Plenary Session 2

09:00 - 10:00, Wednesday, October 6, 2004

From Decoding-Driven to Detection-Based Paradigms for Automatic Speech Recognition

Chin-Hui Lee, Georgia Institute of Technology

Abstracts

We present a detection-based automatic speech recognition (ASR) paradigm that is capable of integrating both the knowledge sources accumulated in the speech science community and the modeling techniques established in the speech processing community. By exploring this new framework, we expect that researchers in the Interspeech community can collaboratively contribute to developing next generation algorithms that have the potential to surpass current capabilities, and go beyond the limitations of the state-of-the-art ASR technologies.

Biographical Sketch

Chin-Hui Lee is a professor at School of Electrical and Computer Engineering at Georgia Institute of Technology. Dr. Lee received the B.S. degree in Electrical Engineering from National Taiwan University, Taipei, in 1973, the M.S. degree in Engineering and Applied Science from Yale University, New Haven, in 1977, and the Ph.D. degree in Electrical Engineering with a minor in Statistics from University of Washington, Seattle, in 1981.

After graduation, Dr. Lee joined Verbex Corporation, Bedford, MA, and was involved in research on connected word recognition. In 1984, he became affiliated with Digital Sound Corporation, Santa Barbara, where he engaged in research and product development in speech coding, speech synthesis, speech recognition and signal processing for the development of the DSC-2000 Voice Server. Between 1986 and 2001, he was with Bell Laboratories, Murray Hill, New Jersey, where he became a Distinguished Member of Technical Staff and Director of the Dialogue Systems Research Department. His research interests include multimedia communication, multimedia signal and information processing, speech and speaker recognition, speech and language modeling, spoken dialogue processing, adaptive and discriminative learning, biometric authentication, information retrieval, and bioinformatics. His research scope is reflected in "Automatic Speech and Speaker Recognition: Advanced Topics", published by the Kluwer Academic Publishers in 1996. From August 2001 to August 2002 he was a visiting professor at School of Computing, The National University of Singapore. In September 2002, he joined the Faculty of School of Electrical and Computer Engineering at Georgia Institute of Technology.

Prof. Lee has participated actively in professional societies. He is a member of the IEEE Signal Processing Society, Communication Society, and the European Speech Communication Association. He is also a lifetime member of the Computational Linguistics Society in Taiwan.

In 1991-1995, he was an associate editor for the IEEE Transactions on Signal Processing and Transactions on Speech and Audio Processing. During the same period, he served as a member of the ARPA Spoken Language Coordination Committee. In 1995-1998 he was a member of the Speech Processing Technical Committee of the IEEE Signal Processing Society (SPS), and later became the chairman of the Speech TC from 1997 to 1998. In 1996, he helped promote the SPS Multimedia Signal Processing (MMSP) Technical Committee in which he is a founding member.

Dr. Lee is a Fellow of the IEEE, and has published more than 250 papers and 25 patents on the subject of automatic speech and speaker recognition. He received the SPS Senior Award in 1994 and the SPS Best Paper Award in 1997 and 1999, respectively. In 1997, he was also awarded the prestigious Bell Labs President's Gold Award for his contributions to the Lucent Speech Processing Solutions product. In 2000, he was named one of the six Distinguished Lecturers by the IEEE Signal Processing Society.

From Knowledge-Ignorant to Knowledge-Rich Modeling: A New Speech Research Paradigm for Next Generation Automatic Speech Recognition

Chin-Hui Lee
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
chl@ece.gatech.edu

Abstract

The field of automatic speech recognition (ASR) has enjoyed a fast technology progress in the last three decades, due to the extensive use of statistical learning algorithms, the availability of a number of large collections of speech and text examples, and fast computing machines. However ASR advances have slowed down quite a bit in recent years. It seems the success of the above *knowledge-ignorant* modeling approach can be further extended if knowledge sources available in the large body of speech science literature can be properly integrated into the statistical modeling paradigm and objectively evaluated. In this paper we explore a *knowledge-rich*, data-driven approach to ASR that serves as a candidate paradigm for developing next generation ASR techniques and systems. A few knowledge-supplementary examples will first be illustrated. Some potential collaborative research scenarios will also be discussed.

