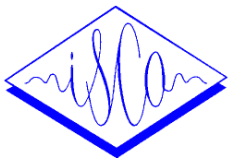
ISCA Archive
http://www.isca-speech.org/archive



Robust Speech Recognition for
Unknown Communication Channels
Pont-à-Mousson, France
,April 17-18, 1997

# Speculations on Knowledge Versus Data in ASR

*Hynek Hermansky*

*Oregon Graduate Institute of Science & Technology, Portland, Oregon*

## ABSTRACT

**For most practical purposes almost unlimited amount of transcribed training data is available in radio and TV broadcast and it appears that there are limits to capabilities of current weakly constrained HMM recognizers. There is a renewed interest in implementing more knowledge into ASR systems. This short communication expands on Chin-Hui Lee's conviction that "...robust speech recognition cannot be solved simply by collecting more data" [1] and attempts to argue for more attention to techniques which could extract reliable, reuseable, and relevant knowledge from currently available large amounts of speech data.**

## 1. Some History

### 1.1. The Seventies

Early attempts for large vocabulary continuous speech recognition emerged from noble beliefs that a variability observed in distributions of parameters of phonetic classes could be dealt with by complete understanding of sources of extralinguistic variability. Intuitively determined features were derived from the signal and elaborate systems were used to combine information for the final classification. These (sometimes called "knowledge-based") techniques were typically hand-crafted on a particular problem and were mostly applicable only in extremely controlled environments.

### 1.2. The Eighties

An important paradigm shift occurred in the early eighties when statistical Hidden Markov Model (HMM) based ASR started to dominate the field. A typical HMM ASR system is powerless without extensive training. However, when exposed to a sufficient amount of training data, it typically outperforms knowledge-based ASR approaches.

The "ignorance-based" [2] statistical approaches are not ignorant at all. They recognize the inherent uncertainty of classification, and attempt to derive class decision boundaries which minimize classification error on training data. When designed and trained well, they contain true knowledge acquired without many preconceptions about the speech communication process from real speech data.

When the statistical ASR was introduced, it appeared that the road to progress was to get hands on as much speech data as possible and to be able to process the training set in a reasonable amount of time.

### 1.3. The Nineties

There has been some progress since the Radio Rex [3]! Thanks to succesful commercial efforts in telephone industries, it seems that we can finally recognize a few words and we are aspiring for much more. The speech community is being challenged by the availability of large amounts of "found speech" data [4] and few problems with current statistical approaches are emerging:

- The classifier trained on large-variance data is only **globally** optimal and not necessarily performing too well on any of the particular sub-problem [5].
- To optimize the performance, the underlying models are relatively weak (multi-gaussian), systems grow large and demands on training data are extensive. Generalization on new unseen problems is quite poor.
- The way the knowledge is represented in the system is not always transparent. Techniques for re-use of the acquired knowledge are not well developed.
- In strive for the best performance, hand-crafting of model structure and empirical "fudging" of probabilities from subsystems is common.

The underlying original philosophy of statistical ASR is that "the data should speak". However, since there is currently no principled way to derive the structure of the model from the data, the hand-crafting of the HMM model is in some way reminiscent to procedures from knowledge-based ASR, and making it vulnerable to similar problems.

[1] Refer to C.H. Lee in these Proceedings.
[2] The origin of this term is claimed by at least two speech researchers.
[3] Refer to JASA 46(4), pp.1049-1050.
[4] Speech recorded from radio and TV broadcasts.
[5] One could in principle compete in both down-hill and cross-country using the same back-country telemark skis, but would it be an optimal strategy?

## 1.4. Today

We realize that at least some sources of undesirable variability can be identified and their effect alleviated [6]. In particular, the effects of variable communication environments are better understood. This leads to current activities in robust feature extraction, and feature and model adaptation. In some cases, the effect of the undesirable variability in a given feature space is trivial and relatively simple techniques are efficient in reducing it. In our laboratory, the efforts aim at exploiting syllable-length correlations for this purpose.

