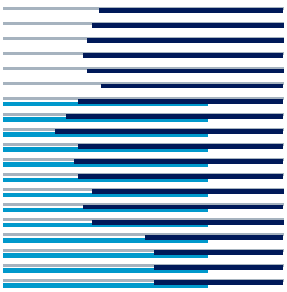
Human Language Acquisition methods in a Machine Learning Task

Nicole Beringer



Technical Report No. IDSIA-19-04
October 2004

IDSIA / USI-SUPSI
Dalle Molle Institute for Artificial Intelligence
Galleria 2, 6928 Manno, Switzerland

This work was published on ICSLP 04 and is part of the conference proceedings.

Human Language Acquisition methods in a Machine Learning Task

Nicole Beringer

October 2004

Abstract

The goal of this study is to develop a psycho-computational model of human phoneme acquisition that includes the knowledge of linguistic universals [1, 2, 3] to "teach" Artificial Neural Nets incrementally. Long Short-Term Memory (LSTM) artificial neural networks are capable to outperform previous recurrent networks on many tasks ranging from grammar recognition to speech [4] and robot control [5]. Together with our psycho-computational model they are supposed to recognize phonetic features in a way similar to humans learning to understand their first language.

1 Introduction

For many applications in speech processing, such as in ASR and speech synthesis (e.g. PSOLA), reliable segmentation and labeling of large speech databases is required. Also, as ASR increasingly uses pronunciation modeling [10, 11, 6, 9, 8, 7] the demand for statistically based pronunciation models in different languages is growing. To segment or recognize databases of a language, we usually have to (mostly statistically) train the system on the relevant acoustic models, which is an expensive and time-consuming process, because a huge amount of data is needed.

Also, these training procedures are more or less "external" approaches (like logical top down/bottom up processing or statistical Hidden Markov Modelling) in that they differ from the way the human brain acquires speech.

All contemporary results in speech processing - especially speech recognition - show that neither the logical nor the statistical approaches will fully succeed. Their way of adapting to the data usually does not consider the biological way of speech processing, namely the processing of speech in human brains, but instead just try to adapt on noise, gender, dialects etc..

Unfortunately, neural nets, which were supposed to be more biological, also failed in the first attempt not only because they had problems with long time lags but also because they were still too "external", i.e. they did not consider how the human brain acquires speech. Using bottom up processing tended both to blow up training time and to cause confusion between too many possible recognition options.

Knowing about this lack of similarity with human speech processing we analyzed human language acquisition. Our goal is now to provide a psycho-computational model of language acquisition with main focus on the acquisition of the sound system, which can be used to train the Long Short-Term Memory (LSTM) artificial neural net iteratively.

Based on two main principles to improve second language acquisition and children's sound acquisition we now can simulate this biological process to LSTM. Looking at human second language acquisition it is known that using the "Pimsleur language learning system" [12] (see section

3), which is based on the way children learn their mother tongue, results in an incredible second language capacity after a very short learning time.

In a second step our goal is to simulate these human principles of language acquisition incrementally with recurrent neural nets. Concretely, we have to combine the advantages of LSTM artificial neural networks, which allow the filtering of relevant information out of noisy time series[1] [4] and which are able to learn unsupervised [5], with the human language acquisition principle.

The following section briefly describes the principle of LSTM networks. Section 3 deals with the "Pimsleur language learning system". Section 4 gives a detailed description of the psycho-computational model of the sound system acquisition. Conclusions and future work are discussed in the last section.

2 Long Short-Term Memory (LSTM) artificial neural networks

"Long Short-Term Memory" [13, 14] is a general purpose algorithm for extracting statistical regularities from noisy time series. LSTM networks are novel RNNs that overcome the fundamental problems of traditional RNNs, and efficiently learn to solve many previously unlearnable tasks like recognition of temporally extended patterns in noisy input sequences, recognition of simple regular and context free and context sensitive languages [15], recognition of the temporal order of widely separated events in noisy input streams, extraction of information conveyed by the temporal distance between events, stable generation of precisely timed rhythms, smooth and non-smooth periodic trajectories, robust storage of high-precision real numbers across extended time intervals, reinforcement learning in partially observable environments [5], metalearning of fast online learning algorithms [13, 16], music improvisation and music composition [17], incremental speech processing [4].

Our hope is that the advantages of the LSTM networks in unprompted speech as well as the handling of contextual effects within the speech data set (the corpus) - in this case the perceptionally critical clusters - will recognize even utterances that humans have perceptual difficulties.

