

# Acoustic Modeling in Automatic Speech Recognition — Overview of Current State and Research Challenges

Li Deng

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

## 1. INTRODUCTION AND OVERVIEW

In most of the useful applications of statistical pattern recognition (e.g., automatic speech recognition (ASR), handwriting recognition, machine translation, image recognition, etc.), the desired recognition tasks are complex and difficult, for which no straightforward physical principles can easily provide satisfactory solutions. As a result, researchers seek to use whatever reliable and relevant sources of knowledge, albeit their imperfection, to make the statistical decisions that are “optimal” given such partial and possibly vague knowledge. The decisions often come down to the minimization of specific types of decision error measures; e.g., the number of errors in a string of words (sentence) or word error rate in ASR.

Given the current state of statistical pattern recognition, ASR in particular, several central issues have been and still being addressed by the research community [2]. Among them, the issue of building high-quality probabilistic models for speech acoustics, i.e., (high-fidelity) acoustic modeling, is arguably the most difficult and important one. Substantial amounts of my personal research effort in the past have been devoted to this area, and in this lecture I would like to provide an overview with the focus on the need to advance scientific knowledge and technological integration.

There have been many types of statistical models for speech acoustics developed over the past three decades. They can be broadly classified into two main categories: 1) generative models, and 2) discriminative models. Generative speech recognizers (e.g., [6, 3]), such as those based on mixture models, HMMs, and stochastic segment models, rely on a learned model of the joint probability distribution of the observed acoustic features and the corresponding speech class membership. They use this joint-probability characterization to perform the decision making task based on the posterior probability of the class computed by Bayes rule. In contrast, discriminative speech recognizers, such as those based on maximum entropy models, neural networks, and conditional random fields, directly employ the speech class posterior probability or the related discriminant function. The discriminative recognizer design philosophy is the basis of a wide range of popular machine learning methods, where some known deficiencies of the HMM are addressed by applying direct dis-

criminative learning and hence replacing the need for a probabilistic generative model by a set of flexibly selected, overlapping features. Current state of acoustic modeling in ASR is that the capabilities and limitations associated with both generative and discriminative approaches discussed above are compromised, leading to practical recognition frameworks where simplistic joint-distribution models (such as HMMs) are established to characterize the statistical properties of speech, with the complexity lower than what is required to accurately “generate” samples from the true distribution. In order to make such low-complexity, low-fidelity generative models discriminate well, it requires parameter learning methods that are discriminative in nature to overcome the limitation in the simplistic HMM structure. This is in contrast to the generative approach of fitting the intra-class data as conventional maximum likelihood based methods intend to accomplish. This type of practical frameworks has been applied to much of the recent work in speech recognition research, where HMMs are used as the low-complexity joint distribution for the local acoustic feature sequences of speech and the corresponding underlying linguistic sequences of sentences, words, or phones.

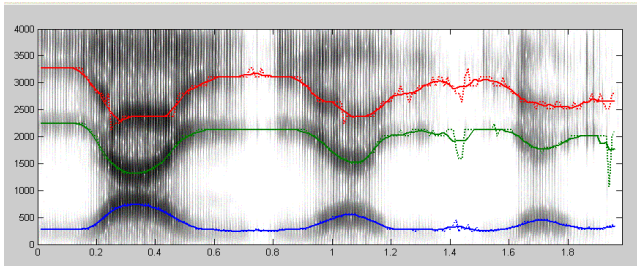
For advancing the state of the art in acoustic modeling for ASR, it is this author’s belief that both the generative and discriminative modeling approaches require acoustic models with higher fidelity than the common approaches seen today, and that their respective strengths as discussed above may be combined to achieve greater effectiveness. Current state of HMM-based acoustic modeling has intended and is able to capture only a subset of the tremendous variability in speech acoustics, often in an isolated, non-systematic way. To achieve true robustness in ASR, we need to handle all sources of the variability, including 1) pronunciation variability; 2) variability due to accent and dialect; 3) variability due to prosodic and phonetic context; 4) variability due to speaking behavior (e.g., style and rate); 5) variability due to the adverse speaking condition that affects articulation; 6) variability due to noisy acoustic environment; 7) transducer variability and distortions; and 8) transmission channel variability and distortions.

