Presented at ESCA workshop of Robust Speech Recog., Pont-a-Mousson (1997), pp. 43-44.

ROBUST FEATURES AND ENVIRONMENTAL COMPENSATION: A FEW COMMENTS

Nelson Morgan
International Computer Science Institute, 1947 Center St, Berkeley, CA 94704
University of California at Berkeley, EECS Department, Berkeley, CA 94720
Tel: (510) 643-9153, FAX: (510) 643-7684, Email: morgan@icsi.berkeley.edu

1. GENERAL COMMENTS

This is a brief note to comment on a few points related to two excellent keynote papers by Greenberg [3] and by Stern et al [5]. In a sense, Stern's paper describes the current technology; in particular, approaches to adjusting ASR systems based on phone or sub-phone-based HMMs in order to improve performance in the presence of noise and linear channel effects. On the other hand, Greenberg's paper gives a direction for the future, focusing on aspects of spoken language that he does not believe our current systems incorporate. At first glance, the papers might seem almost unrelated. Greenberg's paper focuses on characteristics of conversational speech that indicate limitations of current ASR technology. He suggests a wide-ranging multi-tiered strategy as the fundamental solution to the poor performance that is observed for unexpected testing conditions with machine recognizers. Stern's paper is descriptive of the approaches to noise and channel robustness developed at CMU and elsewhere over the last decade, and as such is a good review of what can be done with the techniques that Greenberg criticizes. The papers are not really contradictory; faced with the requirement of improving recognition performance a good engineer will both consider new directions and also maximally exploit the existing ones. The CMU group has placed considerable emphasis on exploiting a range of solutions to linear disturbances, including both model-based and feature-based compensations. When information about the nature of the disturbance (or about the "clean" signal) is available, methods pioneered by the CMU group show the extent to which the problem can be reduced. Other methods show how iterative approaches (EM) can be used to improve the probability estimates despite interfering signals or convolutional error. We do not yet know what engineering techniques will be required in order to implement a system incorporating all the levels that Greenberg suggests, but when we do it is likely that a real implementation will be statistical, and as such will still require mathematical characterizations such as the ones Stern presents (though perhaps not these same ones).

2. FEATURES AND MODELS

Stern's taxonomy of compensation strategies consists of three classes of approaches: feature modification to match an undegraded signal (which he calls *empirical*, and which will not be discussed further here); model-based compensation, in which statistical model parameters are modified during testing; and what he refers to as *cepstral high-pass filtering*, which will be discuss further in the next section. Model-based compensation highlights a deceptively simple though crucial notion: namely, that the statistical models are a critical part of the problem of robust recognition. Methods such as CDCN and the more recent polynomial expansion approaches are techniques that are used to optimize systems based on phone or sub-phone-like HMMs; as we expand to more of Greenberg's "tiers," will we not still profit from such methods? Perhaps even more fundamentally, it may not even be possible to benefit from additional levels of knowledge without the development of the appropriate techniques for adjusting the statistical models under less than ideal conditions. Our field has had many examples of wonderful ideas at the feature or lexical levels that in some sense *had* to be correct, and yet did not provide improvements when simply applied to existing statistical structures.

3. CONNECTION: RASTA, CMN, AND SYLLABLES

There is a strong relationship between the approaches referred to by Stern as cepstral high-pass filtering and the syllabic perspective suggested by Greenberg. Two common forms of the filtering case are bandpass filtering in the log critical band domain (RASTA) or in the log-like domain (J-RASTA, used for noisy speech). Both of these actually use bandpass filters (though some sites have implemented versions with highpass filters), and the time smearing due to the filters is typically significant for hundreds of milliseconds; in some sense the local log spectrum is compared to a reference of the previous syllable or two to reduce the influence of spectral modulations that are slower or faster than the region of interest (typically 1 to 12 Hz for most RASTA implementations). Cepstral mean normalization (CMN) has been implemented in many ways and over many time ranges, but it also implies a comparison between the current time-varying log spectrum and the log spectrum from a time region that can vary from a hundred milliseconds to many seconds, depending on both implementation and the particular speech token. Thus, both RASTA and CMN are simple feature-based approaches to incorporate longer stretches of time in the estimates. They are almost degenerate cases, since the main thing the long time stretches are used for are to (essentially) estimate a degradation to be removed; however, they are a start. As Stern has shown, further improvements can be achieved by incorporating knowledge of the structure of one's statistical models in the design of the compensation algorithm. The next challenge is, can this be done in a more complete way for the kinds of "multi-tiered" structures suggested by Greenberg?