3 Pimsleur Language Learning System

The Pimsleur language learning system (PLLS) [12] is a language acquisition method based on four main principles:

- Anticipation - getting and keeping your attention.
- Graduated interval recall - presenting information at the right time makes it easier to retrain.
- Organic learning - vocabulary, pronunciation and listening comprehension are presented all at once.
- Learning like a child (by imitating others) - reproducing what you hear others say.

These principles make sure how the brain naturally stores the language in long term memory. The mental mechanism converts unintelligible human sound into language. It is also important to know how the brain recreates the language as speech (articulation) and how the words spoken by others are recognized and understood (comprehension).

Once the sound acquisition is modelled, the PLLS (except the first item) is important to implement our Psycho-Computational Model of the Sound System Acquisition to our LSTM network.

[1] 2% WER on the TIMIT corpus could be reduced to 0.5% after a retraining of only six minutes.

4 Psycho-Computational Model of the Sound System Acquisition

Studies in first language acquisition 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 [18, 19, 20, 21].

4.1 Vowels

Following the work of Jakobsen [22] and Smith [23], 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@/.

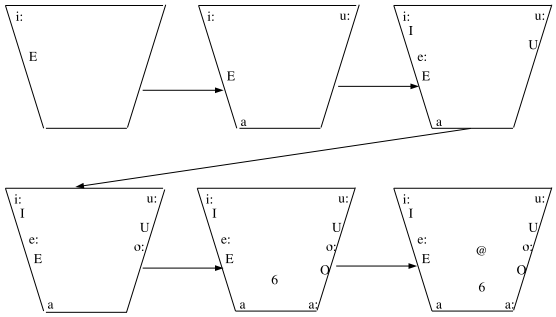


Figure 1: *Vowel distinction. Starting from the leftmost vowel rectangle the acquisition ends with the full vowel rectangle in the right front corner.*

Figure 1 shows the vowel acquisition from the distinction /E/ - /I/ until the full vowel system is reached. As can be seen from the figure, 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.

4.2 Consonants

The consonantal system is developed with regard to:

- **Manner of articulation:** early words mostly contain an open syllable consisting of a consonant and an open vowel. Distinctions are mostly within category, i.e. within nasals, within plosives, rather than cross-category, e.g. fricatives vs. plosives/nasals. The distinction between fricative sounds are generally last to be developed. Generally, one frictionless continuant phoneme like /l/ develops.
- **Place of articulation:** Apart from the manner of articulation also the place plays an important role in language acquisition. In this sense generally labial-apical contrasts occur before contrasts with velars. Distinction between apicals are last to be developed.

- **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.

Figure 2 shows the acquisition of the consonantal system as described above. Generally, it can be said that the more back the place of articulation the later the acquisition of the sound. It is shown by the third dimension in the figure. 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!).

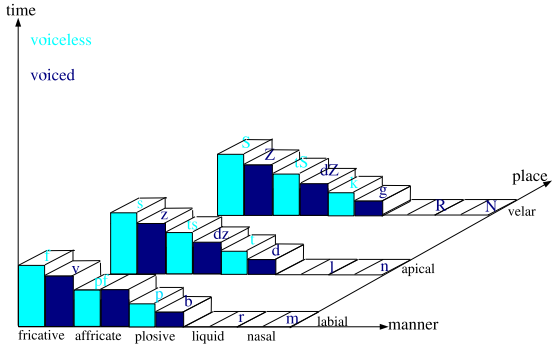


Figure 2: *Consonantal distinction: The higher the column and the more back the place of articulation the later the acquisition.*

Also the distinction between perceptual similar sounds (usually acoustically similar) is learned quite late.

4.3 Prosody

Apart from the level of segmental sounds human language acquisition also requires the development of prosodic features:

- **Intonation:** The contour of the pitch pattern; rising or falling tone at the end of the pattern. Often, intonation is just a co-incidence of voice qualities.
- **Stress:** Where the main accent occurs - which word(s) have emphasis? Children are stress sensitive. Phonetic forms may be generated in a different way, e.g. dropping unstressed syllables before the main accent.
- **Rhythm:** How the words are grouped together. In children's speech there occurs a syllable timed rhythm where each syllable receives an equal amount of time.

It often occurs that the initial consonant cluster is shortened to monoconsonantal.

4.4 The Psycho-Computational Model applied to LSTM

Human language acquisition can now be simulated with LSTM. In a first step the training material has to be adapted to the way humans learn to understand speech, i.e. the above presented psycho-computational model of the sound system acquisition has to be applied.

This means we have to present the acoustic information stepwise (PLLS Graduated intervall recall) according to our Psycho-Computational Model which reflects the PLLS principle "Learning like a child".

