Fast and effective retraining on contrastive vocal characteristics with Bidirectional Long Short-Term Memory nets

Nicole Beringer

3SOFT GmbH, Erlangen, Germany

Nicole.Beringer@3SOFT.de

10.21437/Interspeech.2006-235

Abstract

We apply Long Short-Term Memory (LSTM) recurrent neural networks to a large corpus of unprompted speech - the German part of the VERBMOBIL corpus. By training first on a fraction of the data, then retraining on another fraction, we both reduce time costs and significantly improve recognition rates. Contrastive retraining on the initial vowel cluster fraction of the data according to the Psycho-Computational Model of Sound Acquisition (PCMSA) shows higher frame by frame correctness due to more sparseness and the articulatory position of the sounds. For comparison we show recognition rates of Hidden Markov Models (HMMs) on the same corpus, and provide a promising extrapolation for HMM-LSTM hybrids.

Index Terms: Bidirectional Long Short-Term Memory, recurrent neural networks, retraining on data fractions, Psycho-Computational Model of Sound Acquisition.

1. Introduction

The human brain simultaneously filters all important information out of different aspects, e.g. acoustics, prosody and phonotactics, and adapts and compares this information to its inherent speech database. Without any prior knowledge of the grammatical rules of a language or a predefined language model[1] the human brain is capable of learning statistic regularities within speech and can easily adapt its internal representations of the regularities once the statistics change.

Most current Automatic Speech Recognition (ASR) systems don't handle all important information at a time. They use submodules concentrating on one aspect of speech, e.g. providing the phonotactics of a language in a Language Model, when building acoustic models or providing prosodic models for an extra prosodic recognizer.

It would be desirable to retrain an Automatic Speech Recognition (ASR) system on new data without losing the benefits of previous learning. For example, it may be necessary to adapt quickly to new input, or to use information gained from a previous task, e.g., recognizing read speech, in order to solve the next task, e.g., quasi-spontaneous (= unprompted) speech. In task/domain independent recognition, systems that are (pre-)trained under certain conditions and/or certain dialogue specifications are required to adapt to utterances recorded under different conditions or with different dialogue specifications. It has also become standard practice to train Hidden Markov Models (HMMs) on multiple corpora, in order to improve their robustness also with respect to new data. However, methods for adapting HMM's are complex and time-consuming [1]. Most modern systems use a hybrid of HMMs and maximum likelihood linear regression to adapt to new training material.

Artificial Neural Networks (ANNs) allow a more brainlike approach: learn from scratch and let the learning process decide which information is important and how it has to be connected to other already learned information. An ANN approach works quite well for a small set of unfiltered information but dealing with large speech corpora, ANNs seem to be overtaxed when trying to learn all information out of different aspects of speech. How can we benefit from the ANN brainlike approach also for a large amount of information? Is the division of a large speech corpus in small subsets of unfiltered information a solution? How can we sequentially add the learned information out of these different subsets without loosing the already learned regularities?

In fact, ANNs lend themselves to a very simple form of retraining: train on one dataset, then continue training on another without resetting the weights. Recurrent Neural Nets (RNNs) are particularly promising for speech processing because they have the potential to learn a dynamic model of speech that incorporates multiple time scales without using time windows or fixed time delays. Unlike traditional RNNs, Long Short-Term Memory nets (LSTM) [2] can also handle long time lag correlations between inputs and errors, also in the context of speech applications [3]. They have already been successfully used for speech applications. Recent experiments with plain LSTM on speaker adaptation [4] suggest that retraining is fast and effective on small corpora, and that results of previous learning and generalization improve with retaining on randomly chosen subsets of the data. Recently, we applied this approach to Bidirectional LSTM [5] and a large corpus of unprompted speech [6] with randomly chosen subsets. Retraining seems a reasonable solution of handling unfiltered information of large speech corpora. But is there also any brainlike approach to form the subsets for sequential retraining on different aspects of speech?

A solution for the forming of subsets is to simulate the human sound acquisition as we recently proposed in the Psycho-Computational Model of the Sound System Acquisition (PCMSA) [7]. In this paper we retrain on a vowel subset of the VERBMOBIL corpus - the initial vowel contrast of the PCMSA - and compare recognition results to a randomly chosen subset. The following section briefly describes the VERBMOBIL data used for both LSTM and HMM experiments. Section 3 gives an overview of LSTM. Section 4 describes the experimental setup. Section 5 summarizes the aspects of the Psycho-Computational Model of the Sound System Acquisition. Section 6 analyses the experimental results of baseline and retrained LSTM for framewise phoneme prediction on a randomly chosen subset and the initial PCMSA vowel subset. Section 7 provides an extrapolation of the frame-

[1] Meant are here the grammatical rules which can be found in grammar books and the kind of language models used in ASR. Both are defined by humans. We do not refer to Chomsky's Universal Grammer.

based results for a HMM-LSTM hybrid based on previous comparisons of framewise and phoneme error rates on various corpora of read speech.

