NETWORK TRAINING FOR CONTINUOUS SPEECH RECOGNITION[1]

Issac Alphonso and Joseph Picone

Institute for Signal and Information Processing, Mississippi State University
{alphonso, picone}@isip.msstate.edu

ABSTRACT

The standard training approach for a hidden Markov model (HMM) based speech recognition system uses an expectation maximization (EM) based supervised training framework to estimate parameters. EM-based parameter estimation for speech recognition is performed using several complicated stages of iterative reestimation. These stages are heuristic in nature and prone to human error. This paper describes a new training recipe that reduces the complexity of the training process, while retaining the robustness of the EM-based supervised training framework. This paper show that the network training recipe can achieve comparable recognition performance to a traditional trainer while alleviating the need for complicated systems and training recipes for spoken language processing systems.

1. INTRODUCTION

Standard hidden Markov model (HMM) based speech recognition systems typically use a forced alignment stage during training that produces a single phonetic transcription for an utterance [1]. This phonetic transcription is then used to guide the parameter reestimation process. To optimize performance, one typically generates this transcription several times using an iterative approach involving the best model set available at that stage of the iteration. To a naïve user of speech recognition technology, the need for an automatically generated phonetic transcription of the data appears to create a "chicken and egg" problem when developing a new application.

Further, for many years, we have supplied detailed online support of a public domain speech recognition system [2]. It has been our experience that more than 75% of our support requests involve situations in which the intermediate files required for training, such as a phonetic alignment, do not correctly correspond to the input audio data. Another common problem is that generalization of these context-independent phone transcriptions to context-dependent phone transcriptions often causes similar errors in the experimental setup.

Another significant complication in most speech recognition systems is silence modeling. Approaches that rely on a forced alignment stage during training also often have to identify and introduce silence into the transcriptions. Typically, this is done through the use of a silence phone. Several stages of training are often devoted to identification of silence and training of this silence model.

Hence, it has been our goal to simplify the training process without compromising performance. This is a non-trivial problem since it is well-known that bootstrapping procedures in which complexity is introduced incrementally have been very successful over the years in the development of speech recognition technology.

2. NETWORK TRAINING

A popular approach to speech recognition utilizes a hierarchical network of knowledge sources, as shown in Figure 1. The training paradigm employs maximum likelihood estimation (MLE) within the expectation maximization (EM) framework. The actual parameter estimation is implemented using a computationally efficient algorithm known as Baum-Welch reestimation (also referred to as the Forward-Backward algorithm). A detailed description of the training recipes used in our system and the Baum-Welch reestimation equations used in training can be found in [1,2].

The network training approach on the surface appears identical to the training paradigm used in a traditional HMM trainer. Baum-Welch reestimation is applied to the hierarchical network shown in Figure 1. In this paper, we show that training these networks directly can simplify the training process and provide comparable performance.

When compared to the standard left-to-right training approach used in a traditional system, Baum-Welch training can be viewed as providing the system a capability to make soft decisions. Probabilities about

1. This material is based upon work supported by the National Science Foundation under Grant No. IIS-0085940. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

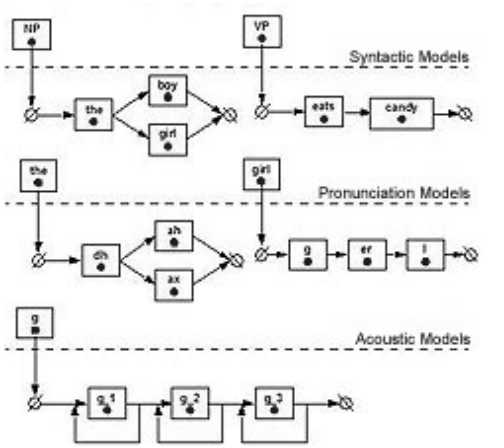


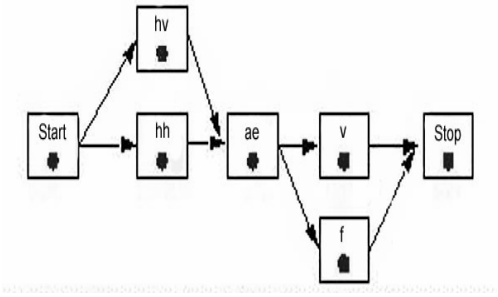
Figure 1: An example of a hierarchical system that contains embedded knowledge sources at each level.

pronunciations, and other alternate paths through the network, receive contributions from all data. It is well known that systems involving soft decisions [3] can provide better performance, though these systems may take longer to converge during training.

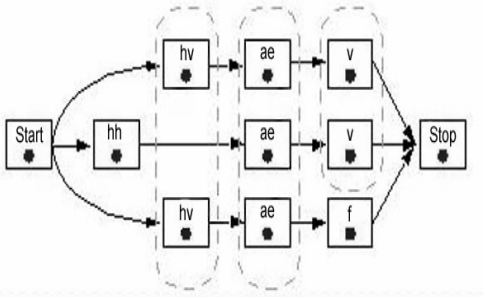
The network trainer directly estimates the parameters of multi-path graphs at all levels of the hierarchical network of knowledge sources. In Figure 2, we compare one level of the graph, a phone-level pronunciation network, to an equivalent expanded graph. Traditional HMM training recipes will often select one of the linear paths and update the corresponding models. The models corresponding to the alternate paths (and the associated arcs) will not be updated, unless a phone is common to both paths (as is the case in Figure 2).

