

LARGE CORPUS EXPERIMENTS FOR BROADCAST NEWS RECOGNITION

Patrick Nguyen, Luca Rigazio and Jean-Claude Junqua

Panasonic Speech Technology Laboratory (PSTL)
3888 State Street, Suite 202,
Santa Barbara, CA 93105, U.S.A.
nguyen@research.panasonic.com

ABSTRACT

This paper investigates the use of a large corpus for the training of a Broadcast News speech recognizer. A vast body of speech recognition algorithms and mathematical machinery is aimed at smoothing estimates toward accurate modeling with scant amounts of data. In most cases, this research is motivated by a real need for more data. In Broadcast News, however, a large corpus is already available to all LDC members. Until recently, it has not been considered for acoustic training.

We would like to pioneer the use of the largest speech corpus (1200h) available for the purpose of acoustic training of speech recognition systems. To the best of our knowledge it is the largest scale acoustic training ever considered in speech recognition.

We obtain a performance improvement of 1.5% absolute WER over our best standard (200h) training.

1. INTRODUCTION

1.1. Data and large vocabulary

Speech recognition is essentially a discipline based on a statistical learning machine which relies first and foremost on observed training data. In almost all speech algorithms, there is a tradeoff between quantization error and uncertainty. As the size of the model increases, its modeling power increases to finer grained phenomena, reducing the quantization error. However, very complex patterns must be learned from a large amount of data. It is not easy to distinguish actual detailed structure from accidental correlation. Therefore, the amount of training data must increase. A good speech recognition system is therefore a system which fringes misdeed on both aspects equally: it is as over-training as much as it is grossly approximate.

In that light, speech researchers have been trying to incorporate complexity on the model parametric shape in order to obtain the best tradeoff with relatively small amounts of data. On the other hand, there is a lurking conviction that LVCSR system operates in a wealth of data. It is generally put forward that the community “does not know how to use all these data”, or “HMMs do not scale up”. Firstly, we believe that

LVCSR, like any other domain, could do better with more data. Secondly, we would also like to investigate whether classical training can scale up to a larger training corpus with almost no modification.

1.2. Upwards trend in speech recognition

Speech recognition has evolved significantly in the last 20 years. It is not only due to vigorous algorithmic developments, but we attribute it also to the availability of large amounts of real-life, found data, and the ambition to tackle real problems. The NIST evaluations crystallize a large part of this effort, as shown on Table 1. If the trend had continued past 1998, there would have been more than 2000h, or 83 full days, available for training in 2002. Considering that transcription by humans is rated at about 20-50 times real-time, this would have constituted an enormous investment.

Name	Task	Year	Data	WER
TIMIT	phonetic	1989	1h	N/A
WSJ0	dictation	1992	12h	12%
WSJ0+1	dictation	1993	70h	8%
Hub4	Broadcast	1996	100h	16%
Hub4	Broadcast	1997	200h	13%
Hub5	Conversational	1998	270h	24%
Projected	N/A	2002	2048h	N/A

Table 1. Amount of data for training, by year, used in DARPA tasks. Approximate best Word Error Rates (WER) are also reported.

2. PSTL'S BROADCAST NEWS SYSTEM

Our baseline Broadcast News system follows a classical architecture. We review its main features. In the NIST RT02 evaluation, the system was scored at 20.4% WER.

2.1. Frontend processing

There were 12 MFCC parameters from c_1 to c_{12} . We appended the energy. A five-tap non-causal filter is applied twice to obtain delta coefficients acceleration coefficients. Then, a sliding window average is computed over 2s and removed from all coefficients including energy. No special processing regarding narrow-band speech was performed.

2.2. Acoustic training

LDC97S44 (train96) and LDC98S71 (train97) served as training data. Overlapping speech and music-only segments are discarded. Word-internal triphonic 3-state HMM models are generated using a decision tree state tying procedure. We enforce a final size of 3200 states by setting a minimum likelihood change between parent and aggregate children score. Models were iteratively split and trained up to 128 Gaussians per state. From the resulting 400k Gaussians obtained thereby, we retain only 192000.

