

## Reflections on artificial intelligence, speech technology, brain imaging and phonetics

Björn Lindblom

Phonetics laboratory, Stockholm University, Sweden

lindblom@ling.su.se

### Whither speech research?

Broadly defined, the field of spoken language research is undergoing rapid change as a result of several factors.

Computer technology can be used to simulate interactive human speech communication in spectacularly diverse and realistic ways.

Since the 90's sophisticated methods have become increasingly available for imaging the real-time processes that take place in the brain during speaking, listening and learning to speak.

It would seem that the present situation offers speech research yet another opportunity to expand its experimental toolbox and orient itself towards neuroscience with a view to acquiring a more profound understanding of the many forms of language use under normal as well as disordered conditions. Already many investigators have done so and have contributed significant new knowledge (Guenther 2016).

Our field presents a diverse mosaic of research interests and special expertise. Has time come for unification? How could such interdisciplinary integration be brought about?

The subject matter of a discipline tends to be defined by the questions it asks, but there are of course other factors such as tradition, ineffective administration and obsolete educational programs that create obstacles.

Those are some of the questions and issues that need to be urgently addressed and dealt with today.

### Speech technology and AI

There is an anecdote about Richard Feynman, the eminent physicist, who had the following sentence written on

his blackboard: *'What I cannot create I do not understand!'* (Gleick 1993). Let us interpret this quote as a requirement to be met by all good science. Let us call it the *Feynman criterion*. If we apply it to speech research, what do we find?

First let us consider some recent achievements in speech technology.

On an almost daily basis, we hear about new, often spectacular, artificial intelligence (AI) developments. For instance, some car models now have autopilots. Medical diagnoses and treatments can be drastically improved thanks to vast data sets and clever methods that detect patterns in the data that escape even the most experienced expert. Furthermore, law breakers get more easily caught with the aid of clever facial recognition algorithms.

And then there is *AlphaGo* - a program that has beaten the world's top human players of Go - a much more complex game than chess (Mozur 2017). Its training consisted in learning not only from records of human games but from discovering, on its own, winning and losing strategies by playing innumerable times against itself.

The list of potential benefits of AI - and disadvantages - is long (Foer 2018, Friedman 2019).

Part of the story is AI-based speech technology.

Recently we learned that attempts to improve communication for paralyzed patients are under way and show considerable promise. One such project (Akbar et al 2019) used a Brain Computer Interface technique to reconstruct input speech signals from a population of evoked neural activity in the human auditory cortex.

The investigators used a deep neural network to make direct estimates of speech synthesizer parameters. This model achieved high subjective and objective intelligibility scores on a digit recognition allowing the authors to conclude that reconstructing speech from the human auditory cortex offers a speech neuroprosthetic for direct communication with a paralyzed patient's brain. The technique is said to work under both overt and covert conditions.

### **The IBM debater project (2019)**

This project has produced a computer program that recently participated in a debate on *'whether preschools should be subsidized'*. Its performance was based on fifteen minutes of preparation, developing an argument in favor of subsidies, producing a text and then giving an oral presentation of it. A large human audience attended the event and was found to prefer the human participant, but the simulated debater came close to winning.

Listening to this AI debater one is struck by its ability to 'reason' and produce fluent and natural sounding speech (see references). The big news is not that the human debater had won the debate. It is that the IBM debater had given its human competitor a good run for his money – despite the complexity of the task. I cannot blame those who respond by thinking that these developments are nothing short of extraordinary. I agree.

But we have to ask: Where do we place them within general science? What are the implications for research on speech and language learning? Exactly how does the debater do it? It meets the Feynman criterion, but is imitation enough?

AI people admit that they do not know exactly how these systems work (McAfee & Brynjolfsson 2016). But they know enough to attribute the success of AlphaGo (and other projects) to the use of *deep learning networks*. These models make it possible to abandon what has been used so far, viz. explicitly specified, computational rules. Instead,

deep neural networks learn to a great extent on their own by accumulating and comparing a massive number of examples of successes and failures – in part found in records of human data, in part from information generated by the algorithm itself.

Such results take steps towards removing a major hurdle in the modeling of human cognition. This hurdle is created by the fact that humans *'know more than they can tell'* – *Polanyi's Paradox*. If this becomes a characteristic also of computers, it would imply a major breakthrough for AI. Machines would then have become even more human-like. Are we already there?

There is an extremely serious downside to these developments. 'Not knowing exactly what deep learning algorithms do' is problematic. It feeds right into people's worst fears about future computers evolving to surpass and dominate their human masters.

The 'secrecy' of the modeling should also disqualify them as scientific tools. They are unquestionably great imitators. But imitation is hardly enough. If these frameworks do not allow us to study the observed behavior, what use would they be to linguists, phoneticians, speech pathologists and communication engineers?

