SPEECH SCIENCE AND SPEECH TECHNOLOGY

Terrance M. Nearey
Department of Linguistics, University of Alberta, Edmonton, AB T6G 2E7

1 INTRODUCTION

Speech synthesis technology has maintained a relatively close relationship with speech science since the beginning. Speech recognition technology has had a more volatile relationship with speech science. The victory of statistical pattern recognition methods (documented by Klatt 1977, JASA **62,** 1345-1366) in the ARPA sweepstakes and the subsequent success of template based and HMM systems had much to do with the development of a gulf between the two disciplines. In the face of failures of expert-system recognizers compared (e.g.) to Bayesian learning automata, many in the engineering community have concluded that it is more productive to accommodate uncertainty than to incorporate knowledge. However, the extension of speech recognition to large-vocabulary, speaker-adaptive systems has lead statistical modelers to develop architectures and heuristics that accommodate phonetic context and speaker variation in ways that are quite interpretable within a phonetic and speech science framework. This trend, if properly appreciated by both camps, may lead to a renewal of ties between the speech recognition and speech science communities.

2 SYMBOLIC CONTEXT

Consider a straw-man Model A that is a perfectly legitimate HMM, but one that no one actually uses. It assumes that the symbolic elements are phoneme-like units, represented by a single family of allophones that has no relation to its context. (This is "free variation" in linguistics 101 vocabulary.) Model A further assumes that each of these phones is realized by a sequence of one or more observation frames. These are typically 10-40 ms spectral sections or measures derived from them. Each observation frame is assumed to be independent of the others. This is the type of naive model phoneticians meet, when they are first exposed to HMMs. But the distance between HMMs in practice and phonetic theory is much less than this.

Although valid as an HMM, no one actually uses Model A because it won't work. Instead, major concessions are made to accommodate *contextual variation* that speech scientists have always insisted on. A rather standard model involves two kinds of concessions to speech science. The first concession uses what amounts to the Linguistics 101 strategy of context-conditioned allophones, in the form of what have been called "Wickelphones" (after Wayne Wickelgren's model of speech production in the 60's.)

For a full triphone implementation, each phoneme has a separate allophone for each combination of left and right phone contexts. Complete triphone sets are rarely used. Often, many triphones are not frequent enough in the training data to allow reliable estimates of their distributions. Some kind of data sharing is imposed between elements, smoothing over elements of a phoneme family. Typically this is done by numerical clustering, pooling estimated distributions over similar allophones. However, more principled, knowledge-driven methods are also sometimes used. In a recent paper by Jouvet, Bartkova and Stouff (ICSLP 94 283-286), a phonetically motivated clustering scheme out-performed a number of standard statistical clustering schemes for triphone smoothing.

Some dialog is clearly possible and it can cut both ways. The importance of clustering in triphone models may bear on the perceptual issue of exemplar-based versus prototype models. In exemplar models, every example is stored and new tokens are classified on the basis of distance to previously learned examples. In a prototype model, only an abstract summary of a category is stored. Speech recognition research demonstrates that enumerating contexts may exact the heavy price of inadequate generalization. Heavy smoothing of triphone models moves them further from exemplar-like toward prototype-like behavior.

3 STIMULUS CONTEXT

Apart from symbolic context (allophones), which are fully legitimate additions to HMMs, there other concessions to context. HMM theory requires that successive observations are conditionally independent of each other. But this assumption (essential to the strict Bayesian interpretation of HMMs) is deliberately violated for better performance. Two "standard violations" are 1) to allow massively overlapping analysis frames 2) to code *delta coefficients*, involving rate of change of properties. These can span up to 50 ms and often cross phoneme boundaries, so that the last state of a consonant HMM gets to preview information about a following vowel and *vice versa*.

Speech perception research provides a rich source of information on how human listeners use context in decoding speech signals. Careful study of existing evidence may help develop perceptually motivated accounts of context that can be engineered more forthrightly into stochastic models. There are already a fair number of hybrid neural-net plus "relaxed" HMM models that make more elaborate attempts to deal with context in a manner that seems more plausible to researchers in speech perception (e.g., Afify, Gong and Haton ICSLP 94, 291-293). This is one of many hopeful signs. Real progress in both camps is likely to be accelerated if the gulf between speech science and speech technology is actively bridged by workers on both sides who are willing to critically consider the others' insights rather than dismissing them as scientifically naive or as ivory-tower dreams.