SPEECH RECOGNITION: CURRENT STATUS AND PROSPECTS

Li Deng

Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

INTRODUCTION

Speech recognition technology has been significantly advanced over the past two decades. The advances can be attributed to the breakthrough use of a consistent statistical paradigm empowered by increasing quantities of speech data corpus, as well as by powerful algorithms developed for model learning from the data. Up until now, the technology has been primarily founded on the principle of statistical "ignorance" modeling, where generally unstructured speech models (mainly the hidden Markov model or HMM) learn their gigantic number of parameters from massive amounts of directly observable speech data.

In this plenary talk, I will first provide a critical review of the state-of-the-art automatic speech recognition technology. This review will include: 1) Analyzing the "Fundamental Equation" of statistical speech recognition; 2) Fundamental architecture of the modern speech recognition systems dissected into their common, basic phonetic and phonological components; and 3) Critical review of the backbone of modern speech recognition technology — HMMs. In this review, I will try to analyze why the current speech recognition technology is successful in certain areas of applications and not successful in other areas. Following such a review, I will address a number of key research issues which are aiming to overcome several fundamental deficiencies in the current speech recognition technology. A potentially fruitful approach will be outlined. This approach replaces the "bead-on-the-string" notion in the linear phonological model (uniformly used in the current speech recognition technology) by a version of nonlinear phonology in which the atomic speech units are constructed from multi-dimensional features overlapping in time. This approach further interfaces such units to phonetic models of speech dynamics, which has a statistical structure generalizing from the conventional HMM.

SPEECH RECOGNITION: FUNDAMENTAL EQN.

First, Let me give a brief description of the statistical framework that underlies much of modern speech recognition research and system development. Let $\mathrm{O} = O_1, O_2, ..., O_T$ be a sequence of observable acoustic data of speech, which can either be speech waveforms, or continuous-valued acoustic vectors (or any other type of general acoustic measurements), and let $W = w_1, w_2, ..., w_n$ be the sequence of words intended by the speaker who produces the acoustic record O above. The goal of a speech recognizer is to "guess" the most likely word sequence $\hat{W}$ given the acoustic data O. Bayesian decision theory provides a minimum Bayes-risk solution to the above "guessing game", and the minimum Bayes risk can be made equivalent to minimum probability of error if the risk is assigned values of one or zero for incorrect and correct guesses, respectively. According to Bayesian decision theory, speech recognition is formulated as a top-down search problem over the allowable word sequences W based on the posterior probability $P(W|\mathrm{O})$:

$$\hat{W} = argmax_W P(W|\mathrm{O}) = argmax_W P(\mathrm{O}|W)P(W), \tag{1}$$

where $P(W)$ is the prior probability that the speaker utters W, which is independent of the acoustic data and is determined by the language model, and $P(\mathrm{O}|W)$ is the probability that the speaker produces (or the microphone of the speech recognizer receives) the acoustic data O if W is the intended word sequence by the speaker. Disregarding the issue of language modeling, the above formulation, or fundamental eqn. (1), of the speech recognition problem can be reduced to two issues: 1) speech generation or production from word sequence to acoustic streams — how to accurately compute the probability $P(\mathrm{O}|W)$? and 2) a search for the word sequence W (the operation $\arg\max_W$ in Eqn. 1) that provides the optimal value of the posterior probability.

CRITICAL REVIEW OF HMMs

There is no doubt that HMMs are currently the most successful technology in many (heavily) constrained speech recognition applications. This success is not so much due to the mathematical formulation of the HMM itself as due to its conformity to the probabilistic analysis-by-synthesis formulation epitomized in Eqn.(1). Implicit in Eqn.(1) are the need to efficiently compute a production probability $P(\mathrm{O}|W)$ and the need to learn "production model" parameters so as to achieve high accuracy in evaluating $P(\mathrm{O}|W)$. HMMs are amenable to efficient computation and parameter learning thanks to Baum's work, and thus would fit naturally into the probabilistic analysis-by-synthesis framework of Eqn.(1). This is entirely consistent with the qualification of an HMM as a speech generator or production model, because embedded in the HMM there is a mechanism for converting a word sequence W directly into acoustic data O.

