RETHINKING OF COMPUTATION FOR FUTURE-GENERATION, KNOWLEDGE-RICH SPEECH RECOGNITION AND UNDERSTANDING

Li Deng
Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
deng@microsoft.com

ABSTRACT

A new trend is emerging in the semiconductor industry that future computation speedups will likely come more from parallelism than from having faster individual computing elements. Most algorithm designers for the current, HMM-based speech recognition systems, which have the recognition performance significantly lower than that of human, have not embraced this trend. This is partly attributed to the state-of-the-art sequential algorithms that have involved extremely clever schemes to speed up single-processor performance developed and matured over many years. This invited presentation advances two arguments. First, much more powerful speech systems in the future generations will likely approach human performance with new architectures that integrate rich knowledge sources and overcome the reasonably well understood limitations of the current HMM-based systems. Second, the success of the above endeavor will require complete rethinking of computation issues, likely disposing of the traditional thinking of HMM-centric sequential processing and embracing parallel computing in the new architectures mimicking key aspects of the human speech processing system. Four case studies are provided in this paper extracted from some recent influential work that may shape the foundation of this potentially active research area.

Index Terms — computation, parallelism, speech recognition, speech understanding, decoding, knowledge integration.

1. INTRODUCTION

Moore's law has been a reliable predictor of the increased capability for computation and storage in computational systems for decades, creating enormous impact on the historical development of automatic speech recognition (ASR) and understanding systems. Larger and larger speech databases and recognition systems have been developed, along with more and more detailed models (such as complex hidden Markov model, or HMM) of speech and human language. Much of the planned future research in the ASR field has implicitly relied on a continued advance in computational capabilities (Baker et. al., 2009a, 2009b). However, the fundamentals of the progression based on Moore's law have recently changed. The power density on microprocessors has increased to the point that higher clock rates would begin to melt the silicon. Consequently, at this point industry development is now focused on implementing microprocessors on multiple cores. The new road maps for the semiconductor industry reflect this trend, and future speedups will come more from parallelism than from having faster individual computing elements. For the most part, algorithm designers for speech systems have ignored investigation of such parallelism, since the advance of scalar capabilities has been so reliable. Building future generation ASR systems, as discussed in (Baker et. al., 2009a, 2009b; Deng and Huang, 2004), will require significantly more computation, and consequently researchers concerned with implementation will need to consider parallelism explicitly in their designs.

The rest of the paper is organized as follows. In Section 2, we provide an overview of the current, HMM-based ASR system design, with a focus on their computation architecture. We will discuss in Section 3 some new trends in ASR research aimed to develop future-generation high-performance ASR systems going beyond HMMs as the architectural basis, and analyze in which areas of ASR research parallelism in computing may be needed. In Section 4, we use four case studies to illustrate some recent work that are representative of such trends.

2. OVERVIEW OF COMPUTATION ARCHITECURE IN CURRENT ASR SYSTEMS

Current ASR systems have been built invariably based on statistical principles, as pioneered by the work of (Baker, 1975; Jelinek, 1976). A source-channel mathematical model or a type of generative statistical model proposed therein is often used to formulate speech recognition problems. Briefly, the speaker's mind decides the source word sequence **W** that is delivered through his/her text generator. The source is passed through a noisy communication channel that consists of the speaker's vocal apparatus to produce the speech waveform and the speech signal processing component of the speech recognizer. Finally, the speech decoder aims to decode the acoustic signal **X** into a word sequence $\hat{\mathbf{W}}$, which is in ideal cases close to the original word sequence **W**.

The "fundamental equation" of speech recognition follows the following Bayes rule (Huang et. al., 2001; Deng and O'Shaughnessy, 2003):

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W} \mid \mathbf{X}) = \arg\max_{\mathbf{W}} \frac{P(\mathbf{W})P(\mathbf{X}|\mathbf{W})}{P(\mathbf{X})} = \arg\max_{\mathbf{W}} P(\mathbf{W})P(\mathbf{X} \mid \mathbf{W})$$

where $P(\mathbf{W})$ and $P(\mathbf{X} \mid \mathbf{W})$ constitute the probabilistic quantities computed by the language modeling and acoustic modeling components, respectively. The acoustic model is typically represented by an HMM, and language model by an N-gram model. As epitomized in the fundamental equation above, the decoding process in a speech recognizer's operation is to find a sequence of words whose corresponding acoustic and language models best match the input feature vector sequence. Therefore, the process of such a decoding process with trained acoustic and language models is often referred to as a *search* process. This process determines the computation architecture in an ASR system.

