ARE AUDIO OR TEXTUAL TRAINING DATA MORE IMPORTANT FOR ASR IN LESS-REPRESENTED LANGUAGES?

Thomas Pellegrini, Lori Lamel

LIMSI-CNRS, BP133, 91403 Orsay cedex, FRANCE
thomas.pellegrini@limsi.fr, lamel@limsi.fr

ABSTRACT

State-of-the-Art speech recognizers are typically trained on very large amounts of data, both transcribed speech and texts. With the recent growing interest in developing speech technologies for languages for which only small amounts of data are accessible, collecting appropriate data is a key issue in building new speech recognition systems. This article reports on an experimental study assessing the performance of a speech recognizer for a less-represented language, as a function of the quantity of texts and transcribed speech data available for model training. The experimental results show that for supervised training with only 2 hours of manually transcribed data, the acoustic models are the weak point. With 10 hours or more of transcribed audio data, the quantity of texts has a larger affect on the error rate than the quantity of speech.

Index Terms— Automatic speech recognition, less-represented languages, broadcast news transcription

1. INTRODUCTION

A key challenge in rapidly porting a speech recognizer across languages is minimizing the effort needed to collect audio and text data. At least a few hours of audio data need to be manually transcribed, and the text data require some normalization, for example, the processing of numbers, amounts and dates transforming them to approximate spoken language. A natural question arises as to where it is most efficient to spend effort? Is it better to invest time collecting texts for example from the Web to estimate language models, or is it better to collect and transcribe a few more hours of audio data? There is probably no one answer to this question, and the answer is likely to depend upon what operational point the system is at with a given quantity of audio and text data.

With the growing interest for technologies allowing multilinguality, more and more languages are concerned by speech technologies and by speech recognition in particular. The famous citation "There is no data like more data" has not been contradicted so far. In the case of less-represented languages, it is often relatively easy to obtain a few hours of audio data for acoustic model training, whereas if can be quite difficult to find sufficient representative texts in electronic form. A recurrent question posed by researchers, users and funding agencies is how much and what types of data are needed to train the models, and what performance can be expected with a given amount of resources.

There are not very many studies in the literature about the relationship between word error rate (WER) and the amount and type of training data. In [1], closed-captions and detailed transcriptions were compared for acoustic model (AM) training. ASR performance with lightly supervised AM training was shown to approach that obtained with supervised training, on a broadcast news (BN) task with a well-trained language model (LM).[1] In [2], the studies were extended by assessing the WER with quantities of training data ranging from 10 minutes to 200 hours of raw data, while dramatically reducing amount of textual data (1.8M words) used for language model estimation. An iterative procedure was shown to improve the quality of the acoustic models despite a very high initial WER. These studies showed that detailed manual transcriptions are not a requirement for acoustic model training.

However, the experiments reported in [1] and [2] were carried out on American English, which is the language the most used in ASR research. Approaches or results on less-represented languages, for which little or no expertise at all is available, may be different.

This work studies the impact on the WER for the different types of training data: speech material used to train acoustic models; texts corresponding to the transcriptions of the speech corpus; and texts collected from newspapers and newswires available on the Web. The transcripts and the texts are used to estimate language models. As a case study, speech recognition experiments are carried out for the Amharic language, for which only relatively small amounts of audio and text data are available.

2. AMHARIC CORPUS

The Amharic language is an example of a less-represented language, for which only small quantities of written texts are available. There are some recent studies on speech recognition and speech processing for Amharic [3, 4], and a web

[1] Over 1 Billion words of LM training data were used including 790M words of newspaper and newswire texts and 240M words of commercial BN transcripts.

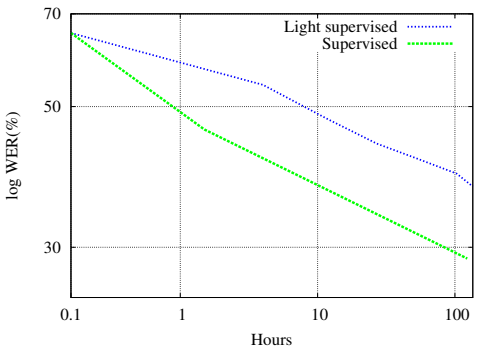

Fig. 1. Word Error Rate (%) as a function of acoustic training data quantity (taken from WERs reported in [2])

resource portal for Amharic corpora has also been created. [2] Amharic has 34 basic symbols, for which there are 7 vocalizations: /ɛ/, /u/, /i/, /a/, /e/, /ə/ and /o/, referred to as the seven orders. The basic symbols are modified in a number of different ways to indicate the different vocalizations. 85% of the syllables represent a CV sequence (C for consonant and V for vowel), one symbol represents the complex sound /ts/V and the reminder represent CwV sequences (where w is a semi-consonant).