The theoretical treatment of the HMM as a production model is one thing; how reasonably and effectively it behaves as a production model is another thing. To examine this latter issue, let us first examine the production probability $P(\mathrm{O}|W)$ which appeared in Eqn.(1) into two factors:

$$P(\mathrm{O}|W) = \sum_{\mathcal{P}} P(\mathrm{O}|\mathcal{P})P(\mathcal{P}|W) \approx max_{\mathcal{P}} P(\mathrm{O}|\mathcal{P})P(\mathcal{P}|W), \tag{2}$$

where $\mathcal{P}$ is a discrete-valued *phonological model* and specifies, according to probability $P(\mathcal{P}|W)$, how words and word sequences W can be expressed in terms of a particular organization of a small set of "atomic" phonological units; $P(\mathrm{O}|\mathcal{P})$ is the probability that a particular organization $\mathcal{P}$ of phonological units produces the acoustic data for the given word sequence W. We shall call this latter mapping device from phonological organization to speech acoustics *phonetic model.*

In view of the factorization in Eqn.(2), state-of-the-art speech recognizers using phone-based HMMs can be analyzed as follows. The phonological model $\mathcal{P}$ is essentially a linearly-organized multiple-state phone sequence governed by a left-to-right Markov chain, and the phonetic model is simply a temporally independent random sampling from a set of (trainable) acoustic distributions associated with the states in the Markov chain. Both of these model components are highly simplistic descriptions of the true speech process, and such simplicity limits the success of the current technology in free-constrained speech recognition applications. Nevertheless, such simplicity permits efficient model learning (training) from data, which is responsible for its success in strongly-constrained speech recognition applications that contain only a sparse space of phonetic confusion.

FEATURE-BASED PHONOLOGICAL MODEL

One approach to revolutionizing the phonological model $\mathcal{P}$ in speech recognition is to adopt speech units which are based on overlapping phonological features. The features are common across languages. They exploit relations and similarities of feature components across languages, thereby offering opportunities to share observation data across languages and to generalize the observations from a source language(s) to a different, target language. One key element in constructing the feature system for use in speech recognition is to appropriately represent the possible feature sequences with their temporal evolution or statistical feature-overlapping pattern responsible for producing the speech utterances corresponding to word sequences (for any arbitrary language). For American English, the rules are based on the syllabic structure as we have implemented them. The feature overlap pattern within consonant clusters pertaining to onset and coda are rather regular, as are the overlap patterns between onset and nucleus and those between nucleus and coda. Our current rule set disallows spreads in Tongue features between onset and coda (i.e., cross nucleus) within a syllable. For Velum and Labial features, the cross-nucleus feature spread patterns are constrained to be from coda to onset only and not from onset to coda. Feature spreads are permitted, with constraints determined by the prosodic constituent boundaries, between adjacent syllables. Once a syllable is broken down into its constituents, the size of these constituents becomes countably small and hence they are easily enumerated.

The feature-overlapping pattern can be described computationally as a finite-state automaton, where each state in the automaton corresponds to a feature bundle with no precise timing information specified. Within this framework, the mathematical operations permitted in computational phonology can be successfully applied. In particular, if a sufficient amount of data is available with detailed annotation on such information as syntactic, prosodic, morphological, lexical-stress levels, syllabic, phonemic, allophonic, and articulatory events, then a probabilistic parsing strategy can be developed to automatically construct the feature-bundle based finite-state automaton. This strategy enables optimal use of a comprehensive set of linguistic constraints imposed at multiple levels of the general hierarchical structure of speech.

PHONETIC MODELS OF SPEECH DYNAMICS

The symbolic nature of the feature-based phonological model by itself does not permit an accurate description of the observed dynamic behavior in speech patterns. An integration mechanism between the discrete valued phonological model and continuous valued phonetic model must be developed. There is a general consensus that, in human speech production, the phonological component acts in a discrete fashion to control the running of the phonetic (physiological and physical) production "machinery" which, in turn, ensures correct implementation of the phonologically defined speech production goals. This phonetic machinery has a number of distinct components including motor controller (neuromotor command generator), articulatory system (smooth motion of several articulatory organs driven largely by separate neuromotor commands), vocal tract (VT) acoustic system (speech signal generator), and the auditory system (speech signal transformation). Needless to say, these components in human speech communication need to be drastically simplified at the current stage in any functional, computational model, but the key dynamic character of the process must be faithfully preserved and the dynamic model's parameters must be carefully and accurately learned from observable data in as much a physically meaningful way as possible.