1. Introduction

Speech is considered as the most natural means of communication among human beings. There is also a rich set of human information embedded in speech beyond just word transcription of the spoken utterances. Mining of speech information is therefore of great importance both in theory and in practice. One way to extract such information from spoken languages by machine is to convert speech to text through *automatic speech recognition* (ASR). With the increasing research and application interests, the field of ASR has enjoyed a fast progress in the last three decades, due to the extensive use of statistical learning techniques, the availability of large collections of speech and text examples, and fast computing machines. However ASR advances have slowed done quite a bit in recent years. It seems the past successes of the prevailing *knowledge-ignorant* modeling approach can still be further extended if knowledge sources available in the large body of literature in speech and language sciences can be objectively evaluated and properly integrated into statistical modeling. In this paper we explore a *knowledge-rich*, data-driven modeling paradigm that is capable of going beyond the current limitations of the state-of-the-art ASR technology, and gradually bridging the performance gap between ASR and HSR (*human speech recognition*). The implied *detection-based approach to ASR*, taking advantage of both knowledge-based and data-driven modeling paradigms, serves as a candidate paradigm for developing next generation ASR algorithms and systems.

2. Knowledge-Ignorant Modeling

The statistical modeling approach to ASR is motivated by expressing spoken utterances as stochastic *patterns*. ASR is then accomplished by finding the sequence of words that maximizes the joint probability, $P(S, W)$, of a given spoken utterance, S, and the corresponding word sequence, W, assuming that $P(S, W)$ is known. Based on statistical decision theory, it can be shown that the above ASR solution agrees with the *optimal Bayes rule* that minimizes the total risk of the expected sentence, word, or phone error rates, depending on the underlying problems of interest [1].

Next we consider the *Shannon's channel modeling paradigm* that a given sequence of input symbols, I, is passed through a noisy channel, and converted into an output signal, O. We are interested in recovering I from O, by designing an optimal channel decoder that maximizes the joint probability of I and O, $P(I, O)=P(O|I)P(I)$, where $P(O|I)$ is the conditional probability of O given I, and $P(I)$ is the prior probability of I. Therefore the two ASR perspectives, one based on designing the optimal Bayes decision rule, and the other on optimal channel decoding [2], give the same solution in implementing the *maximum a posteriori* (MAP) decoding policy that maximizes $P(W|S)$, or similarly $P(I|O)$. This powerful tool has been applied to many speech and language processing applications. Since we don't have an exact knowledge of the joint probability distributions for most practical problems, the forms of the distribution functions of $P(O|I)$ and $P(I)$ are often assumed, and their corresponding distribution parameters are then estimated from a large collection of application specific training examples. Therefore, the recognized input sequence is solved by implementing a *plug-in MAP decoding policy* [1] that plugs the estimated parameters into the assumed distributions in order to evaluate the required probabilities, or likelihood, in MAP decoding. Clearly, the optimal policy in decision and channel decoding no longer holds.

In most speech and language processing problems, the input symbol sequences, such as words, concepts, or part-of-speech tags, can often be approximated by Markov chains, while the output observations are either continuous signals, like speech, or discrete signals, like word sequences. Due to the fact that the input symbols are hidden and not directly observed Markov sources, the output observations can now be considered as discrete or continuous density *hidden Markov models* (HMMs) [3]. This partly explains the success of using HMMs in many recently reformulated pattern classification applications, even if the observed signals are not necessarily generated by hidden Markov sources. Designing the channel decoder, or computing the acoustic model, $P(S|W)$ and language model $P(W)$, demands modeling of $P(O|I)$ and $P(I)$, which is often accomplished by collecting a large training set of input-output pairs, and applying statistical learning techniques to estimate all the distribution parameters required to evaluate the two probabilities, $P(O|I)$ and $P(I)$. It is clear that the abovementioned approach does not require any detailed specifications of the input symbols being decoded. With such a knowledge-ignorant modeling approach, it is now quite straightforward to demonstrate ASR capabilities of new tasks for almost any spoken language, without using detailed descriptions about the language. There are now available a vast collection of speech and language corpora, sponsored by many business and government-funded projects in many countries. Advances in hardware, algorithms and data structures have also made implementation of large vocabulary, continuous speech recognition (LVCSR) systems affordable.