2. Corpus description

Our present investigation uses a database of *unprompted* speech-the VERBMOBIL (VM) corpus [8].The VM corpus is divided into VM1 and VM2. Both sets differ in recording conditions and tasks. The corpus consists mainly of three language portions: German, American English and Japanese. The German VM portion contains sufficient speech data for training and testing (35136 turns[1]). For this study only the German portion was used. The database-scenario deals with scheduling appointments with a business partner: real-life-situations with currently used speech. The "formal situation" setup ensures that speech contains fewer and weaker regional variants than it would contain if personal affairs were discussed.

The training (train), development (dev) and test (test) sets currently used in our experiments on the VM corpus contain the following constraints [6]: each speaker is allowed in only one set (hard constraint), for each speaker there must be at least one complete dialogue (to allow speaker adaptation algorithms to be applied; hard constraint), speakers should be distributed equally across sexes in all sets (soft constraint), recordings should be distributed equally across recording sites in all sets (to cover possible accents preferences in one site; soft constraint).

The HMM system uses the full data for training and testing. The LSTM classification network uses only one fourth of the training set in its baseline training, another fourth of the training set is used for retraining. The full test set is used as described above.

The HMM phone recognizer was built up with the Hidden Markov Toolkit [9]. It uses the above defined subsets and a bigram trained solely on the training corpus.It was tested on the development sets with the corresponding lexicon (total: 5540 lexical entries). The acoustic models are based on 12 Standard MFCC + Energy + velocity + acceleration (39), Diagonal covariance matrices, 3-5 states per phoneme, 43 phoneme classes (extended German SAMPA) + garbage + voice garbage + silence + laugh + breath (48), Models initialized using the Munich Automatic Segmentation (MAU) tier of the BAS Partiture Format (BPF) from 1/4 of TRAIN, Re-estimation and splitting mixtures after 6 iterations on total TRAIN, testing after every two iterations on DEV, weight of language model fixed to 6.5; beam search width 100.0.

Table 1 shows the recognition results of a plain HMM phone recognizer which was trained both on monophones and triphones (also across words).

Table 1: For comparison: Phoneme error rate for plain HMMs

System	training set size	phoneme error rate on the test set	epochs
Monophone	full	34.29%	52
Triphone crossword	full	35.49%	37

Monophones contain 512 Gaussian mixtures per state. Triphones have the same number of parameters as the monophone system, 8 mixtures per state and are trained also across word boundaries. HMMs were trained on the full training set.

3. LSTM

"Long Short-Term Memory" [2] is a general purpose algorithm for extracting statistical regularities from noisy time series. It learns from scratch, typically with more adjustable parameters (the weights), a larger search space, and less initial bias [10] than HMMs, which incorporate prior linguistic knowledge.

[1] One turn in the VM database has about 22.8 words in average.

3.1. Bidirectional LSTM (BLSTM)

The output of typical RNNs is based on the complete history of *previous* inputs. However, there are many sequence processing tasks where future inputs are also useful because reverse correlations exist. In speech, for example, the articulatory system is already preparing future utterances as it shapes the current one. A solution is bidirectional training [11, 12]: the input is presented forwards and backwards to two separate recurrent nets, both of which are connected to the same output layer. In this way, errors can be injected as normal and backpropagated through the nets. Current results with BLSTM [5] show that it outperforms normal LSTM, as well as previous bidirectional RNNs on speech recognition tasks.

3.2. Retraining with bidirectional LSTM

An in-depth investigation of retraining with LSTM [4] (i.e. presenting new data to an already trained network) showed that LSTM is capable of fast and effective relearning on speakers with widely varying vocal characteristics. The net was trained and successively retrained on disjoint subsets of the TIDIGITS database. The retraining time and difficulty diminished with repetition, and the net was able to transfer knowledge across several datasets. The final performance of the net was generally raised by having been previously trained on different datasets, and this improvement persisted over multiple retrainings.