The most obvious step is to exploit the linguistic universals [1, 2, 3]: train a speech recognizer on the most general speech sounds (i.e. easiest to understand for humans), which are vowels and other continuous sounds starting from the central vowels unless the full distinction of the vowel rectangle in Figure 1 is reached. Then retrain it on the consonants according to the above model. If we use a speech recognizer that does allow for incremental learning, then retraining should be much faster than training from scratch, due to the common regularities.

LSTM already clearly outperformed previous recurrent networks on many sequential processing tasks ranging from grammar recognition to robot control [5]. Also, LSTM seems ideal to follow the Pimsleur language learning principle: the retraining process follows the graduated intervall recall.

Former studies [4] have shown that retraining a trained net on new data results in a very large reduction of the error rate after some minutes[2]. In this sense, adapting the trained net to the more difficult sounds (consonants) in the time range humans learn to distinguish consonantal sounds shown in Figure 2 is promising.

5 Conclusions and Outlook

We discussed language acquisition methods in general and developed a psycho-computational model which applies human language acquisition to machine learning. We presented the different stages of the acquisition of the sound system and how it is to be implemented to LSTM.

Future work consists of implementing also prosodic features in the LSTM network. Further, the psycho-computational model for the sound system acquisition is going to be tested on a multilingual database with LSTM.

References

[1] N. Chomsky: Om sproket. Problem och perspektiv (orig.tit. Reflections on language). Norstedts, 1978.

[2] N. Chomsky: Lectures on government and binding. Dordrecht: Foris, 1982.

[3] N. Chomsky: Knowledge of language: its nature, origin and use. Praeger, 1986.

[4] A. Graves, D. Eck, N. Beringer, J. Schmidhuber: Biologically Plausible Speech Recognition with LSTM Neural Nets. *Proc. Bio-ADIT*, 2004.

[5] B. Bakker: Reinforcement learning with Long Short-Term Memory. *Advances in Neural Information Processing Systems*, 14, 2002.

[6] N. Beringer: Rule-based categorial analysis of unprompted speech - a cross-language study. *Proc. PaPI Conference*, 2003.

[7] N. Beringer, M. Neff, T. Ito: Generation of pronunciation rule sets for automatic segmentation of American English and Japanese. *Proc. ICSLP*, 2000.

[2] 2% WER of an LSTM network pre-trained on a subset of the TIMIT corpus could be reduced to 0.5% on another subset after just a few minutes of retraining

[8] N. Beringer, M. Neff: Regional pronunciation variants for automatic segmentation; *Proc. LREC*, 2000.

[9] N. Beringer, F. Schiel: Independent Automatic Segmentation of Speech by Pronunciation Modeling. *Proc. ICPhS*, pp. 1653-1656, 1999.

[10] N. Beringer, F. Schiel, P. Regel-Brietzmann: German Regional Variants - A Problem for Automatic Speech Recognition? *Proc. ICSLP*, Vol. 2, pp. 85-88, 1998.

[11] K. Ma, G. Zavaliagkos, R. Iyer: Pronunciation Modeling for Large Vocabulary Conversational Speech Recognition. *Proc. ICSLP*, Paper No. 866, 1998.

[12] P. Pimsleur: Testing foreign language learning. *A. Valdman (Ed.), Trends in Language Teaching*, McGraw-Hill, pp. 175-214, 1966.

[13] S. Hochreiter, J. Schmidhuber: Long Short-Term Memory. *nc*, Vol. 9, No. 8, pp. 1735-1780, 1997.

[14] F. A. Gers, J. Schmidhuber: Long Short-Term Memory learns simple context free and context sensitive languages. *Proc. IEEE TNN*, 2001.

[15] F. Gers, J. Schmidhuber, F. Cummins: Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, Vol. 12, No. 10, pp. 2451-2471, 2000.

[16] A. S. Younger, S. Hochreiter, P. Conwell: Meta-Learning with Backpropagation. *Proc. IEEE IJCNN*, pp. 2001-2006, 2001.

[17] D. Eck: Finding Downbeats with a Relaxation Oscillator. *Psychological Research*, Vol. 66, Nr. 1, pp. 18-25, 2002.

[18] J. Dore, M.B. Franklin, R.T. Miller, A.L.H Ramer: Transitional phenomena in early language acquisition. *Journal of Child Language 3:*, pp. 13-28. 1976.

[19] L. Menn: Development of articulatory, phonetic and phonological capabilities. *Butterworth, B. (ed)*, Language Production vol 2, Academic Press pp. 3-50, 1983.

[20] M.A.K. Halliday: Learning how to mean: Explorations in the Development of Language. *Edward Arnold*, 1975.

[21] 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, pp. 273-97, 1976.

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

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