3. Advances in Hidden Markov Modeling

Knowledge-ignorant modeling is considered as one of the most fruitful areas in characterizing speech and language in recently years. Key advances [1, 4] can be summarized in three broad topics, namely: (1) *Detailed Modeling* - Software packages [5] are available now in public domains to establish acoustic models with hundred of thousand *Gaussian mixture* components, and language models with hundred of million of *n-gram* probabilities. The previous limitation imposed by the *curse of dimension,* widely known in the pattern recognition community, was alleviated with many advanced modeling techniques that take parameter sharing into account, such as the commonly-used tied-state tree learning strategy in phone modeling; (2) *Adaptive Modeling* - Adaptive learning of HMM parameters, such as MAP adaptation for parameters of phones and their corresponding structures, is now a standard practice in many systems. Online adaptation of HMM parameters has also been developed to improve learning efficiency and effectiveness with little data; and (3) *Discriminative Modeling* – Using learning criterion that is consistent with speech recognition and verification objectives, *minimum classification error* (MCE) and *minimum verification error* (MVE) learning algorithms for HMM parameters have been shown quite effective in improving model separation, system accuracy, and performance robustness.

4. Limitations with Knowledge-Ignorant Models

Based on the above statistical modeling paradigm we have learned a great deal about how to build practical ASR systems for almost any spoken language without the need of a detailed understanding of the language. However these existing systems are often overly restrictive, requiring that their users have to follow a very strict set of protocols to effectively utilize spoken language applications. Furthermore, the ASR system accuracy often declines dramatically in adverse conditions to an extent that an ASR system becomes unusable, even for cooperative users. When compared with HSR [6], the state-of-the-art ASR systems usually give much larger error rates even for rather simple tasks operating in clean environments. In highly noisy conditions, such as those in moving vehicles, ASR often gives an error rate more than one-two orders of magnitude higher than HSR [7]. Such a performance gap is unacceptable to users and makes the work of application designers extremely difficult.

In addition to the robustness problem with adverse *acoustic mismatches*, in most real-world environments spoken messages are usually conveyed with spontaneous speech which is often "ill-formed", with many utterances containing *out-of-task, out-of-grammar, and out-of-vocabulary* speech segments. Since we don't have a complete specification to characterize all possible ways of such spoken expressions using finite state network representations of all needed knowledge sources, the commonly-adopted MAP decoding policy for ASR is no longer applicable to this so called "*open-set*" robustness problem with adverse *linguistic mismatches*. Therefore the conventional ASR notion, of a complete word transcription of any spoken sentence by any person speaking in any acoustic condition with any language, cannot really be achieved using the current state-of-the-art ASR paradigm. Instead, we need to explore new formulations that can facilitate *partial understanding* of spoken languages, just like in the case of human speech understanding without the need of recognizing every single sound in a spoken utterance.

However efforts in integrating detailed knowledge, from acoustics, speech, language and their interactions, are hampered by the current ASR formulation as a "blackbox" of models trained to "remember" the training data, because it is not straightforward to integrate all available knowledge sources into the current top-down, knowledge-ignorant modeling framework. This makes it difficult for the ASR community to take advantage of the vast body of literature developed in the speech and language science communities to improve the performance and robustness of ASR systems. Thus, a collaborative paradigm is needed to facilitate innovations and lower entry barriers to ASR research for every single individual or group interested in contributing to ASR advances.

5. Knowledge-Supplemental Modeling

In the last few years, we have witnessed an increasing awareness of attempting to integrate limited knowledge sources into state-of-the-art HMM-based ASR systems to improve recognition accuracy. Three such examples are illustrated as follows, namely: (1) *"Sound-Specific" Features* [8] **-** A single voice onset time measurement, or VOT, was shown to be more powerful than 39 spectral features for discriminating a stop pair, such as /d/ vs. /t/. By re-ordering recognition candidates according to VOT, a two-stage alphabet recognizer gave a 50% error rate reduction over state-of-the-art ASR results; (2) *Key-Phrase Spotting Mimicking "Foreign Ears"* [9] - Human listeners are very good in detecting relevant keywords, buried in utterances of a foreign language. This suggests using a detection approach to mimic keyword spotting by "foreign ears", or poor ASR systems. For a spontaneous speech application, such a combined *keyword detection* and *utterance verification* [10] strategy maintained good accuracy for in-grammar utterances and greatly reduced errors for ill-formed sentences; and (3) *"Knowledge-Based" Front-End* [11] **-** An LVCSR system was built based on speech attributes produced by artificial neural network detectors. These "knowledge-based" features were then used to train a set of HMMs. By merging the MFCC baseline system with systems built with 60 attributes and 44 phone features using a ROVER combination [12], we obtained a word error rate of 3.7% for the WSJ 5K test used in Nov92 evaluation, about 20% relative error reduction over the best baseline system.