Compared to other languages for which models and systems have been developed [5], the Amharic audio corpus is quite small. It comprised of 37 hours of broadcast news data from two sources, *Deutsche welle* and *Radio Medhin*. The data were transcribed by native Ethiopian speakers, and contains a total of 247k words with 50k distinct lexemes. Two hours of data taken from the latest shows of each source were reserved for development test. This development data contains 14.1k words, of which almost 15% do not appear in the training portion, as shown in table 2 with the text size of zero and the 35h training condition.

In addition to the transcriptions of the audio data, about 4.6 million words of newspaper and web texts have been used for language model training. Over 340k distinct words are found in these texts.

3. IMPACT OF AUDIO TRAINING DATA QUANTITY

In [2], the performance of a state-of-the-art LVCSR system was studied as a function of speech training material, with quantities ranging from 10 minutes to 200 hours. Figure 1 represents the WER reported in Table 4 of [2], as function of the amount of raw acoustic model training data. According to [6], there is a linear relationship between word error rate and the quantity of training material. The lightly supervised acoustic modeling led to a 37.4% WER when using 135h of automatically transcribed data with a language model trained on 1.8M words whereas the initial WER, achieved with only 10 minutes of manual transcribed data was 65.3%. With the same 123h of data trained in a supervised manner, a WER of 28.8% has been obtained with the same language model.

[2] http://corpora.amharic.org/

4. IS IT MORE IMPORTANT TO COLLECT TEXTS OR TRANSCRIBE SPEECH?

In order to measure the impact of the acoustic and the syntactic components of a speech recognizer on the WER, several systems have been built varying the quantity of training data. Acoustic model sets have been trained on four sized subsets of the training corpus: 2, 5, 10 and 35 hours of data. For each of these subsets, component trigram language models are trained (modified Kneser-Ney smoothing) on the corresponding transcriptions. Other component trigram LMs are trained on different quantities of texts from the 4.6M word corpus, selecting 10k, 100k, 500k, 1M and 4.6M words. The LMs used in the systems result from an interpolation between the LMs estimated on the transcriptions and the LMs estimated on the texts. All LMs are 3-grams with modified Kneser-Ney smoothing. Since only limited transcribed data were available, the development corpus was also used to optimize the interpolation coefficients for each system by averaging the coefficients obtained on randomly selected subsets of the dev data. The weight of the audio transcripts ranges from 0.2 to 0.9, and is generally higher when the text corpus is small.

4.1. Audio training data selection

The audio data were recorded over a 1-year period from January 2003 to January 2004. The development corpus is comprised of the last shows from each source, and date from December 2003 and January 2004. The 2-hour (h) subset of the audio training corpus was selected from November 2003. Then additional data, anterior in date, were added to obtain the 5h and 10h subsets. Finally the entire 35h audio training corpus has been used. The 2, 5, 10 and 35-hour corpora contain 17k, 35k, 70k and 240k words respectively.

4.2. Text subset selection

In total there are about 4.6 million words of LM texts. A first 10k word text is extracted by selecting one sentence out of 470. The every 51st sentence of the remaining texts has been added to this 10k word corpus to obtain the 100k word corpus. And similarly for the larger subsets.

Table 1 gives the lexicon sizes for each test configuration. No selection criterion was applied to build the lexica for the text subsets – all words from the transcripts and the texts were included. However, when the complete text corpus of 4.6M words was used, a minimum occurrence count of three in the texts was applied.

Transcripts		*Words (texts)*				
Hours	*Words/Types*	*10k*	*100k*	*500k*	*1M*	*4.6M*
2h	*17k / 7k*	11k	36k	96k	142k	114k
5h	*35k / 12k*	16k	39k	98k	144k	116k
10h	*70k / 21k*	24k	45k	103k	148k	119k
35h	*240k / 50k*	52k	69k	120k	163k	133k

Table 1. Lexicon sizes (in number of word types) for each audio/text configuration.

