



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

J. Mariani, J.L. Gauvain, L. Lamel

*LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France*

Received 9 February 1996

The goal of this article, an extended version of the presentation at Eurospeech'95, is to incite the speech recognition community to sit back and take a look at where we were, where we are, and where we are going. We certainly agree that this type of reflection could benefit the field as a whole, and wholeheartedly support the basic idea of needing new methods to solve the outstanding problems in automatic speech recognition. In particular, we appreciate the discussion of problems linked to currently used approaches, as well as the summaries of the selected techniques which may eventually help address these problems. This being said, we have some reservations about the manner in which the authors present research in the field and in their negative view of “popular research themes” over the last five years. In fact, it appears to us that the authors had some of the same reservations, as evidenced by the contradictions they have in the paper.

The work presented deals mainly with problems in classification and acoustic modeling – but there are many more outstanding research areas such as speaker variability, effects of speaking rate, phonological variants, lexical modeling and language modeling. We will come back to some of these points later, but first we summarize our understanding of the view put forth by the authors.

Their proposed research strategy is the following:

1. Stop improving your state-of-the-art system, because you probably reached a local minimum.

2. Try new models based on your expertise.
3. Improve your new model for a while hoping that you will get better results than the original state-of-the-art system.
4. If you do, go back to step 1, if you don't go back to step 2.

Steps 2 and 3 usually mean changing something in your system by something else that seems more promising and try to make it work. This is a very common strategy used by any researcher in the field. What is different is the scale and nature of the change. It is also our experience that most of our ideas result in an increase in error rate. So we may conclude that the message of the authors is that in step 2 we should make radical changes to play for the jackpot and don't hesitate to pursue those with very high error rates.

Unless you have strong prior knowledge about the power of your new model, this can be considered as a game theory problem, where you need to optimize the long-run expected return. All depends on the shape of the error rate surface, which cannot be explored exhaustively.

The authors point of view is not new and is shared by many other researchers. Conducting research on innovative approaches should be encouraged. But it is also necessary to test the results at some point in order to assess the interest of the approach in an objective way. What can't be avoided is confronting the new model with data.

The evaluation paradigm is often criticized, mostly by those who never used it. It has been said for example at the NAS (National Academy of Science) colloquium on Human–Machine Communication by Voice (1993) that using it is like using a very powerful lamp to look for an object in a dark room, focusing on a spot while the object stays in the dark beside. But refusing to use objective measures of performance is like looking for the object without any lamp!

This was obvious in the period between the end of the first ARPA project (1976) and the start of the new one (1984). It seems that many potential benefits of a decade of research effort worldwide may have been lost, as there were no ways for comparing systems developed in the first ARPA project or after. Having such evaluation tools would have greatly reduced the time to accept the new HMM paradigm, compared with “pattern-matching DTW based” or “knowledge-based” approaches (as mentioned by the authors themselves at the end of Section 3.1!).

It is certainly acceptable (and necessary!) to criticize the evaluation paradigm in order to improve the way it is used. But it is important to keep in mind that the paradigm offers many advantages. Crudely criticizing the evaluation paradigm may induce people who are unfamiliar with it to reject it a priori, without taking the time to discover the positive aspects such as objective evaluation, contrast of methods, and sharing reliable information among participants.

We draw from the authors criticisms of the “standard approach in our field” that commonly used approaches, such as those based on HMM are stuck in a local minimum and there will be an eventual platforming of performance, from which further improvements will not be made. However, as demonstrated in the yearly ARPA evaluations, it is not at all clear that such a local minimum has been reached – or will be reached. Each year significant progress has been made on increasingly more difficult tasks by building on existing tools. Why change a winning team? (The authors attribute this improvement to the use of more training data, but this is only partly true. Even on tasks where the training data are limited, improvements have been reported.)

We note that most of the approaches proposed by the authors are based on speech perception, in con-

trast to HMM which is a model (albeit crude) of speech generation. This fundamental issue is unlikely to be agreed upon, but there is no reason to believe that models based on perception should outperform those based on production. Neither the ability to hear nor to create sound was particularly developed for speech. Speaking and listening have to make do with the existing apparatus, developed for other reasons.

We would like to assure the authors that they are not the only ones able to increase the error rates, we do it too, but we are usually not too happy about it. Most of our ideas result in word error increases not only the first time we try them. The art in research is to decide, after a certain amount of time (measured in experiments), whether to continue pursuing this idea, or to figure out why it doesn’t seem to work and to move on to the next one.

The most striking point of the paper is that more scientific (and radical) ideas should be pursued, rather than small perturbations on something that seems to be working. The authors say that “approaches that reduce the word error rate ... effectively leads to the suppression of innovation”. Of course, this is untrue. It is perfectly acceptable to conduct research on innovative approaches. But it is also necessary to test the results at some point in order to assess the interest of the approach in an objective way. We believe, like the authors, that progress will result from both approaches. Another title for the paper could be “Here are some ideas that can be interesting to pursue” however this title is certainly less “catchy”.

There are many other factors that are observed to influence the speech recognition performance that are not addressed in the paper. Some of these unsolved problems are:

- *Inter-speaker variability*. Even today’s best systems have a huge difference in performance (sometimes as much as a factor of 30) between the word error of the best speaker (1–2%) and the word error of the worst speaker (25–30%). What demonstration is there that novel techniques will be able to reduce this huge difference? What about dialect modeling or dealing with the speech of non-native speakers? As humans we usually are able to quickly adapt to different accents.
- *Speaking rate*. Speakers that are much faster or slower than the norm tend to have much higher

word error rates. How do we account for these differences? Differences in speaking rate affect not only the acoustic level, but also the phonological level and maybe even word level.