4. The Psycho-Computational Model of the Sound System Acquisition (PCMSA)

Recently, we developed a Psycho-Computational Model of the Sound System Acquisition (PCMSA) [7]. This is based on studies in first language acquisition [13, 14, 15, 16, 17, 18] which claim that children start using their sound system with the onset of vocal gestures - sounds or sound sequences produced with consistency by the child in different situations.

Vowel contrasts start between the open front /E/ and the closed front /I/ followed by a third degree of opening or by a back /u:/ and finish by /e/ vs. /E/ vs /E@/. First only /i:/ and /E/ are distinguished, in a second step we find the distinction also with /a/ and /u:/, the third to fifth step deal with distinguishing phonetically closer phonemes until the full vowel distribution in the last rectangle is reached.

The consonantal system is developed with regard to **Manner of articulation:** (Distinctions are mostly within category, i.e. within nasals, within plosives, rather than cross-category, e.g. fricatives vs. plosives/nasals.), **Place of articulation:** (Generally labial-apical contrasts occur before contrasts with velars. Distinction between apicals are last to be developed.), and **Voicing:** (Given the same manner and place of articulation voicing occurs when the vocal cords in the larynx vibrate. In the language acquisition task children first learn the voiceless counterparts in most languages.). Generally, it can be said that the more back the place of articulation the later the acquisition of the sound. Also, according to the voiced-voiceless contrast it can be seen that voiceless counterparts are learnt after the voiced sounds and that nasals and liquids are learned before plosives, affricates and fricatives (in this order!). Also the distinction between perceptual similar sounds (usually acoustically similar) is learned quite late.

5. Experimental setup

Preliminary experiments with LSTM standard nets with 25, 50, 100 and 200 blocks (2 cells each) showed that although the duration of the epochs doubled each time, comparable results occurred in far fewer epochs. Nevertheless all experiments converged at around 50% framewise phoneme correctness. When comparing LSTM bidirectional nets to standard nets with comparable weights (50 000) we found that BLSTM needs less epochs to obtain comparable results to standard nets and reaches higher framewise phoneme correctness (58.87%). Both bidirectional and standard nets reach their peak around the 120th epoch.

5.1. Randomly chosen subsets for BLSTM training

Based on these findings we used a two-step retraining procedure as follows: LSTM training and retraining sets were each around 1/4 of the whole VM training set. Both training and retraining set are distinct from each other but were randomly chosen from the whole training set. The whole VM test set was used. The framewise phoneme error rate was calculated for all sounds of the test set and for the vowels of the PCMSA subset. Our bidirectional LSTM network contained two hidden LSTM layers (for the forward and reverse nets), each with 200 blocks of 2 cells. It had 26 input nodes and a softmax output layer containing 52 nodes. A cross entropy objective function was used. The input layer was connected to the hidden layers, both of which were connected to themselves and to the output layer. There were 907112 weights in total. Note that unlike HMMs BLSTM has no structural bias and more weights - a disadvantage according to the bias-variance dilemma [10].

5.2. PCMSA vowel contrasts

The BLSTM net described above was also used to train and retrain on the PCMSA subset. Preliminary experiments on a smaller BLSTM net and single contrastive training on consonants and vowels showed that the more periodical structure a sound has, the easier to recognize for BLSTM. Therefore, we decided to concentrate on the initial vowel cluster of the human vowel acquisition and trained our system subsequently on /E/ vs /i:/ according to section 4:. In the experiment we trained on the single vowels and retrained on the vowel contrasts described in [7] until the frame by frame error on the development set improved insignificantly. Both training and development subsets are a portion of the training and development sets described in section 2.

6. Experimental results

Our experiments are divided into two main parts: The first gives the plain LSTM classification for frame by frame recognition results. Part two shows the frame by frame recognition results of the consecutive single vowel training and retraining on the PCMSA vowel contrast and the frame by frame recognition results for the PCMSA vowel contrast phonemes trained on the randomly chosen subset. Both systems use the same test set. Table 2 shows that BLSTM retraining led to a 5% improvement on the full test set. Using 1/4 of the training set at a time greatly reduces total training time.

Table 2: Recognition results: frame by frame phoneme error rate for plain BLSTM

System	training set size	frame by frame phoneme error rate on the test set	epochs
baseline	1/4[2]	38.40%	50
retraining	1/4[2]	33.36%	67

Table 3 shows the main results of the PCMSA vowel contrast cluster and the randomly chosen subset of the BLSTM net.