Fo situations when a good model of the competing information source (e.g. noise) is available, an interesting and apparently efficient technique is the Parallel Model Combination which, rather than reducing the unwanted variability, treats its source as another simultaneous legal information source which can be recognized and then discarded in the final decision.

# 2. Future?

It still may be that the speech signal interpreted by a system of human speech perception may carry enough information for a reasonable self-normalization and practical phonetic classification in presence of nonlinguistic variability. After all, the fact that we communicate by voice could support this belief.

The dominance of statistical approaches is partly due to the fact that our knowledge of human speech communication processes is at best incomplete and sometimes outright wrong. I believe that we should not accept that as a permanent state of affairs. Today's explosive developments of computing and data acquisition technology are mainly utilized for blind training of ever larger and larger ASR systems. Applying these new resources to the old problem of acquiring permanent and reuseable knowledge of human speech communication process which could be used in improving practical ASR may be the right thing to do. Changing the ways we do and reward ASR research may also help here. The premium should be paid on efforts and achievements which aim for advancing our knowledge, applicable to realistic ASR problems.

We may need to accept the fact that not all sources of undesirable variability are simple to model and easy to identify. There is no doubt that stochastic classification techniques present powerful means for dealing with such sources of random variability and as such they need to be respected, studied, and utilized. However, they should not serve as the lazy man's tool for alleviating the need of better understanding and better utilization of speech-specific knowledge. Some of the knowledge is common sense, e.g.:

- It is unlikely that significant amounts of linguistic information could be found in speech components which are hard to hear by an average listener, difficult to produce or control by an average speaker, or vulnerable to natural environmental noises and casual pronunciations. Thus, better knowledge of the process of human speech production and perception may point to natural constraints for sensible speech processing.
- Feature affected by truly random perturbances are unreliable for the classification. Here, techniques for handling partially degraded and missing data need to be studied and applied to avoid catastrophic failures in realistic communication environments.

[6] Both keynote papers in this session serve the evidence.

## 2.1. Towards knowledge from data

One place to put the knowledge is a signal processing module. My bias here still is towards auditory-like processing using available relevant knowledge of human speech perception. However, current efforts in data-driven design of speech processing in my laboratory at OGI go beyond the straightforward application of speech perception knowledge. We aim for improvements in acoustic modeling by analysis of relatively large amounts of speech data and for finding links between this improved acoustic processing, properties of speech data and properties of human auditory perception [7].

While the deterministic signal processing module is one obvious part of the system where the knowledge can be applied, there is also a lot which can be done in re-structuring the current phoneme-based statistical recognizer The piecewise-steady-state representation of speech is not realistic. Temporal information is mishandeled by time allingment techniques. Coarticulation is not well modeled. Variable temporal structure of sub-word units is not respected. Overall, the dynamic nature of the dialogue is not accomodated.

It may be that spending time on studying trained statistical recognizer could pay off in enhancing the structure of the next one. For all I know this may be routinely done by some but we should devote more time to discussing and interpreting the obtained results, reporting efforts which failed to perform to expectations and discuss reasons for such failures.

The knowledge hardwired into a design of the recognizer is the knowledge which does not have to be re-acquired from the data every time the recognizer is used for a new task. Based on our experience we feel that available speech databases contain more information (knowledge) then currently being utilized for structuring of ASR systems. Such knowledge acquired from data could serve as a knowledge source for building more structure into current ASR systems so that they are better able to handle variability introduced by realistic communication channels.


## Acknowledgments

The paper is based on discussions with Carlos Avendano, Noboru Kanedera, Narendranath Malayath, Misha Pavel, Sangita Tibrewala, and Sarel van Vuuren, and motivated by keynote papers by Chin-Hui Lee and by Mark J.F. Gales to the ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-a-Mousson, France, April 1997. Many things presented in this short response paper were said before elsewhere by others, still I feel they can be said again. The limited space is the reason for no references to literature in this article. Autor's perspective on data-driven acquisition of knowledge benefited from article by Fred Jelinek in Speech Communications, Vol. 18, No. 3, May 1996.


[7] There are two papers from our lab in these proceedings describing some of our efforts in data-driven design of temporal processing of speech.