Other notable efforts include two 2003 symposiums and a recent study at the JHU Summer Workshop: (1) *Perspectives on Speech Separation*, http://www.ebire.org/speechseparation/; (2) *Symposium on Next Generation Automatic Speech Recognition*, http://www.ece.gatech.edu/~chl/ngasr03; and (3) *Landmarks-Based ASR*, http://www.clsp.jhu.edu/ws2004/groups/ws04ldmk/.

6. Human-Based Models for Speech Processing

It is interesting to note that human beings perform HSR by integrating multiple knowledge sources from bottom up. It has long been postulated that a human determines the linguistic identity of a sound based on detected evidences that exist at various levels of the speech knowledge hierarchy, from acoustics to pragmatics. For example, Klatt [13] studied the so-called *acoustic landmarks* that are assumed invariant to changes in speakers and speaking environments. Stevens [14] and Fant [15] have consistently advocated the approach of detecting and recognizing distinctive features in speech sound from an acoustic-phonetic framework. Indeed, people do not continuously convert a speech signal into words as an ASR system attempts to do. Instead, they detect *acoustic* and *auditory* evidences, weigh them and combine them to form *cognitive* hypotheses, and then *validate* the hypotheses until consistent decisions are reached. This process has been successfully demonstrated in *spectrogram reading* by trained experts based on knowledge in acoustic-phonetics [16]. Furthermore, a *phonological parsing* paradigm [17] for ASR has been proposed by assuming all the distinctive features can be exactly detected. However these features are not widely used in speech recognition due to the fact that they cannot be reliably detected in continuous speech, especially in adverse acoustic conditions.

In order to bridge the performance gap between ASR and HSR systems, it seems clear that the narrow notion of speech-to-text in ASR has to be expanded to incorporate all related human information "hidden" in speech utterances. This collection of information includes a set of fundamental speech sounds and their linguistic interpretations, the speaking environment that describes the interaction between speech and acoustics, a speaker profile that encompasses gender, accent and other speaker characteristics, such as the emotional state, etc. Collectively, we call this set of speech information, *speech attribute*s. They are not only critical for ASR but also useful for many other applications, including speech coding, speaker recognition, language identification, speech perception, and speech synthesis. Based on this set of speech attributes, ASR can be extended to *Automatic Speech Attribute Transcription*, or ASAT, a process that goes beyond the current notion of just word transcription.

7. A Detection-Based ASR Paradigm

The above human-based model of speech processing suggests a candidate framework for developing next generation speech technologies that have the potential to go beyond the current limitations. The missing link in utilizing acoustic and linguistic knowledge sources to recognize speech lies in designing a *bank of feature detectors* that is mathematically rigorous and capable of producing consistent detection results, even in adverse conditions. These robust event detectors are designed using acoustic-phonetic knowledge but are stochastic in nature so that the principles of statistical hypothesis testing and data-driven modeling techniques (as successfully adopted in the *top-down*, knowledge-ignorant ASR systems) can be extended to such a *bottom-up*, *knowledge-rich* modeling approach. These detected "events" can then be combined into higher level knowledge and evidences for phone and word recognition in a probabilistic manner, using hypothesis testing theories and computationally efficient algorithms. Furthermore, there is no need to restrict feature extraction at a fixed frame rate. Analog detectors and other related biologically-motivated processors [18] can also be incorporated into this flexible framework. The methodology distinguished itself from prior acoustic-phonetic approaches [13] practiced in the 1970's in the consistent use of data-driven designs for *speech event detection* and in the way the detected cues are fused into higher level evidence in *linguistic knowledge integration* with hypothesis testing and *pattern verification* [19].