Gender-dependent are trained using the ML criterion. Variance were fixed to Speaker-Independent (SI) ones. Unseen Gaussians were left untouched (SI), as they seem to provide background modelling necessary for classification.

2.3. Language model

On Table 2 we show what data were incorporated in LM training. We interpolate counts with the given weight factor. These TDT transcriptions contain *both* rush transcripts and closed captions. The language model includes 53514 words,

Name	Size (M words)	Weight
1996 CSR Hub-4	140	3
North American News	500	1
TDT2 + TDT3	31	3
Acoustic training	1.6	12

Table 2. LM training data: amount and weighting.

19007163 bigrams, and 68189884 trigrams in a standard back-off topology. Backoff mass was held out using Good-Turing formulae.

2.4. Decoding

The BN-STT (Broadcast News Speech To Text) system proceeds in two stages. The first-pass decoding uses gender-dependent models according to the labels provided by the segmentation/clustering step. There were 192000 Gaussians as described previously. The decoder is a fast trigram context-dependent word-internal Viterbi decoder with bigram lookahead and histogram pruning [1].

The most likely transcription is used for MLLR adaptation. Block-diagonal matrices (3 blocks) constitute the transformation. The 7 regression classes were allocated to silence(1), vowels (4), and consonants (2). In degenerate cases we allowed ourselves to reduce the number of classes to three (one for each of silence, vowel, and consonants) or one global class.

In this configuration, the total processing time including I/O and startup time was about 6 times real-time on an Intel Pentium IV Xeon running at 3.06GHz, with 3GB of DDR memory connected on a 533MHz frontside bus.

3. A LARGE CORPUS FOR BROADCAST NEWS

3.1. Corpus description

We propose to inject an additional 1000h of data in the standard 200h training data set. Table 3 summarizes basic information about the corpora. The baseline training corpus includes both Hub4 databases. (We have removed all data recorded in December 1998, from which the test corpus was extracted.) The times indicate raw material. As can be seen from the table, we increase the number of uttered words by one order of magnitude. The Topic Detection and Tracking corpus has been avail-

Name	M words	Hours	LDC audio/text
Hub4 '96	0.6M	97	LDC97S44/97T22
Hub4 '97	0.6M	100	LDC98S71/98T28
TDT2	12M	600	LDC99S84/2001T58
TDT3	9M	480	LDC2001S94/2001T57
Total	23M	1200	N/A

Table 3. Training corpora

able for long time. However, until recently it has never been used to training acoustic models. To the best of our knowledge, no one has ever published a successful attempt at training on all TDT data, or to improve the baseline system using additional TDT data.

3.2. Transcriptions

Unlike Hub4 training data, TDT does not come with accurate verbatim transcriptions. These include speaker identity, sex, sometimes level of background noise, and focus condition. Also, the text depicts exactly what is on the sound file, including filled pauses, breath noises, laughter, dysfluencies and word fragments.

On the other hand, in TDT we find associated rush transcripts or closed captions. We chose to select closed captions when rush transcripts were not available [2]. This approximate text is then normalized. This step converts human-readable text into text suitable for language modeling. For instance, numbers, currency signs, and abbreviations are expanded into corresponding words. Our normalizer was tuned

to minimize the error rate on the careful transcription part of TDT (LDC2000T44). We measured a word error rate of about 15%.

Then, the TDT corpus is decoded using a trimmed down version of our recognizer. The decoder ran within real-time, i.e. fewer than 1000 CPU-hours were required to decode the entire database. It used a spectrum-entropy speech/non-speech detector followed by a cascade of morphological filters. The outcome comprised lattices for further MMI processing, the most likely transcription, and utterance boundaries. Then, transcriptions were filtered using dynamic programming alignment against the normalized captions, followed by an erosion filter. About 70% of the decoded words had a matching entry in the captions. The erosion filter discards words that do not precede or follow a word with a matching caption word. We kept about 50% of the words. It has been observed that misrecognized words usually have a poor time-alignment. Also, they might corrupt cross-word contexts. Finally, the alignment is not reliable in areas of word-error rates, where frequent short words (e.g. “a”) inserted by the language model are matched randomly within the stream. This resulted in about 350’000 utterances.