Our research group at Univ. of Waterloo have been guided by this general principle during the past several years in pursuing research on various forms of the dynamic phonetic-interface model. The three main forms, differing with respect to the distinct levels at which the object of dynamic modeling is posited, have been developed. First, the acoustic-dynamic model based on trended HMM attempts to condition the properties of the dynamics directly on specific feature-coded speech production mechanisms. In that model, the underlying articulatory-feature based phonological units are used to determine either a dynamic or a static trajectory (via the differing orders of polynomial used as the trended function) that describes the acoustic correlates of the phonological units, and substantial phonetic recognition performance improvements have been demonstrated. Second, the articulatory-dynamic or stochastic target model aims at accounting for detailed movement behavior of biomechanical articulators guided by the multi-dimensional target distributions defined in the biomechanical articulator coordinate space. Third, the statistical task-dynamic model posits the object of dynamic modeling in the space of the "task" variable which is functionally significant for phonetic implementation of phonologically defined speech production goals.

The statistical task-dynamic model we have developed is based on the use of either VT constriction or VT resonance as the "hidden" dynamic variable. The dynamic process can be written as a second-order, target-directed state equation, with the continuous valued state providing the input to a static nonlinear function that results in observation speech acoustics. This statistical nonlinear dynamic system model is employed to describe aspects of the physical process of spontaneous speech production, where a large amount of speech knowledge about the VT constriction or resonance dynamic behavior in speech production is naturally incorporated into the model design, training, and decoding/scoring. The statistical nature of model design allows the computation of the probability for acoustic observations of speech, in a more accurate fashion than the conventional HMM has provided. Such a model consists of two separate components which accommodate separate sources of speech variabilities. The first component is a smooth dynamic one, linear by nonstationary. The nonstationarity is described by left-to-right regimes corresponding to phonological units. This way of handling nonstationarity is very close to that by conventional HMMs, but for each state (discrete as in HMM), rather than having an i.i.d. process, we have a phonetic-goal-directed linear dynamic process with physical entity of the state variable (continuous). Equipped with the physical meaning of the state variable, variabilities due to phonetic contexts and to speaking styles are naturally represented in the model by varying model parameters. (This contrasts with the conventional HMM approach where the variabilities are handled by expanding the total number of model parameters.) The second component is static and nonlinear. This component handles other types of variabilities including VT anatomical differences across speakers and channel/microphone variations. The two components combined form a nonstationary, nonlinear dynamic system whose structure and properties are well understood in terms of the general process of human speech production. The learning algorithms include ones from system theory, neural-network theory, and statistical optimization theory. The VT-resonance version of the model has been successfully used by my student Jeff Ma and myself in the six-week 1998 summer workshop at Johns Hopkins University where we demonstrated the effectiveness of this model in reducing word error rate for unconstrained, spontaneous, telephone-line speech recognition task defined from the Switchboard corpora.

SUMMARY AND PROSPECTS

The current state of speech recognition technology based on HMMs can be characterized as being successful in highly constrained tasks while experiencing greater and greater hardship as the tasks are becoming less and less constrained. For example, for conversational speech recorded in telephone lines for which human listeners typically have no difficulty in comprehension, the best recognizers in the world still produce over one third errors in the recognized words. The new paradigm of speech recognition outlined in this paper aims to overcome some fundamental difficulties of the current speech recognition technology. This paradigm is founded on a statistical learning strategy driven by linguistic (phonological) and physical (phonetic) principles of speech-pattern formation, as well as by functional and computational modeling of such speech patterns. It stands in contrast to the prevailing technology characterized by blind, data-driven and "ignorance" modeling where phonetic and linguistic knowledge sources are used, at best, as external constraints, rather than as intrinsic elements of the model for speech patterning. The proposed stochastic model of speech contains a compactly parameterized structure which jointly represents contextual and speaking-style variations manifested in the speech acoustics, and it provides a natural mechanism for omni-lingual speech recognition.

The research program described here emphasizes the notion of structural learning of speech-data generation mechanisms for use in designing statistically based speech recognition systems. This notion breaks away from the blind, data-driven approach currently dominating the speech recognition technology. Our current research efforts are devoted to demonstrating the potential success of integrating structural knowledge from speech science with the statistical models used in speech technology. The research is pursued both at the theoretical and algorithmic development levels and at a practical level aiming at advancing core speech recognition technology.