Table 3: Recognition results: frame by frame phoneme error rate for PCMSA vowel contrast

System	training set size	frame by frame phoneme error rate of /E/ - /i:/ on the test set	epochs
baseline	sequential[3]	29.60%	74
retraining	contrast[3]	21.75%	71
plain BLSTM	1/4[2]	37.70%	67[4]

As can be seen in table 3 already a consecutive training on all sounds of the PCMSA initial vowel contrast cluster in the baseline shows an improvement of 8.10% on the test set compared to the randomly retrained BLSTM. Note that the test set in both cases is the same subset of the test set described in 2 which includes only the vowels of the vowel contrast cluster /E/ - /i:/. Retraining on the PCMSA vowel contrast cluster results in another 7.85% reduction of the frame by frame phoneme error rate.

7. Predicting the phoneme error rate: an extrapolation for a HMM-LSTM hybrid approach

Although we cannot compare the framewise phoneme error of BLSTM directly with the phoneme error of the HMM we expect that a BLSTM-HMM hybrid (under construction) will outperform both plain BLSTM on frame by frame and plain HMMs on the phoneme level, inheriting the best of both worlds, namely reduction of training material and training time (BLSTM), as well as more built-in structural bias (HMMs). This expectation is encouraged by experiments on **read speech** by Chen and Jamieson [19], Shire [20], Waterhouse, Kershaw and Robinson [21], and Elenius and Blomberg [22]. They all achieved better results on the phoneme level using an ANN-HMM hybrid approach, as shown in table 4 for framewise and phoneme error rates for several systems on various corpora. *improvement factor* shows the relative ratio of framewise and phoneme error. *LIN* stands for Linear Input Network, *MLIN* for Mixtures of LINs for adaptation (*_2* = 2 experts; *_4* = 4 experts). *MLP* stands for Multilayer Perceptron nets[5].

As can be seen from table 4 the framewise errors are quite high for noisy input sequences (several microphones or enriched with background noise) as opposed to clean speech. The HMM part of the hybrids is able to drastically reduce the error on the

[2]The portion of the baseline was randomly chosen. The portion for retraining is distinct from the baseline portion

[3]The portion of the baseline includes only /E/ and /i:/. Training was sequential. The portion for retraining is distinct from the baseline portion. It includes the vowel contrast cluster /E/ - /i:/ according to PCMSA.

[4]The underlying BLSTM system is the retrained BLSTM of table 2

[5]MLPs are supervised feedforward neural networks trained with the standard backpropagation algorithm. With one or two hidden layers, they can approximate virtually any input-output(= the desired response) map. They are widely used for pattern classification and can approximate the performance of optimal statistical classifiers in difficult problems.

[6]Swedish speakers

[7]MUM[3] Task

[8]clean speech: NUMBERS95

[9]clean sp. no border: NUMBERS95

[10]factory noise: ARPA 1995 H3 multiple unknown microphones

[11]factory noise no border: NUMBERS95

[12]TIMIT

Table 4: Framewise and phoneme errors on read speech corpora

System	frame (plain ANNs)	phoneme (ANN-HMM hybrids)	improvement factor
Backprop[22][6]	30.0%	24.5%	1.22
RNN 0 pass [21][7]	22.8%	18.1%	1.26
LIN 1 pass [21][7]	20.1%	16.5%	1.22
LIN 2 pass [21][7]	19.9%	15.9%	1.25
MLIN_2 1 pass [21][7]	19.2%	16.5%	1.16
MLIN_2 2 pass [21][7]	18.9%	16.1%	1.17
MLIN_4 2 pass [21][7]	18.2%	15.8%	1.15
MLIN_4 3 pass [21][7]	18.0%	15.7%	1.15
MLP[20][8]	28.97%	7.3%	3.97
MLP[20][9]	29.80%	7.7%	3.87
MLP[20][10]	42.84%	15.5%	2.76
MLP[20] [11]	42.88%	15.0%	2.86
RNN [19] [12]	26.3%	20.21%	1.30

phoneme level due to structural bias of the HMM. This means that on unprompted speech with background noise, speaker overlaps and other perturbations we can expect a much lower phoneme error. With the **worst** improvement factor (1.15) of table 4 we can conservatively predict a phoneme error rate of 29.01% for a retrained BLSTM-HMM hybrid on VERBMOBIL (33.39% for the standard BLSTM respectively). An optimistic calculation with the **best** improvement factor (3.97) for read speech in table 4 would give us 8.4% for the retrained BLSTM-HMM hybrid (9.67% for the baseline respectively). If we do a similar extrapolation to predict the phoneme error rate of the PCMSA driven subset, the **worst** improvement factor (1.15) of table 4 would result in a conservative prediction of the phoneme error rate of 18,91% for a retrained PCMSA-BLSTM-HMM hybrid (25,74% on the baseline PCMSA-BLSTM-HMM hybrid, 32.78% for the randomly retrained BLSTM respectively). The **best** improvement factor (3.97) for read speech in table 4 would result in optimistic prediction of 5.46% for the retrained PCMSA-BLSTM-HMM hybrid (7.46% for the baseline PCMSA-BLSTM-HMM hybrid, 9,50% for the randomly retrained BLSTM respectively).