4. MULTIPLE SAVANTS

Greenberg suggests that the use of multiple strategies (for example, incorporating representations of such levels as syllabic and phrasal length acoustics) is the essential element in the robustness of human speech recognition to the vagaries of both natural speech and environmental acoustics. One could imagine that each of a group of knowledge resources might be very ignorant outside of a limited range of expertise,[1] but whose merged knowledge was quite robust to a range of potential sources of variability; Greenberg suggests that the overall process is one of deduction from these coarse sources of evidence. In a more limited framework, the subband perspective suggested by Allen [1] [2]. is a recent example in which each "savant" is an estimators that provides recognition information given a fraction of the acoustic spectrum. While multi-tiered approaches beyond the current limited mainstream paradigm have certainly been proposed before (Hearsay's "blackboard" approach comes to mind), the emphasis has been on high level sources of information, such as the pragmatics of a limited domain. We now can also incorporate statistical regularities for a range of acoustic levels, and the application of statistical modeling to these levels should give us opportunities that were not available before. We should not minimize the significance of having orders of magnitude more computing power than was available to the Hearsay researchers, as well as having large amount of available natural speech. Another exciting aspect of the current situation is that we now are starting to learn more about natural human speech, for instance speech recorded from telephone conversations. Greenberg has given us some preliminary distributional information from such a corpus (Switchboard); we need to proceed from this to develop an understanding of the acoustical models that will be required to represent longer stretches of time than we have been accustomed to modeling. Perhaps most importantly, we will need to develop algorithms for optimally combining these different levels.

5. ROBUSTNESS TO SPEAKING STYLE VS ENVIRONMENT

A final point concerns the relationship between strategies to combat degradations to recognizer performance due to (a) environmental acoustics and (b) speaking style. [5] refers to the former while [3] refers to the latter. Given the specific models that were developed for particular linear signal degradations in [5], there would appear to be no way to generalize to conversational speech style as a source of recognition errors. On the other hand, the future directions suggested in [3] could potentially improve recognition under many conditions. Should we focus on building a better overall model and disregard the specific effects of a particular degradation, or should we take each degradation one at a time and attempt to fix it? Prudence suggests a hybrid (naturally). It seems necessary to expand our models to include richer sources of information; however, it also seems necessary to include periodic sanity checks by testing on a range of degradations. Perhaps it is time to develop a set of standard tests that are performed for new approaches - like the DRT, perhaps, but with a range of bandwidths, noises, reverberations, speaking style ... training would always need to be cross-vocabulary so that the effects were not dominated by high level linguistic concerns. With such tests we might find it easier to proceed with the challenges suggested by these papers.

REFERENCES

[1] Allen, J.B., "How do humans process and recognize speech?," IEEE Trans. on Speech and Audio Processing, vol. 2, no. 4, pp.567-577, 1994.

[2] Bourlard, H., Dupont, S., and Ris, C., "Robust Speech Recognition Based on Multi-Stream Features" Proc. of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels (this volume), April, 1997.

[3] Greenberg, S., "On the Origins of Speech Intelligibility in the Real World," Proc. of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels (this volume), April, 1997.

[4] Hermansky, H., "Multi-Stream Classifiers and Syllable-Length Temporal Evidence in Handling Unkown Sources of Non-Linguistic Information" Proc. of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels (this volume), April, 1997.

[5] Stern, R., Raj, B., and Moreno, P., "Compensation for Environmental Degradation in Automatic Speech Recognition," Proc. of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels (this volume), April, 1997.

[1] One is tempted to call these resources *experts*, but the term has become "loaded" of late, often referring to a specific set of statistical approaches that have been developed recently.

[2] See, for example experiments by Bourlard [2] and Hermansky [4] from this proceedings.