Transcripts	*Words (texts)*					
	0	*10k*	*100k*	*500k*	*1M*	*4.6M*
2h	36.2	30.6	18.8	11.5	9.0	8.1
5h	28.5	25.7	17.3	11.1	8.8	7.9
10h	22.7	21.4	15.7	10.6	8.4	7.7
35h	14.5	13.9	12.1	9.1	7.5	7.0

Table 2. OOV rates of dev data as a function of the number of words in the audio and text corpora.

4.3. Lexicon size

With the combined 2h transcript and 10k word corpora, there are only 11k distinct words types. The number of word types grows rapidly with the number of word tokens, which is characteristic of languages with a rich morphology such as Amharic [7]. For the combined 100k word text and the 35h transcripts (240k total words), the lexicon contains 69k distinct entries. Lexicons including the entire text corpus are smaller than those including the 1M word texts because of the frequency cut-off used in the former case. For the 1M word texts and the 35h transcripts, the lexicon contains 163k whereas with the entire 4.6M word text, the lexicon contains 133k words [3]. Without the frequency cutoff, the full corpus contains 350k word types.

4.4. Out-Of-Vocabulary words

Table 2 gives the out-of-vocabulary (OOV) rates for each configuration. With a 2h transcript corpus and a 10k word text, 30.6% of words of the devset are OOV. With the full 4.6M word text and the 2h audio transcripts, this rate is reduced to 8.1%. Including all transcripts, the OOV rate is 7.0%. Figure 2 illustrates the evolution of the OOV rate vs the text corpus size. The curves correspond to three of the subsets of the audio transcripts: 2h, 10h and 35h. The 5h curve has not been drawn to simplify the figure, but the OOV rates are given in Table 2. The OOV rates with 5h are seen to split the difference of the 2h and 10h subsets. With 2h of transcripts, increasing the texts from 10k to 100k words reduces the OOV rate by 11.8% absolute. With the 10k word text corpus, the OOV reduction is 9.2% absolute when increasing the audio transcripts from 2h to 10h. Increasing the amount of audio transcripts is seen to have a large effect on the OOV, particularly for smaller sized text corpora. It can be seen that doubling the texts from 500k to 1M has a much larger effect on the OOV rate than increasing by over a factor of 4 from 1M to 4.6M words.

[3] Applying even a cut-off of 2, that is removing singletons, reduced the lexicon size by more than 50%.

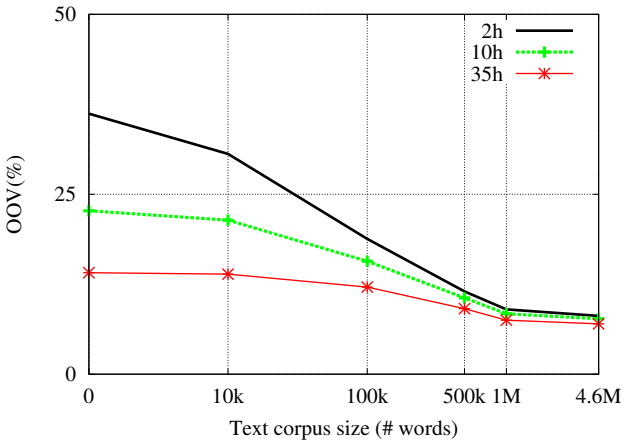

Fig. 2. OOV rates as a function of the text corpus size. The three curves correspond to the three distinct transcript corpora: 2h, 10h and 35h.

Hours	*2h*	*5h*	*10h*	*35h*
# Contexts	3027	4557	6323	10726
# Tied states	1187	2286	3861	8554

Table 3. Number of modeled phone contexts and tied states for all audio training subsets (2h, 5h, 10h and 35h).

4.5. Recognition experiments

This section reports recognition results obtained with systems trained for each transcript/text configuration. The speech recognizers all have two decoding passes, with unsupervised acoustic model adaptation after the first decoding pass [8]. Specific acoustic models were built for each of the three

Transcripts	*Words (texts)*				
Hours / Words	*10k*	*100k*	*500k*	*1M*	*4.6M*
2h / 17k	64.4	59.6	55.5	53.0	48.5
5h / 35k	45.5	38.7	33.1	30.9	28.0
10h / 70k	40.4	35.5	30.9	28.7	26.7
35h / 240k	31.4	29.9	27.0	25.7	24.4

Table 4. Word Error Rates (%) for all combinations of audio data (2h, 5h, 10h and 35h) and texts (10k, 100k, 500k, 1M, 4.6M).