A block diagram of the proposed ASAT paradigm for collaborative speech research is shown in Figure 1. It outlines a four-fold proposition that: (1) We can build a bottom-up speech recognition system that combines information from articulatory and acoustic-phonetic features to form phones, words, and word sequences; (2) We can do it in a modular fashion that can facilitate plug-'n'-play interoperability, allowing for close collaboration across a wide range of groups; (3) We can encourage new speech applications beyond the traditional notion of word transcription, such that researchers in both speech science and speech processing communities can contribute to technology advances; and (4) We can continue to practice the objective evaluation methodology [20] commonly adopted in the ASR community to track technology progress by developing similar evaluation strategies for individual modules and overall system to monitor detection performance, evidence combination effectiveness, as well as feature, phone and word accuracies.

Furthermore, the proposed bottom-up detection paradigm implies a new approach to solving the robustness problems by using a "divide and conquer" strategy. It also enables us to take advantage of the vast body of literature developed outside the ASR community. Knowledge in speech production and auditory processing and perception can also be applied to detection-based ASAT and ASR. By providing a plug-'n'-play platform, we hope to encourage the many researchers that have worked on the rule-based induction, acoustic-phonetic feature detection, and machine learning techniques in ASR and other speech areas to combine their efforts into such a single system. The proposed approach, when applied to auditory processing, attempts to simulate the human auditory process by assuming that speech is first converted to a collection of auditory response patterns (feature detection), each modeling the probabilistic activity level of a particular acoustic-phonetic event of interest. Knowledge sources and computational models in neuroscience [21] can also be extensively utilized. Detection of the next level of events or evidence, such as phones and words, is accomplished by combining relevant features. Each activity function can be modeled by a corresponding neural system. Both the activation levels and firing rates have been used in neural encoding. Artificial neural networks [22] provide a convenient tool to model neuron combinations. *Feedforward neural networks* have been used to encode and decode temporal information. *Recurrent neural networks* have also been used to provide feedback loops to simulate neural processing. Simulating perception of temporal speech events is of particular interest.

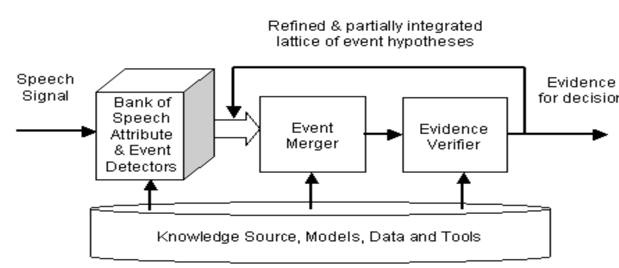

Figure 1 Bottom-up ASAT based on speech attribute detection, event merging and evidence verification

8. Summary

In summary, the ASR community is now at a crossroad searching for new directions. We are exploring a new knowledge-rich, data-driven modeling approach to next generation automatic speech recognition under a recently awarded NSF ITR grant: *Automatic Speech Attribute Transcription (ASAT): A Collaborative Speech Research Paradigm and Cyberinfrastructure with Applications to Automatic Speech Recognition*. It is clear that we have a long way to go before we can develop a complete ASR system that is competitive with the state-of-the-art performances. However, we do believe such a detection-based paradigm is flexible and rich in features, and provides an excellent vehicle for collaborative speech research. To facilitate such a community effort we intend to develop an open and sharable platform, and make all the system modules and tools available to the broader speech research community.

It is believed that by incorporating knowledge sources into speech modeling and processing, the set of recognized attribute sequences, event lattices, and evidences for decision in Figure 1 provides an instructive collection of diagnostic information, potentially beneficial for improving our understanding of speech, as well as enhancing speech recognition accuracy. From some of our preliminary results, we found that the recognized errors produced in knowledge-based, data-driven modeling systems often corresponded to more meaningful confusion of sounds in the same broad phonetic class than those errors obtained with knowledge-ignorant modeling systems, although the features are not necessarily more discriminative in classifying these speech sounds. Based on this set of information, we believe a better set of attribute detectors can be designed and they will contribute to improving both modular and overall system performances.

It is also noted that the performance in the proposed system is "additive". For example a better module for a feature will not produce poorer performance for the individual module and other modules related to this attribute, including the overall system. To facilitate a community effort to monitor research progress we will design a collection of evaluation sets for each speech attribute. Corresponding performance history will also be documented. Everyone is welcome to participate in this effort. We hope to eventually obtain a collection of "best" modules collectively provided by the community for all the needed features, so that they can be collaboratively incorporated into the "best" overall ASR system of the next generation.