Speech recognition search is usually carried out with the Viterbi decoder, which is a sequential algorithm based on dynamic programming (DP). The reasons for choosing the Viterbi decoder involve arguments that point to speech as a left to right process and the efficiencies afforded by a time-synchronous process. One obvious (and brute-force) way is to search all possible word sequences and select the one with best posterior probability score. This, however, is not practically feasible.

The Viterbi search can be executed efficiently via the trellis framework. It is a time-synchronous search algorithm that completely processes time t before going on to time t+1. For time t, each state is updated by the best score (instead of the sum of all incoming paths) from all states in at time t-1. When one update occurs, it also records the backtracking pointer to remember the most probable incoming state. At the end of the search, the most probable state sequence can be recovered by tracing back these backtracking pointers. Viterbi algorithm provides an optimal solution for handling nonlinear time warping between HMMs and the acoustic observation, word boundary detection and word identification in continuous speech recognition. This unified Viterbi search algorithm serves as the fundamental technique for most search algorithms in use in continuous speech recognition.

Time-synchronous Viterbi search can be considered as a *breadth first search* with DP. Instead of performing a tree search algorithm, the DP principle helps create a search graph where multiple paths leading to the same search state are merged by keeping the best path. The Viterbi trellis is a representation of the search graph. Therefore, all efficient techniques for graph search algorithms can be applied to time-synchronous Viterbi search. Note that it is not necessary for the entire search space to be explored before the optimal path can be found. When the search space contains an enormous number of states in the HMM, it becomes impractical to pre-compile the composite HMM entirely and store it in the memory. It is preferable to dynamically build and allocate portions of the search space which is sufficient to search the promising paths. By using the graph search algorithm, only part of the entire Viterbi trellis is generated explicitly. By constructing the search space dynamically, the computation cost of the search is proportional only to the number of active hypotheses that is independent of the overall size of the potential search space. Therefore, dynamically generated trellis is a key to heuristic Viterbi search for efficient large-vocabulary continuous speech recognition.

Unfortunately, the core computation in current ASR systems, DP- based decoding algorithms, has been based on intrinsically sequential algorithms described above. These algorithms have exploited extremely clever schemes to speed up single-processor performance (e.g., Huang et. al. 2001; Jelinek, 1997). Recent work has appeared on using vector architectures with parallelism to implement the decoder, with the finding that the traditional HMM-based decoding algorithm is difficult to vectorize as the vocabulary grows (Janin, 2004). More recent work (Chong et. al., 2008) successfully parallelized the HMM-based decoding algorithm for a large vocabulary system (50,000 words and one million states). These works, however, are confined with the conventional HMM framework and are thus difficult to break the fundamental limitations imposed by the framework.

3. COMPUTATION ARCHITECTURE FOR FUTURE-GENERATION, KNOWLEDGE-RICH ASR

There have been strong consensus in the research community that the future-generation ASR will need to overcome fundamental limitations of the current HMM framework and that more powerful speech systems in the future will likely approach human performance with new architectures that integrate rich knowledge sources (e.g., Baker et. al., 2009a, 2009b; Lee, 2003). This provides opportunities to rethink many of the computation issues in ASR. HMM-bound sequential processing may not need to dominate the computation. The multi-stream analysis for ASR as discussed in (Baker et. al., 2009a, 2009b) may require heterogeneous parallelism in both the algorithms and the computational architecture.

While the incorporation of new types of multiple knowledge sources has been on the research agenda for decades for ASR, we are coming into a period where the resources are available to support this strategy in a much more significant way. For instance, it is now possible to incorporate both larger sound units than the typical phone or sub-phone elements even for large vocabulary recognition, while still preserving the advantage of the smaller units. Further, more fundamental units such as articulatory features can be considered (Sun and Deng, 2002). At the level of the signal processing "front end", we no longer need to settle on the single best representation, as multiple representations have been shown to be helpful (e.g., Morgan

et. al. 2005). Such multiple representations or streams of information can be even more heterogeneous, e.g., coming from different modalities such as bone-conducted vibration, cameras, or low-power radar. In all of these cases, new computation architectures are required that can aggregate all of the modules' responses, quite different from the current HMM systems which typically use singe-stream information. Various approaches for this have been tried for some time, but we are only now beginning to tackle the task of integrating so many different kinds of sources, due to the emerging availability of the kind of resources required to learn how to best do the integration.

4. CASE STUDIES

In this section, we further discuss some new trends in ASR research that are characterized by integrating rich, multi-stream knowledge sources and the associated computation issues involving parallel computing. In particular, we use four case studies to illustrate some recent work that are representative of such trends.