The network trainer performs Baum-Welch reestimation across all networks simultaneously. It can be used to train probabilistic language models, pronunciation models, or acoustic models. Since the Baum-Welch algorithm is used at each level, this approach effectively makes soft decisions about symbol assignments. Such a feature is particularly useful when modeling pronunciations. It leads to better generalization during recognition, since unseen pronunciations can potentially occur in the network training paradigm.

This rather simple difference in the training paradigm has more profound implications for silence modeling. Silence modeling is one of the more crucial aspects of building a good speech recognition system. In traditional approaches, silence is often inserted into a word pronunciation as a phone. The lexicon will contain two entries for a word, one terminating in a short pause, and one terminating in a long silence (e.g., "have" would be represented as "hh ae v sp" and "hh ae v sil"). These silence models in turn contain a topology that allows



(a) A pronunciation network for the word "have".



(b) An expanded network.

Figure 2: In a traditional HMM training recipe, one of the three alternate pronunciations shown in (b) will be selected and trained. In the network trainer, the network in (a) is reestimated directly.

silence to be skipped. During the forced alignment stage of the training procedure, these silences must be identified explicitly in the transcriptions.

In the network training recipe, the forced-alignment stage is eliminated. Silence is simply treated as an alternate word in the pronunciation. The topology of the silence word allows for a long and a short path through the model, as shown in Figure 3. The multi-path silence word model removes the need for making a hard decision as to the duration of the silence after a word. During

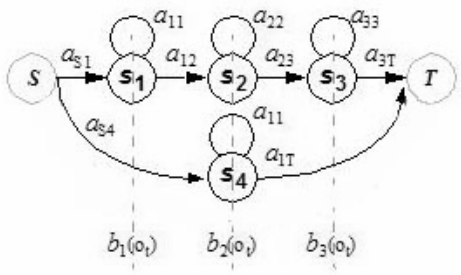


Figure 3: An example of a multi-path silence model topology used by the network trainer.

training, arbitrary amounts of silence are allowed as an optional word everywhere in the network: before and after an utterance, and between words. The trainer makes an optimal decision about the placement of silence by estimating probabilities the same way these probabilities are estimated for states in the acoustic models.

The combined impact of this flexibility is that training is driven entirely from the word-level transcription. No intermediate forced alignments are needed. A transcription is augmented with optional silence, and then a standard Baum-Welch training is performed across the entire hierarchy. In the next section, we will analyze the impact this has on overall performance.

3. EXPERIMENTS AND ANALYSIS

To prove our hypothesis, experiments were conducted on three corpora representing industry-standard tasks: TIDigits (TID) [4], OGI Alphadigits (AD) [5] and Resource Management (RM) [6]. All experiments were conducted with a context-independent (CI) speech recognition system since this is the stage most often impacted by the proposed changes. The baseline CI system was based on a context-dependent (CD) system that achieves near state of the art performance on the three tasks included in this study [7,8]. The CI system did not use any forms of adaptation or normalization in this study beyond that implied by a standard speaker independent system.

Introducing the full flexibility of the network trainer at the start of the training process can often backfire. One instance of this is when we allowed too much flexibility when training the silence model. For example, using an optional silence at the beginning and end of the transcription resulted in poor recognition performance on TIDigits. The word alignments in Figure 4 show the degree to which an underestimated silence can misalign the segment boundaries. The alignments compare the hypothesis for a fixed and an optional silence to the reference transcription. While the fixed silence hypothesis comes close to the reference, the optional silence hypothesis is misaligned.

We experimented with several ways to fix this including different training recipes. The source of the problem is that the silence model needs some amount of seeding before it can converge to the correct result. Letting both the silence and speech models learn simultaneously results in a suboptimal solution. Our best solution was to force silence at the beginning and ends of an utterance using the three-state long silence path through the silence model. This is something that is common in many training systems. We still allow silence to be optional between words. These two silence models share emission probabilities so the final parameter count is kept virtually the same. Using a fixed silence restricts the flexibility of the network trainer; however, this is a small price to pay compared to what we gain in terms of the overall flexibility of the system.

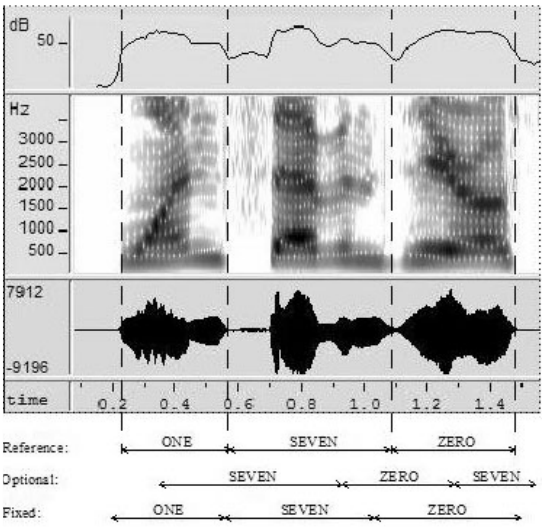


Figure 4: A comparison of word-alignment using fixed and optional silence models at transcription bounds.