How to systematically handle all these types of speech variability in the discriminative modeling framework appears

to be less straightforward than in the generative modeling framework, partly because the much longer history of development of the latter. Even within the generative modeling framework, the HMM framework in particular, much research remains to represent and to compensate for all the main sources of variability in a principled and systematic way. In this lecture, I will use two case studies within this framework to illustrate how to account for two specific types of variability that is difficult for the conventional techniques.

## 2. CASE STUDY 1: SOLUTION TO VARIABILITY DUE TO SPEAKING BEHAVIOR

The main underlying cause for the variability in the acoustic observation of speech due to factors related to speaking behavior lies in the “hidden” domain of un-observed articulation and its control. One aspect of the variability is reflected in the phenomenon called formant target undershooting, shown in Fig. 1.



**Fig. 1.** Spectrograms of three renditions of the same (short) utterance.

The computer simulation results to be presented in this lecture show that the acoustic model which embeds the knowledge of articulatory-like constraints can effectively account for the speech variability due to a range of speaking behavior. The conventional HMMs, which may not naturally use such constraints, have difficulties in capturing this type of speaking-behavior variability in a parsimonious manner. So far, the most comprehensive implementation and evaluation of the model have been applied to the standard phonetic recognition task of TIMIT, a relatively small task due mainly to the high computational cost in decoding (not in training). The results presented in [4] show the significantly higher phonetic recognition rate (75.1%) than a state-of-the-art HMM system (71.4%). Error analysis shows that the improvements are most significant in the sonorant class, followed by the stop-consonant class.

## 3. CASE STUDY 2: SOLUTION TO VARIABILITY DUE TO ADVERSE ACOUSTIC ENVIRONMENT

In the second case study, we focus on another major type of speech variability, that due to the adverse acoustic environ-

ment with both additive and convolutive (with short-term impulse responses) distortions. Handling this type of variability has high practical value since it is directly related to the deployment of speech recognizers. Environment robustness in speech recognition remains an outstanding and difficult problem despite many years of research and investment. The difficulty arises due to many possible types of distortions, including a varying degree of additive and convolutive distortions and their mixes, which are not easy to predict accurately during recognizer deployment. As a result, the speech recognizer trained using clean speech often degrades its performance significantly when used under noisy environments if no environment-robustness strategy is applied.

Traditionally, the acoustic model for environmental distortion ignores the phase asynchrony between the clean speech and the mixing noise. Such a “low-fidelity” model has been improved, over the past several years, to achieve “higher fidelity” that removes the earlier simplifying assumption by including random phase asynchrony in the distortion model [1, 5]. In this lecture, I will first give detailed derivation of the phase-sensitive model. Then, I will summarize the existing experimental results that illustrate performance gain by moving from the “low-fidelity”, phase-insensitive model to the “high-fidelity” phase-sensitive model, and offer insight to understanding the roles of incorporating the phase information by analyzing the experimental results.

## 4. REFERENCES

- [1] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, “HMM adaptation using vector Taylor series for noisy speech recognition,” Proc. ICSLP, Vol.3, pp. 869-872, 2000.
- [2] J. Baker, L. Deng, S. Khudanpur, C. Lee, J. Glass, and N. Morgan. “Historical Development and Future Directions in Speech Recognition and Understanding,” MINDS Report of the Speech Understanding Working Group, NIST, 2007. <http://www.itl.nist.gov/iad/894.02/MINDS/FINAL/speech.web.pdf>
- [3] L. Deng. Dynamic Speech Models — Theory, Algorithm, and Application, Morgan & Claypool Publishers, LaPorte CO, USA, 2006.
- [4] L. Deng, D. Yu, and A. Acero. “Structured speech modeling,” IEEE Transactions on Audio, Speech and Language Processing (Special Issue on Rich Transcription), Vol. 14, No. 5, Sept 2006, pp. 1492-1504.
- [5] L. Deng, J. Droppo, and A. Acero. “Enhancement of log Mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise,” IEEE Trans. on Speech and Audio Processing, Vol.12 (2), Mar 2004. pp. 133-143.
- [6] L. Rabiner and B.-H. Juang. Fundamentals of Speech Recognition. Prentice Hall, 1993.