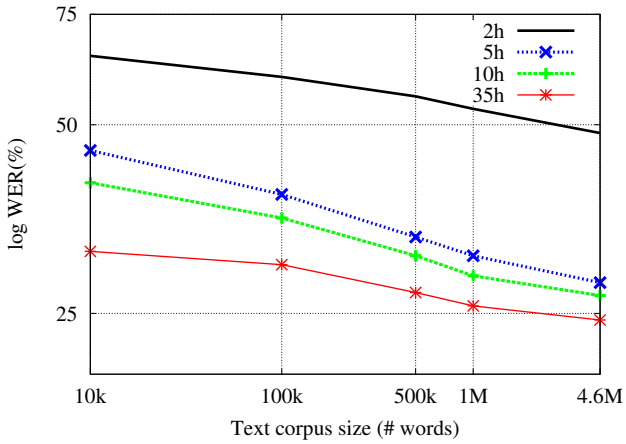

Fig. 3. Word Error Rates (WER in %) as a function of the text corpus size. The four curves correspond to the different sized transcript corpora, with 2h, 5h, 10h and 35h of audio data.

audio training corpus subsets. The models are all tied-state HMMs, covering both intra-word, and cross-words contexts with 3 states per model and 32 Gaussians/state. Grapheme to phoneme conversion is straighforward in Amharic, and the pronunciations are represented with 33 phones and an additional 3 non-speech units. The number of modeled phone contexts and corresponding number of tied states are given in Table 3.

Table 4 and Figure 3 give the word error rates for the systems built with 2h, 5h, 10h, and 35h sets of acoustic training data as a function of the text corpus size used to estimate the LMs that are interpolated with the transcript LMs. The WERs of the 2h system are about twice those of the 35h system for all training text corpus sizes. The 5h system behaves more like the 10h system. For the 35h system the relative WER reduction is 22% when increasing the texts from 10k words to 4.6M words. The largest gain comes when going from 100k words to 500k words (10% relative) whereas the other relative gains are about 5%. For the 2h system, the relative WER reduction is 25% when going from 10k words to 4.6M words to for LM estimation. Nevertheless, the best system built with 2h of audio training data has a very high WER of 48.5%. The middle curve in Figure 3 for the 10h system behaves like lowest curve for the 35h system. With the largest text sets (1M and 4.6M words) the absolute differences between the 10h and 35h systems are 3% and 2% respectively (10% and 8% relative). For the small text sets (10k and 100k) the relative differences are 22% and 16%. It thus appears that the acoustic training data are particularly important for both acoustic and language modeling when the text corpora are small. (In this range the number of words in the transcripts are on the order of or larger than the text corpora).

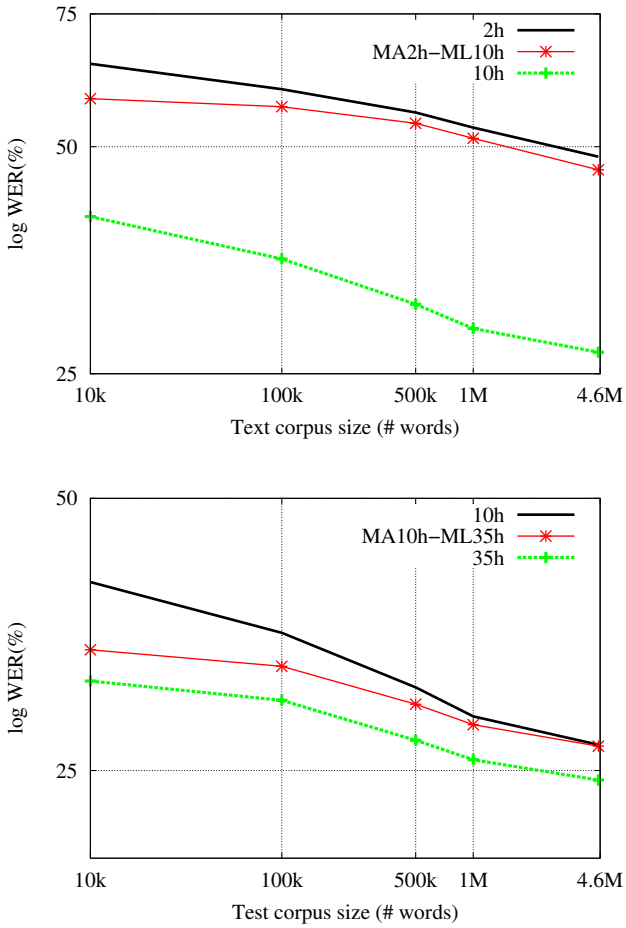

Fig. 4. Log of the Word Error Rates (%) as a function of the text corpus size, corresponding to Table 5.

