



Comments on: “Towards increasing speech recognition error rates” by H. Boullard, H. Hermansky, and N. Morgan

Stephanie Seneff

MIT Computer Science, Cambridge, MA 02139, USA

Received 19 January 1996

I fully embrace the main message of this paper, that we need to allow creative alternative approaches to speech recognition to flourish even in the face of predictable degradation in recognition accuracy. The authors made many points that I resonate with. Among the issues they mention that are of primary significance to me are the use of technologies justified from human auditory perception, the use of neural network approaches, particularly in conjunction with temporal dynamics, and the attempts to focus attention on points of maximal spectral change discussed in Section 5.2.

I was fascinated with the discussion of frame likelihoods and priors in Section 3.5.1. I had not been aware of the problem with Viterbi training of ANN's in which the alternatives with the most probable priors tend to overwhelm all other candidates. I do share with the authors a certain mystification with respect to proper normalization of scores. Certainly it is clear that neural nets can't distinguish between “linguistic” and “acoustic,” so it would seem that if “linguistic” probabilities are introduced independently, they should somehow be extracted from the output scores, although in practice this is not a simple goal to achieve.

My main criticism of the paper is that the authors were not sufficiently bold in their concept of what constitutes revolutionary, as opposed to evolutionary, ideas. For the most part they seem to be happy with a frame-based approach, and happy with an N -gram language model. Both of these aspects are problem-

atic, and alternatives to these traditional methods deserve serious consideration. In the next several paragraphs I will mention a number of different ideas for areas to explore, many of which necessitate a major change in the recognition framework.

At the beginning of Section 4.1, the authors alluded to the issue of higher level cognitive processing. It seems clear to me, however, that the linguistic message of speech must be decoded by humans through a hierarchical framework in which the lowest level would be raw phonetic features of some sort and the highest level would be something like discourse state/subject matter. In between would be layers that encode syllabification, morphology, syntax, semantics, prosody, and pragmatics. Whether this encoding is done in a single hierarchy or in multiple inter-threaded hierarchies, it is clearly an essential part of human speech understanding. I am disappointed that the authors did not at least pay lip-service to a message of this sort.

Since it is my belief that recognition will be of greatest use as a conduit to computer speech understanding, it seems that we should design systems that support a tightly integrated relationship between the recognizer and the linguistic analysis component. In fact, I would welcome the appearance in the literature of successful systems where the bigram language constraint is completely replaced with a probabilistic natural language understanding component. If higher level concepts are explicitly represented in the language model, then it becomes much more

feasible to dynamically adjust the language model probabilities to reflect the dialogue context, particularly in mixed initiative platforms. Again, I was disappointed that the authors seemed to assume that N -gram language models were the only game in town.

I would have liked the authors to have devoted more space to alternate forms of feature representation. For instance, Ken Stevens' notion of distinctive features could potentially turn out to be a breakthrough idea, especially if it were to be adapted to a stochastic framework. For example, if every phoneme were characterized in terms of manner and place, then manner-manner and place-place transitions across phoneme boundaries would be an interesting feature set to consider. By sharing common classes across multiple phonemes, we can potentially win in terms of sparse data problems. Phonological rules are also conveniently expressed in terms of feature spreading, with "don't-cares" licensing anticipatory coarticulation in an elegant way. I think it's still likely that spectral peaks play a more prominent role than other parts of the spectrum, and we should not be ready to give up yet on the idea of a modified form of formant tracking as a means towards robust recognition.

It is well known within the speech communities that the "goals" are often on the outer tails of the "speaking-rate" distribution. I would guess that the phonetic realizations for some subset of the phones are highly dependent on speaking rate. The frequency of firing of certain phonological rules is surely speaking-rate dependent as well – for example, flapping of alveolars is probably far more prevalent in fast speech, and slow speakers probably insert many epenthetic schwas after final nasals and other consonants. Certainly we need to promote some serious scientific studies into the effects of speaking rate, before we can hope to incorporate them into our recognizers.

A probabilistic phonological rule framework in which rule probabilities can be adjusted on the fly based on speaking rate would offer a hope of overcoming some of the speaking rate difficulties. With a sufficiently sophisticated system, there would potentially be high payoff in the area of accents and dialects as well. For example, the British accent includes a characteristic vowel deletion along with

syllable reduction for many four-syllable words with stress on the first syllable. A typical example is "military" pronounced as "military". If there is a rule to capture this phenomenon, whose probabilities can be adjusted on the fly, and if there is a corresponding meta-rule for British accent that enhances its probability, then adjustments for this kind of change can be made without altering the acoustic models. This would represent a novel attack on the problem of phonetic variability.

I was also hoping to see more discussion on the issue of spectral normalization. The point was certainly made that temporal derivatives are important, at least at an 80 msec time frame, but I think other forms of spectral adjustment may also have high payoff. For instance, most systems typically take the divide-and-conquer approach to male–female differences. However, if we could instead shift the spectrum/cepstrum in an intelligent way (e.g., Bark scale spectral shift based on mean F_0), we might be able to pool all data into shared normalized models, thus ameliorating the sparse data problem and improving efficiency.

I have seen many examples of speech recognition hypotheses where the raw acoustic confusions are forgivable, but the prosody/stress pattern is completely wrong. This is perhaps not surprising, since current systems typically ignore stress markings and make no explicit use of fundamental frequency. Even duration is way under-utilized in current systems, and this is because factors that affect duration – speaking rate, stress, prepausal lengthening, consonant clusters, gemination, etc., are not conveniently available within frame-based frameworks. I suspect that the relative duration of higher level constituents (for instance the stressed vs unstressed syllables of a word) may be very useful parameters for improving recognition, if a suitable framework could be devised for identifying them. The fundamental frequency of voicing is surely an important parameter for human speech perception, but I know of no recognizer for English that utilizes it at all. Perhaps we should begin by pursuing recognition of tone languages such as Chinese, where many low level phonemic distinctions are based almost exclusively on F_0 differences.

Incorporation of many of the higher level constraints discussed above would necessitate a recogni-

tion framework that explicitly acknowledges segmental units well above the terminal phone units. This means a major overhaul of the current recognition system structure, in favor of a much more

sophisticated framework based on multiple intercommunicating layers of organization, progressing up the hierarchy towards a gradually widening temporal and linguistic scope of influence.