3.3. Experiments

The lack of data affects almost every component of the speech recognition system. Additionally, these components interact with each other. It would be naive to try to list and study all components exhaustively. The most cited are:

- Decision tree state clustering, and in general tri- or pentaphone contexts, cross-word and word-internal;
- Total number of Gaussians of the system;
- Gender- or condition-dependent modeling;
- Maximum Mutual Information estimation;
- Pronunciation variants.

We have resigned ourselves to restrict our study to only the following topics.

3.3.1. Gender-dependent modeling

Many sites have reported significant improvements with gender-dependent (GD) modeling. In principle, GD models could be trained completely separately from scratch. They need not share decision tree topology, pronunciation variants, or Gaussians.

In practice, due to the lack of data, gender-dependent models are considered as an outgrowth of gender-independent (GI) models. LIMSI adapts GD models from GI models using MAP. Many other sites participating to the evaluation moved to GI in 2002. This indicates that scarcity of data is a major blocking factor to GD modeling.

On Table 4, we report the amount of data used to train GD models. We experimented with several training schemes 5.

Gender	Raw (h)	Filtered(h)
Male	795	478h
Female	293	149h

Table 4. Amount of training data per gender. Our gender classifier makes 8-12% error.

First, we trained models on Hub4 training data only. Then, we tried retraining the means, or means and variances on Hub4 and TDT. We also tried training each gender separately from 1 Gaussian per state and using iterative splitting. Doubling the number of Gaussians from 192k to 384k did help, but resulted in much slower training and decoding. We observed a minor improvement by “mixing” the training data. In addition to gender-specific data, we inserted the full Hub4 training data. It seems to improve models by providing a fall-back background model. All models were trained to convergence. If we remove means that were not seen in a gender during Hub4 training, there is a significant loss. Mixture weights were never

Name	WER
GI	20.0%
GD (means + var)	20.4%
GD (means)	19.6%
GD (seen means)	19.9%
GD (means), 384k	19.7%
GD, tdt-retrain, m	19.4%
GD, tdt-retrain, m+v	18.9%
GD, tdt-retrain, mixed	18.7%
GD, tdt-train	18.9%
GD, tdt-retrain, m+v, 384k	18.5%

Table 5. GD training strategies. Systems use 192k Gaussians unless specified (384k).

adapted.

We have verified that BN models can be indeed improved by the mere addition of data.

3.3.2. Maximum-mutual information estimates

Since Woodland’s breakthrough with MMI [3], many systems have now abandoned GD modeling for MMI. It seems that MMI does not work satisfactorily in setups where the amount of data is scarce.

Table 6 shows the results. Our implementation of MMI builds a state graph from the unigram-weighted word lattice. For performance, lexical trees are built out of all outgoing arcs of each state, and then cached for further reference. Weight pushing was applied for efficient pruning. Times are

GD	TDT	MMI	WER
No	No	No	20%
No	No	Yes	18.8%
Yes	No	No	19.6%
Yes	No	Yes	19.0%
Yes	Yes	No	18.9%
Yes	Yes	Yes	17.3%

Table 6. MMI and GD training

blurred to produce more compact lattices with more modeling potential. The state graph takes into account pronunciation variants. Word graphs with more than 200k word arcs are discarded. Since the denominator works on the full 1200h whereas the numerator is computed on only 700h, accumulators of the denominator were rescaled by a factor of about 0.65.