Transcripts	*# Words (texts)*				
	10k	*100k*	*500k*	*1M*	*4.6M*
2h	64.4	59.6	55.5	53.0	48.5
AM2h/LM10h	57.9	56.5	53.7	51.3	46.6
10h	40.4	35.5	30.9	28.7	26.7
AM10h/LM35h	34.0	32.6	29.6	28.1	26.6
35h	31.4	29.9	27.0	25.7	24.4

Table 5. Word Error Rates (%) as a function of the text corpus size, for the 2h and 10h systems and two crossed systems 'AM2h-LM10h' and 'AM10h-LM35h', with AMs trained on 2h/10h and the transcript LM component trained on the 10h/35h, respectively.

4.6. Influence of LM vs Acoustic models

Figure 3 shows large performance differences between the 2h and the other systems, for all text corpus sizes. These systems differ in the speech data used to train the acoustic models and the transcripts used to train the LM. To evaluate whether one of these has a larger influence than the other on the WER, a "crossed" system was built, where the acoustic models are those of the 2h system, but the LM transcript component is that of the 10h system. Similarly a crossed system was built using the AM of the 10h system and the LM transcript component of the 35h system.

Table 5 summarizes the WERs as a function of the text corpus size, the Figure 4 plots the corresponding log WERs. The smallest "crossed" system, named 'AM2h-LM10h' on the left graph has WERs closer to the full 2h system than to the 10h system. This seems to indicate that with very small amounts of audio training data, adding a few more transcribed hours is crucial. Adding the transcripts only to the language model corpus, only improves the WER by about 5% relative to the 2h system, whereas using the data for both AM and LM training reduces the WER by 30-40% compared to the 2h system. As shown on the right side, the WER of the 'AM10h-LM35h' system behaves more like the 35h system for small text quantities (10k and 100k words) and more like the 10h system for the LM components trained on the larger text sets. This shows that as can be expected the influence of the transcript component of the LM is reduced when larger text corpora are available. Recall though that the interpolation weight of this component is high.

5. SUMMARY

This paper has presented an experimental study assessing the performance of a speech recognizer for a less-represented language, Amharic, as a function of the quantity of texts and transcribed speech data available for model training. The experimental results show that with supervised training, if only 2 hours of manually transcribed data are used for acoustic model training, the WER rate remains quite high, even when the text corpus increases up to 4.6M words. With more than 10 hours of transcribed audio data, the quantity of texts has a larger affect on the error rate. Systems with acoustic models trained on 10 hours and 35 hours have somewhat similar performances (about 25% word error) when a minimum of 1M words are used to train the language models, suggesting that it at this operating point collecting text data may improve performance more than additional audio data. Other studies have shown that automatic transcription can be used with success for acoustic modeling, thus a further study would be to combine new text collection and non-supervised acoustic modeling.

6. REFERENCES

[1] L.Lamel, J-L. Gauvain, and G. Adda, "Lightly supervised acoustic model training," in *Proceedings of ISCA ITRW Workshop on Automatic Speech Recognition: Challenges for the new Millenium*, Paris, 2000, pp. 150–154.

[2] L.Lamel, J-L. Gauvain, and G. Adda, "Unsupervised acoustic model training," in *Proceedings of ICASSP*, Orlando, 2002, pp. 877–880.

[3] S.T. Abate, W. Menzel, and B. Tafila, "An Amharic Speech Corpus for Large Vocabulary Continuous Speech Recognition," in *Proceedings of INTERSPEECH*, Lisboa, 2005.

[4] T. Pellegrini and L. Lamel, "Investigating Automatic Decomposition for ASR in Less Represented Languages," in *Proceedings of INTERSPEECH*, Pittsburgh, 2006.

[5] L.Lamel et al, "Speech Transcription in Multiple Languages," in *Proceedings of ICASSP*, Montreal, 2004, vol. 3, pp. 757–10.

[6] R K Moore, "A comparison of the data requirements of automatic speech recognition systems and human listeners," in *Proceedings of EUROSPEECH*, Geneva, 2003, pp. 2582–2584.

[7] K.Kirchhoff and R. Sarikaya, "Processing morphologically rich languages," in *Workshop Interspeech*, Antwerp, 2007.

[8] JL. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News transcription system," *Speech Communication*, vol. 1-2:37, pp. 89–108, 2002.