A summary of the experimental results is given in Table 1. On the TIDigits and AlphaDigits tasks, performance was slightly better for network training. On the Resource Management task, performance was slightly worse. All differences are not statistically significant according to the NIST MAPSSWE test [9]. Further, we expect these differences would converge as subsequent stages of context-dependent training and Gaussian mixture splitting were performed.

The error modalities for the two systems were not significantly different. Introducing increased flexibility into the training process at the early stages of training can often backfire. Hence, we examined the convergence of the overall likelihood of the data given the models as a function of the number of training iterations. An example for Resource Management is shown in Figure 5. The convergence of the likelihood was similar for all three databases.

Database	Traditional HMM	Network Trainer
TID	7.7%	7.6%
AD	38.0%	35.3%
RM	25.7%	27.5%

Table 1: A summary of results on three popular databases that represent tasks ranging from digit recognition to medium-sized vocabulary command and control recognition.

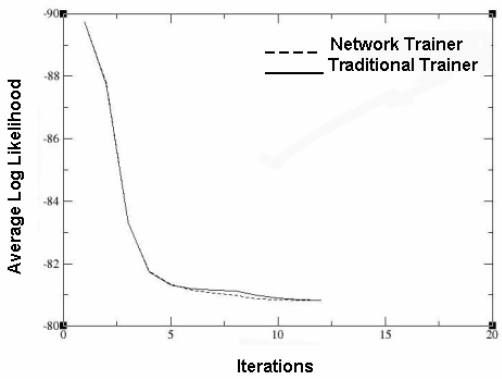


Figure 5: A comparison of the convergence in log likelihood between the two training paradigms for the Resource Management task. All three databases exhibited similar behavior.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we have explored the effectiveness of a network training approach that simplifies the training process. The basis of this approach is a hierarchical implementation of the Baum-Welch training algorithm. Network training allows any level of a hierarchical system to be trained using supervised learning. This approach was evaluated on three different databases: TIDigits, OGI Alphadigits and Resource Management. No significant change in word error rate was observed for a speaker independent speech recognition system that used single Gaussian mixture context-independent phone models.

Since the context-dependent stages of the training process are a direct extension of the context-independent stage, there are no significant changes to the training process once the context independent stages of training have been completed. However, a hierarchical lexical tree decoder is needed to decode the cross-word models. Such a decoder is currently under development for the system that incorporates the network training approach.

We are also independently pursuing the incorporation of discriminative training into our system, as well as new statistical models based on support vector machines and relevance vector machines. These approaches will make use of the network training paradigm. It will be interesting to determine the benefits of network training of multi-path acoustic models using discriminative training.

The features described in this paper are available in the latest releases of our public domain speech recognition system [10]. A more detailed description of this work is available at [11].

5. REFERENCES

1. X. Huang, A. Acero and H. Hon, *Spoken Language Processing,* Prentice Hall, Upper Saddle River, New Jersey, USA, 2001.

2. N. Deshmukh, A. Ganapathiraju, J. Hamaker and J. Picone, "An Efficient Public Domain LVCSR Decoder," *Proceedings of Speech Transcription Workshop*, Linthicum Heights, MD, USA, Sept. 1998.

3. Y. Hua and A. Waibel, "Flexible Parameter Tying for Conversational Speech Recognition," ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, Japan, April 2003.

4. R. Leonard, "A Database for Speaker-Independent Digit Recognition," *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 328-331, San Diego, CA, USA, March 1984.

5. R. Cole, *et al*, "Alphadigit v1.3," Center for Spoken Language Understanding, Oregon Graduate Institute, Oregon, USA, 1997 (available at http://www.cse.ogi.edu/cslu/corpora/alphadigit).

6. P. Price, W.M. Fisher, J. Bernstein and D.S. Pallett, "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 1, pp. 651-654, May 1998.

7. R. Sundaram, J. Hamaker, and J. Picone, "TWISTER: The ISIP 2001 Conversational Speech Evaluation System," *Proceedings of the Speech Transcription Workshop*, Linthicum Heights, MD, USA, May 2001.

8. T. Rotovnik, *et al*, "A Comparison of HTK, ISIP and JULIUS in Slovenian Large Vocabulary Continuous Speech Recognition," *Proc. of the Int. Conf. of Spoken Lang. Proc.*, pp. 681-684, Denver, CO, USA, Sept. 2002.

9. L. Gillick and S. J. Cox, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms," *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 532-535, Glasgow, Scotland, May 1989.

10. J. Picone, "Internet-Accessible Speech Recognition Technology," *http://www.isip.msstate.edu/projects/speech,* Institute for Signal and Information Processing, Mississippi State University, December 2002.

11. I. Alphonso, *Network Training for Continuous Speech Recognition*, M.S. Thesis, Department of Electrical and Computer Engineering, Mississippi State University, December 2003.