3.3.3. Pronunciation variants

Many linguistics agree on the fact that words may be pronounced in many more different ways that a single pronunciation lexicon can encompass. We attempt to train lexical pronunciation probabilities in our lexicon. There are 1.22 pronunciations per word in our lexicon on average. In the training set, words which can be pronounced in multiple ways amount to about 28% of the words.

We trained pronunciation variants probabilities by directly observing counts, and applying an absolute discounting scheme with flooring:

$$\hat{c}(q) = \max(30, \#c - 30) \quad (1)$$

where q is a pronunciation variant. Each q appears $\#c(q)$ times in the training corpus. This count was replaced by $\hat{c}(q)$.

The Broadcast News task is a rather peculiar example for pronunciation modeling. Most anchor speakers are trained to speak clearly with a homogeneous pronunciation. In (UK) English, this is generally known as “received pronunciation” or BBC accent. Note also that during training, we use a flattened neg-logprob semiring, i. e.,

$$P(w|o) = \sum_q P(q|w)P^\alpha(o|q)P(w), \quad (2)$$

where α is equal to the inverse of the grammar weight during MMI. During decoding, we use the so-called tropical semiring, or max approximation:

$$P(w|o) \approx \max_q P^{1/\alpha}(q|w)P(o|q)P^{1/\alpha}(w). \quad (3)$$

This approximation is clearly wrong if probabilities are close. We believe that these factors prevented us from clearly studying the effect of pronunciation variants probabilities.

Unfortunately, systems with or without pronunciation variants gave statistically indistinguishable error rates. Both

scored at 17.3% WER, with the same percentage of insertions, deletions, and substitutions.

3.3.4. Scalability Issues

Computational resources have long been a decisive factor in the design of training acoustic models. We review the main characteristics of the training. Times are measured on a cluster of 44 Athlon AMD CPUs running at 1200-1500 MHz, each with 512 MB of RAM.

Hub4 MLE training (200h)	1 day
TDT decoding	2.5 days
TDT MLE re-training (1200h)	1 day
MMI TDT training	5 days
Hub4: Number of utterances	123k
TDT: Number of utterances	469k

Table 7. Orders of magnitude with different corpora

4. CONCLUSION AND FURTHER WORK

We have described PSTL’s Broadcast News recognition system. We have shown that a significant gain in performance can be achieved by simply adding more data to the standard training procedure. This is verified explicitly with GD and MMI experiments. Preliminary experiments with pronunciation modeling seemed to indicate that pronunciation variant probabilities are not useful in this task.

The additional corpus was obtained by filtering the TDT corpus. Careful normalization and filtering rules helped dynamic programming matching of the closed caption against recognition hypotheses. The new TDT4 corpus will be incorporated when it becomes available.

We focused on studying the effect of enlargement of training on classical training. We have verified that classical training can scale up. We believe that there is much more to gain by leveraging more data greedy algorithms.

5. REFERENCES

- [1] P. Nguyen, L. Rigazio, and J.-C. Junqua, “EWAVES: an efficient decoding algorithm for lexical tree based speech recognition,” in *Proc. of ICSLP*, Beijing, China, Oct. 2000, vol. 4, pp. 286–289.
- [2] L. Lamel, J.-L. Gauvain, and G. Adda, “Lightly supervised and unsupervised acoustic model training,” *Computer Speech and Language*, vol. 16, no. 1, pp. 115–229, 2002.
- [3] P. C. Woodland and D. Povey, “Large Scale MMIE Training for Conversational Telephone Speech Recognition,” in *NIST Speech Transcription Workshop*, 2000.
- [4] M. Saraclar, M. Riley, E. Bocchieri, and V. Goffin, “Towards Automatic Closed Captioning: Low Latency Real-Time Broadcast News Transcription,” in *ICSLP*, 2002, pp. 1741–1744.
- [5] P. Nguyen, L. Rigazio, Y. Moh, and J.C. Junqua, “Rich Transcription 2002 Site Report, Panasonic Speech Technology Laboratory (PSTL),” 2002.