test-clean	dev-other	test-other
3-gram prn. thresh. 3e-7	7.54	8.02	18.51	19.41
3-gram prn. thresh. 1e-7	6.57	7.21	16.72	17.66
3-gram full	5.14	5.74	13.89	14.77
4-gram full	4.90	5.51	12.98	13.97

Table: LM rescoring results for the 960 hour DNN model

Thank you!