Acknowledgement

The ASAT team includes M. Clements, K. Johnson, E. Fosler-Lussier, L. Rabiner, and C.-H. Lee. The author benefits greatly from many interactions with B.-H. Juang, and appreciates his continuing collaboration. The author also owes his sincere gratitude to his past colleagues at Bell Labs, Murray Hill, for endless and stimulating discussions. Such precious opportunities will be forever missed. Finally, the author thanks his Georgia Tech students, Yu Tsao and Jinyu Li, for generating some of the figures and results used in his oral presentation. Part of their efforts was supported under the NSF SGER grant, IIS-03-96848.

Reference

[1] C.-H. Lee and Q. Huo, "On Adaptive Decision Rules and Decision Parameter Adaptation for Automatic Speech Recognition," *Proc. IEEE*, Vol. 88, No. 8, pp. 1241-1269, 2000.

[2] L. R. Bahl, F. Jelinek and R. L. Mercer, ``A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. Pattern Analysis, Machine Intelligence*, Vol. 5, No. 2, pp. 179-190, 1983.

[3] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[4] B.-H. Juang and S. Furui, ``Automatic Speech Recognition and Understanding: A First Step Toward Natural Human-Machine Communication," *Proc. IEEE*, Vol. 88, No. 8, pp. 1142-1165, August 2000.

[5] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book* (for HTK Version 2.1), Cambridge University, 1997.

[6] J. Allen, "How Do Humans Process and Recognize Speech," *IEEE Trans. Speech and Audio Proc.*, Vol. 2, No. 4, pp. 567—577, 1994.

[7] R. Lippmann, ``Speech Recognition by Human and Machines," *Speech Communication*, *22*, pp. 1-14, 1997.

[8] P. Niyogi and P. Ramesh, ``A Detection Framework for Locating Phonetic Events," *Proc. ICSLP-98*, Sydney, 1998.

[9] T. Kawahara, C.-H. Lee and B.-H. Juang, "Key-Phrase Detection and Verification for Flexible Speech Understanding," *IEEE Trans. on Speech and Audio Proc.*, Vol.6, No. 6, pp.558-568, 1998.

[10] R. A. Sukkar and C.-H. Lee, "Vocabulary Independent Discriminative, "Utterance Verification for Non-Keyword Rejection in Subword Based Speech Recognition," *IEEE Trans. on Audio and Speech Proc*, Vol. 4, No. 6, pp. 420-429, 1996.

[11] B. Launay, O. Siohan, A. C. Surendran and C.-H. Lee, ``Towards Knowledge-based Features for HMM Based Large Vocabulary Automatic Speech Recognition", *Proc. ICASSP-2002*, Orlando, 2002.

[12] J. G. Fiscus, ``A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," *Proc. 1997 ASRU Workshop,* pp. 347-352, Santa Barbara, 1997.

[13] D. Klatt, ``Review of the ARPA Speech Understanding Project," J. Acous. Am., Vol. 62, No. 6, 1977.

[14] K. Stevens, *Acoustic Phonetics*, MIT Press, 1998.

[15] G. Fant, *Speech Sounds and Features*, MIT Press, 1973.

[16] V.W. Zue, ``Acoustic-Phonetic Knowledge Representation: Implications from Spectrograms Reading Experiments," *NATO ASI on Speech Recognition*, Bonas, France, 1981.

[17] K. W. Church, *Phonological Parsing in Speech Recognition*, Kluwer Academic Publishers, 1987.

[18] S. Shamma, "On the role of space and time in auditory processing", *Trends in Cognitive Sciences*, Vol. 5, No. 8, pp. 340-348, 2001.

[19] E. L. Lehmann, *Testing Statistical Hypotheses*, John Wiley & Sons, 1959.

[20] D. S. Pallett, ``A Look at NIST's Benchmark ASR Test: Past, Present, and Future," *Proc. IEEE ASRU Workshop 2003*, pp. 483-488, Virgin Islands, Dec. 2003.

[21] P. Dayan and L. F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*, MIT Press, 2001.

[22] S. Haykin, *Neural Networks: A Comprehensive Foundation*, McMillan, 1994.