4.1. Parallel detectors and detection-based ASR

In the research direction proposed in (Lee, 2003) and reported in (Bromberg, et. al., 2007; Lee et. al., 2007), the authors argued that although we have learned a great deal about how to build practical HMM-based ASR systems for almost any spoken language without the need of a detailed understanding of the language, the existing technology is fragile in that careful designs have to be rigorously practiced to overcome technology deficiencies. Furthermore, the accuracy often declines dramatically in adverse conditions to an extent that the ASR system becomes unusable. When compared with human speech recognition, the state-of-the-art ASR systems usually give much larger error rates even for rather simple tasks. They note that human beings perform speech recognition by integrating multiple knowledge sources from bottom up and that a human determines the linguistic identity of a sound based on detected evidences that exist at various levels of the speech knowledge hierarchy. A human listener detects acoustic and auditory evidences, weigh them and combine them to form cognitive hypotheses, and then validate the hypotheses until consistent decisions are reached. The above human-based model of speech processing motivated the authors to develop a candidate framework for developing future-generation ASR technology that has the potential to go beyond the current limitations.

In this detection-based framework, rich knowledge sources are exploited, including a set of fundamental acoustic attributes of speech sounds and their linguistic interpretations, a speaker profile that encompasses gender, accent and other speaker characteristics, the speaking environment. Instead of the conventional top-down, HMM- and DP-based decoding for ASR which we reviewed in Section 2, they developed a bottom-up, event detection and evidence combination paradigm for future-generation ASR, where a large set of acoustic events are detected in parallel before being combined for detecting word sequences. The computation architecture associated with this paradigm can make much more effective use of parallelism than the conventional HMM-based sequential processing paradigm.

4.2. Integrating speech recognition and understanding

One recent trend in ASR research is to use its results as an intermediate step to achieve the ultimate goal of speech understanding. This provides the opportunity to redesign ASR systems so that it is not the word error rate but the speech understanding error rate which is subject to minimization. Traditional methods of speech understanding adopt two independently trained phases. First, an ASR module returns the most likely sentence for the observed acoustic signal. Second, a semantic classifier transforms the resulting sentence into the most likely semantic class. Since the two phases are isolated from each other, such traditional systems are suboptimal. In the work described in (Yaman et. al., 2008), a novel integrative and discriminative learning technique for a speech understanding system was developed to alleviate this problem, and thereby, reduces the semantic classification error rate. The new approach makes effective use of the N-best lists, which are processed in parallel, generated by the ASR module to reduce semantic classification errors.

4.3. Structured speech modeling and transformations

There has been a long tradition of research on overcoming one fundamental, incorrect assumption of the HMM and associated limitations --- conditional independence assumption for the input sequence (Ostendorf et. al., 1996, Deng et. al., 1993, 1994). Recent work in this direction embeds substantial knowledge related to speech articulation and constraints therein to explicitly provide the temporal correlation in the observed speech sequences (Deng et. al., 2006). The resulting structured speech model has a few more "hidden" layers than the HMM, and has greater computation cost while gaining higher recognition performance than the HMM system.

Like the HMM, the decoding algorithm developed for the structured speech model is also sequential in nature, making the model difficult to use in practice due to its huge computation cost. The future direction in this work is to flatten out the previous deep structure in the speech model into a large set of parallel transformations. The latter may then enable efficient parallel computing for model learning and decoding.

4.4. ASR system combination via ensemble learning

Another recent trend in ASR is to effectively combine the outputs from a set of different systems so as to achieve higher recognition performance than each of the individual

systems can achieve (Breslin and Gales, 2009). Currently, many large vocabulary ASR systems use a combination of multiple systems to obtain the final hypothesis. These complementary systems are typically found either in an ad-hoc manner, or by the use of ensemble learning methods.

Different from the commonly adopted approach of optimizing a single classifier, ensemble learning methods achieve pattern discrimination through synergistically combining many classifiers that are complementary in nature. In ASR applications, combining output word hypotheses from multiple speech recognition systems is being increasingly used for boosting the accuracy performance. The complexity of speech sound distributions also warrants the exploration of using ensemble methods to build robust and accurate acoustic models. For example, the component models of an ensemble can be combined in computing the acoustic scores during decoding search at the speech frame level. Recently, some innovative progresses have been made in this direction, producing promising results and revealing attractive properties of ensemble acoustic models (Chen and Zhao, 2009).

Using multiple systems and model components in either speech decoding or acoustic model construction provides a rich opportunity for parallel computing.