8. Conclusions and outlook

We examined the retraining ability of LSTM recurrent nets in a frame by frame phoneme classification task of unprompted speech. We compared recognition results of a normally trained BLSTM system to those of a retrained one. We adapted the experiment by applying human language acquisition to BLSTM retraining. Retraining both significantly reduced time costs and training set size and improved recognition results. The contrastive BLSTM retraining on the initial PCMSA vowel cluster showed that PCMSA seems to be a reasonable method to make ANNs more sparse. An extrapolation based on read speech promises significant additional improvements on the phoneme level through a BLSTM-HMM hybrid. Future work extends a BLSTM-HMM hybrid on a contrastive PCMSA retraining on the whole phonemic representation of a language.

9. Acknowledgements

This work was developed as part of my postdoctoral research at IDSIA supported by the SNF (grant number 200020-100249).

10. References

[1] John McDonough and Alex Waibel, "Performance comparisons of all-pass transform adaption with maximum likelihood linear regression," *Proc. ICSLP*, 2004.

[2] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, 1997.

[3] A. Graves, D. Eck, N. Beringer, and J. Schmidhuber, "Biologically plausible speech recognition with LSTM neural nets," *Proc. Bio-ADIT*, 2004.

[4] A. Graves, N. Beringer, and J. Schmidhuber, "Rapid retraining on speech data with lstm recurrent networks," Tech. Rep. IDSIA-05-05, 2005.

[5] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm networks," 2005.

[6] N. Beringer, A. Graves, F. Schiel, and J. Schmidhuber, "Classifying unprompted speech by retraining lstm nets," *W. Duch et al. (Eds.): Networks ICANN05, LNCS 3696*, 2005.

[7] N. Beringer, "Human language acquisition in a machine learning task," *Proc. ICSLP*, 2004.

[8] K. Weilhammer, F. Schiel, and U. Reichel, "Multi-Tier annotations in the Verbmobil corpus," *Proc. LREC*, 2002.

[9] S. Young, *The HTK Book*, Cambridge University Press, 1995.

[10] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Computation*, 1992.

[11] Mike Schuster and Kuldip K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, 1997.

[12] Jinmiao Chen and Narendra S. Chaudhari, "Capturing long-term dependencies for protein secondary structure prediction," in *Proc ISNN 2004*, 2004.

[13] J. Dore, M.B. Franklin, R.T. Miller, and A.L.H Ramer, "Transitional phenomena in early language acquisition," *Journal of Child Language*, 1976.

[14] L. Menn, "Development of articulatory, phonetic and phonological capabilities," *Butterworth, B. (ed), Language Production, Academic Press*, 1983.

[15] M.A.K. Halliday, *Learning how to mean: Explorations in the Development of Language*, Cambridge University Press, 1987.

[16] C.A. Ferguson, "Ferguson, c. a. learning to pronounce: the earliest stages of phonological development," *Minifie, F. D., Lloyd, L. (eds) Communicative and Cognitive Abilities: Early Behavioural Assessment, University Park Press*, 1976.

[17] R. Jakobson, *Child language, aphasia and phonological universals*, Mouton, The Hauge, 1968.

[18] N. V. Smith, *The acquisition of phonology: a case study*, Cambridge University Press, 1973.

[19] R. Chen and L. Jamieson, "Experiments on the impementation of recurrent neural networks for speech phone recognition," *Proc. Thirtieth Annual Asilomar Conference on Signals, Systems and Computers*, 1996.

[20] M. Shire, "Relating frame accuracy with word error in hybrid ann-hmm asr," *Proc. EUROSPEECH*, 2001.

[21] S. Waterhouse, D. Kershaw, and T. Robinson, "Smoothed local adaptation of connectionist systems," *Proc. ICSLP*, 1996.

[22] K. Elenius and M. Blomberg, "Comparing phoneme and feature based speech recognition using artificial neural networks," *Proc. ICSLP*, 1992.