



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

Melvyn J. Hunt

Dragon Systems UK Ltd, Cheltenham, United Kingdom GL52 4RW

This stimulating paper exhorts researchers to “listen to their own internal sense of confidence” in the teeth of received wisdom and initial negative results. In terms of simulated annealing, we are thus invited to avoid local minima by raising our temperatures.

This may well be good advice, but the evidence on which it is based is probably biased. “They all laughed at Christopher Columbus” as the song says; but they also laughed at thousands of other people who really did have bad ideas. We remember and celebrate the successes, but forget the failures. If you turn out to be a gifted speech researcher like the authors of this paper, then you will look back on your previous work and conclude that having confidence in one’s own unconventional ideas pays off. We do not hear from the, perhaps larger, group of less gifted researchers who usually do not stay in the field and whose best way of contributing may well be through making marginal improvements to existing techniques. The really difficult trick is in knowing which kind of researcher you are. I remember John Bridle saying about fifteen years ago that he wished that he had more confidence in his own original ideas, but he feared that when he finally acquired that confidence he would be at an age when his ideas were no longer any good!

An early attempt to use linear discriminant analysis (LDA) in speech recognition certainly supports the authors’ call to have more confidence in one’s own ideas. In the late Seventies at BNR I was

experimenting with acoustic probability measures for speech recognition and sent off an abstract for an ASA meeting [1]. The text produced for the meeting described not only the use of LDA but also state-dependent variances and dynamic spectral features. As the date of the meeting approached, however, I had no positive results and my boss, Paul Mermelstein, asked if I had considered withdrawing the paper. I persisted in presenting it, but felt so embarrassed that the presentation was very poor and I dropped the ideas immediately afterwards.

A decade later, working on an auditory model for speech recognition, I needed a metric that would suitably combine two different representations, neither of them sufficiently similar to a conventional log spectrum to make the cepstrum the obvious choice. I was consequently driven back to using LDA. When we presented a paper at ICASSP-88 [2] favourably comparing recognition results from the auditory model with LDA against those from a conventional mel-cepstrum representation, someone in the audience (Per Heddellin?) correctly objected that in a fairer comparison the non-auditory-model representation would include the use of weighted cepstrum coefficients and dynamic spectral parameters. Within a week of returning from ICASSP we obtained results using LDA to combine dynamic and static spectral parameters and weight them suitably that were far better than anything we could achieve with the, computationally much more expensive, auditory model. We went on to discover that by sub-

jecting the training data to various distortions before deriving the LDA transformation we could make the representation surprisingly robust to a related range of distortions. We published this at the next ICASSP [Ref. [55] in the main paper].

We should mention that the use of LDA in speech recognition was developed independently by the IBM Speech Recognition Group [3]. They and we used LDA to provide a uniform representation for all states in all models, while George Doddington described [Ref. [30] in the main paper] the use of different transformations for each state.

Perhaps dropping the auditory model when the results with a conventional spectral representation were better is another example of this researcher's lack of persistence. Certainly, the analogy quoted in the main paper ("Airplanes do not flap wings; therefore automatic recognizers should not have ears!") contains a fallacy. The atmosphere was in existence long before birds evolved to fly in it – and before humans developed aeroplanes. But speech was *not* around before there were human vocal tracts to produce it or human ears to hear it. In an accurate interpretation of the analogy, ears correspond to the

air and speech to birds flapping their wings. The task of designing an automatic speech recognizer is therefore not like finding a way to fly in air but rather like finding a suitable medium for birds to fly in. It might be difficult to find anything more suitable for this activity than the earth's atmosphere – or anything more suitable for perceiving speech than the human ear.

References

- [1] M.J. Hunt (1979), "A statistical approach to metrics for word and syllable recognition", Fall 79 Meeting, Acoust. Soc. Amer., Salt Lake City, *J. Acoust. Soc. Amer.*, Vol. 66, pp. S535–536.
- [2] M.J. Hunt and C. Lefebvre (1988), "Speaker dependent and independent speech recognition experiments with an auditory model", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-88*, New York, April 1988, Vol. 1, pp. 215–218.
- [3] L.R. Bahl, P.F. Brown, P.V. de Souza and R.L. Mercer (1988), "Speech recognition with continuous-parameter hidden Markov models", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-88*, New York, April 1988, Vol. 1, pp. 40–43.