5. SUMMARY AND CONCLUSION

In this paper, I argue that the current ASR system design is limited by the use of the structure-poor and knowledge-ignorant HMMs in performance and by the sequential Viterbi-like decoding algorithms in computation. More powerful, future-generation ASR systems will need new paradigms that integrate rich knowledge sources. This will provide ample opportunities for new computation architectures, such as island-driven decoding vs. the current sequential decoding, where parallel processing will likely play an important role in implementing the new paradigms.

REFERENCES

[1] J. Baker, "Stochastic Modeling for Automatic Speech Recognition", in *Speech Recognition*, edited by D. R. Reddy, Academic Press, 1975.

[2] J. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy, "Updated MINDS Report on Speech Recognition and Understanding Part I", IEEE Signal Processing Magazine, Vol. 26, No. 3, May 2009a.

[3] J. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy, "Updated MINDS Report on Speech Recognition and Understanding Part II", IEEE Signal Processing Magazine, Vol. 26, No. 4, July 2009b.

[4] C. Breslin and M. Gales, "Directed decision trees for generating complementary systems," Speech Communication, Vol. 51, 2009, pp. 284-295.

[5] I. Bromberg et. al. "Detection-Based ASR in the Automatic Speech Attribute Transcription Project," Proc. Interspeech, 2007.

[6] X. Chen and Y. Zhao, "Data sampling based ensemble acoustic modeling," Proc. *ICASSP,* 2009, Taipei, Taiwan.

[7] J. Chong, Y. Yi, A. Faria, N. Satish, and K. Keutzer. "Data parallel large vocabulary continuous speech recognition on graphics processors," Technical Report No. UCB/EECS-2008-69, University of California at Berkeley, 2008.

[8] L. Deng and X. D. Huang, "Challenges in Adopting Speech Recognition," Communications of the ACM, vol. 47, no. 1, pp. 11-13, Jan. 2004

[9] L. Deng and D. O'Shaughnessy, *SPEECH PROCESSING --- A Dynamic and Optimization-Oriented Approach*, Marcel Dekker Inc., New York, NY, 2003.

[10] L. Deng, A stochastic model of speech incorporating hierarchical nonstationarity, IEEE Transactions on Speech and Audio Processing, vol. 1, no. 4, pp. 471-475, 1993.

[11] L. Deng, M. Aksmanovic, D. Sun, and J. Wu. "Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states," IEEE Trans. Speech & Audio Proc., Vol. 2, 1994, pp. 507-520.

[12] L. Deng, D. Yu, and A. Acero. "Structured speech modeling," IEEE Transactions on Audio, Speech and Language Processing, Vol. 14, 2006, pp. 1492-1504.

[13] X.D. Huang, A. Acero, and H. Hon. *Spoken Language Processing --- A Guide to Theory, Algorithms, and System Development*, Prentice Hall, 2001.

[14] A. Janin. Speech Recognition on Vector Architectures, Ph.D. thesis, University of California at Berkeley, 2004.

[15] F. Jelinek. Statistical Methods for Speech Recognition, MIT Press, Cambridge, MA, 1997.

[16] F. Jelinek, "Continuous Speech Recognition by Statistical Methods", Proc. of the IEEE, 64(4), 1976, pp. 532-557.

[17] C.-H. Lee, "On Automatic Speech Recognition at the Dawn of the 21th Century," *IEICE Trans. on Information and Systems, Special Issue on Speech Information Processing,* Vol.E86-D, No. 3, pp. 377-396, March 2003.

[18] C.-H. Lee, et. al. "An overview on automatic speech attribute transcription," Proc. Interspeech, 2007, pp. 1825-1828.

[19] N. Morgan, Q. Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, O. Cetin, H. Bourlard, and M. Athineos, "Pushing the Envelope – Aside", IEEE Signal Processing Magazine, Sept 2005, pp. 81-88.

[20] M. Ostendorf, V. Digalakis, and J. Rohlicek. "From HMMs to segment models: A unified view of stochastic modeling for speech recognition", IEEE Trans. Speech Audio Proc., Vol. 4, 1996, pp. 360-378.

[21] J. Sun and L. Deng. "An overlapping-feature based phonological model incorporating linguistic constraints: Applications to speech recognition," Journal of the Acoustical Society of America, Vol. 111, No. 2, February 2002, pp.1086-1101.

[22] S. Yaman, L. Deng, D. Yu, Y. Wang, and A. Acero, "An integrative and discriminative technique for spoken utterance classification," IEEE Trans. Audio, Speech, and Language Processing, vol. 16, pp. 1207-1214, 2008.