- *Phonological variants.* Different speakers make use of different phonological rules. For most speakers, the choice of rules is systematic. No system that we know of is able to make use of this consistency.
- *Lexical representation and modeling.* The most common way to deal with pronunciations variants is to include these directly in lexicons. In many systems this requires manual work and careful verification. In other systems, pronunciations variants are generated by rules, but this has the well-known problem of overgeneralization.
- *Language modeling.* The  $n$ -gram language models which are reasonably successful for English, are less efficient for more highly inflected languages (such as French and German). There is certainly a lot of research that needs to be carried out in this area. A related issue is how well will we be able to recognize speech without understanding? We don't know this limit.

We now give some more specific comments on the paper including some positive aspects of the research carried out over the last 5 years which were not expressed by the authors.

In the introduction the authors provide 3 conditions under which “Permitting an initial increase in error rate can be useful ...”, but they are missing a 4th condition. (4) evaluation on common tasks so as to be able to demonstrate the increase in error rate!

We are quite sensitive to the issues the authors have raised regarding funding problems. This of course is a problem that we all face – even what may be referred to as “demonstrated technology” is difficult to get funded (either it is too advanced – industry should now be funding the research - or the problem is too hard, and can never be solved).

However, it doesn't appear that non-HMM approaches are not funded by the CEC. On the contrary, Wernicke was funded as one of the very few (2) Esprit BRA speech projects in FP3, and as its continuation, Sprach, is the only on-going FP4 Esprit LTR project on speech. The first author managed the Wernicke project, and is managing the second.

Proposals with systems based on an HMM ap-

proach were turned down on the basis that it was not research! This is not fair, as various approaches should be investigated and compared, not only the HMM/ANN one. What is missing is an evaluation infrastructure at the European level, in a multilingual environment, such as a continuation and extension of the Sqale project which successfully tried it on a small scale.

In Section 2.1 the authors argue that increasing the size of the training database does not constitute research, that simply having more data improves performance. While it is certainly true that increasing the size of the database usually improves performance, the criticism is too simplistic. The amount of improvement depends on what you do with the data. As a community we can still learn from the improvements obtained by “simply increasing the amount of training data”. Experiments carried out in the Nov93 ARPA test clearly demonstrated the capability of certain techniques (CDHMM) to better take advantage of the additional training data than other techniques (tied-mixture and discrete HMM).

Perhaps a more important point that can be made, is that constructing corpora that are representative, complete, yet at the same time not too big is an open research area. This is an area considered “un-glamorous” and even “soft-science” by many researchers in the field – it is hard to publish work done in the area and extremely hard to demonstrate the effects of different corpus design strategies. Yet at the same time, the performance of all recognition systems is acknowledged to be quite dependent on the training data.

In the same section, the authors discuss smoothing criteria. Given that our data will never be representative enough, we will always have the problem of parameter smoothing. There are different techniques to deal with the problem, such as parameter sharing, not only smoothing criteria. In fact, many of the state of the art systems do not use smoothing, as parameter sharing has been shown to be more promising to deal with the problems.

Regarding the lexicon size, the authors ignore the most important issues in selecting recognition lexicons. The vocabulary list needs to be selected so as to minimize the expected errors due to out-of-vocabulary words, and pronunciations must be derived for all the lexical entries. Issues in lexical design have

not received much attention from the community at large.

We agree that “knowing when we don’t know” and “real-time recognition” are important research topics (but who wouldn’t agree?). We do believe that being able to address those two issues well enough may conduct per se in large increase in performance, as it would allow for labelling large amount of spoken language data in order to build language models, in the case when large enough text corpus doesn’t exist.

As we stated at the beginning, there are some evident contradictions in the article, that suggest to us that the authors had a hard time sticking to the theme of the paper – that is techniques to increase the error rate.

The authors take as a joke the improvements allowed by using more data, while they also mention that more data is necessary for their own hybrid approach!

The authors also take as a joke the use of the evaluation paradigm, while they use it to point out success of an approach similar to theirs (the Abbot system), based on the good results obtained by this system in formal tests (ARPA and Sqale).

However, the authors are quick to point out the improvements they made with the SPAM model on

recognition of isolated digits. They went from twice the error rate to “now as good as” their best phonetic based system in a few months.

This is just one example of reported results that seem to be in contrast with their main point that radical changes can really improve system performance. In all the examples given in the paper “comparable” or “slightly better” (slightly not being defined nor necessarily significant) results are obtained, yet the authors do not say that they have now decided to throw away these approaches as they did not have the big payoff? Can we expect this big payoff to come now, or, as we expect is more likely, slight perturbations of their approaches will yield small improvements. We would conclude that the authors are just as happy as we are whenever we obtain improved performance with unbiased testing.

Referring back to the opening Montaigne citation – what do “*new discovery*” and “*they say*” refer to? “New discovery” could well stand for HMMs and “they” for the paper’s authors!

And as for Thomas Edison, at the time only 1/1000 light bulbs worked, now it is surely less than 1/1000 that don’t work, but how many millions of steps have been carried out to improve the basic technology?