Instead of helping us understand our own brains, AI projects would create an additional problem perhaps even more forbidding: Understanding artificial brains.

### **Explanation: Holy Grail of research**

Reasoning from first principles goes back to Aristotle. We can illustrate this time-honored method by taking a look at the acoustic theory of speech production (Fant 1960, Stevens 1998).

At first one might assume that the term *'formant'* refers to an empirical notion derived from observations and analyses of large corpora of recorded speech, but it is not. Anyone who has struggled with making formant frequency measurements from short-term spectra and spectrograms knows that voice quality,

nasalization and a high fundamental frequency often make that exercise difficult. Those and other factors also cause errors for automatic methods such as LPC whose results must be checked manually for accuracy. It is true that inverse filtering can provide solutions to such problems – provided that you know what the shape of the residue glottal pulse is supposed to look like.

We love formants, but they can be very elusive!

The ‘formant’ is an independently motivated concept deduced from physics. It is not a product of a data-driven approach. Rather the acoustic theory of speech rests on the following first principles (Beranek 1954): *Newton’s second law*, *Boyle’s law* and the *Conservation of mass*. They provide the foundation for the wave equation that can be solved for arbitrary vocal tract configurations (Flanagan 1965). When the vocal tract shape is known its resonances (formants) can be identified and their frequencies can be calculated.

When someone like me (with a liberal arts education) invokes first principles, it is not an expression of ‘*physics envy*’, nor a wish to reduce everything to physics.

[I would perhaps be willing to plead guilty of ‘*science envy*’ because science is undoubtedly a good thing to be striving towards].

There are no absolute explanations, only deeper and deeper accounts. Individual fields choose their *explanans principles* at different levels. In that sense, every field adjusts its own depth according to the state of the art. However, across disciplines there are converging views on what constitutes an explanation (Miller 1990).

The thing to remember is that practical applications and general scientific knowledge both need the same fuel: Explanations.

My take is that AI projects can imitate/synthesize natural sounding speech in near-perfect ways, but unlike the acoustic theory of speech production, they do not explain anything about

human speech. Rather they hide what they do.

Consequently they leave a big piece of the scientific task undone and might force upon us the epiphenomenal task of understanding - not only our own brains - but also artificial brains and their deeply embedded organization.

### Neurobiology of speech

Students of language and speech are not the only ones trying to get used to new technology. Neuroscience - now attracting our attention - is itself in fact a current instructive example: *The Human Brain Project* (HBP) is a gigantic EU-sponsored effort in which some 500 scientists from over 100 universities participate (2012).

To motivate such a large project, the applicants argue that time has come to bring medicine, neuroscience and computing together in response to the greatest challenge of the 21st century: *the understanding of the human brain*. Their application underscores that, despite significant progress neuroscience in recent years, the field still suffers from fragmentation.

To open the door to new treatments of brain diseases – more than 500 have been identified so far – the application states that it will be necessary to show how the ‘*parts fit together in a single multi-level system*’. A two-way process is envisioned. New computing technologies will be discovered as more is learned about the brain, and conversely, simulations combined with empirical observations are expected to bring novel insights into the workings of both normal and disordered brains. With such a strategy it should be possible to come up with deeper accounts of many challenging topics including learning, memory, language, consciousness, awareness and psychiatric conditions.

Since speech research presents a diverse mosaic of research interests, applying the philosophy of the HBP project to our own field makes a lot sense. However, we need to ask ourselves: Have current models of real-time speech

reached a sufficient degree of realism to be effective for interpreting large and complex volumes of brain data? Also since brain imaging relies heavily on Big Data approaches and does not come with a theory paralleling the acoustic theory of speech, we need to proceed with caution always remembering: Priority #1: Explanation!

## Conclusion

The quality of our theories of spoken language and practical applications – clinical, educational, technological – will be a function of how well we understand how ‘humans do it’.

That, in my opinion, should be the future niche for phonetics and spoken language research.

## Acknowledgements

I thank Rolf Carlson for helpful comments on a draft of this paper.

I am indebted to Olle Engwall for sketching a number of neuroimaging projects on L2 language learners and dyslectic subjects, and for making me aware of Akbar et al’s paper.

I am grateful to Joakim Gustafsson and Jonas Beskow for giving me a comprehensive update on various impressive AI/speech-tech projects.

My correspondence with Martti Vainio gave me a valuable perspective on neurophonetics - much needed now that Stockholm University will soon be able to offer its students, faculty and guests a new facility for brain imaging: Stockholm University Brain Imaging Centre (SUBIC). I thank him for his thoughtful reply to my questions.

## References

- Akbar H H, Khalighinejad B, Herrero J L, Mehta A D & Mesgarani N (2019): "Towards reconstructing intelligible speech from the human auditory cortex", *Scientific Reports* 9:874
- Beranek L L (1954): *Acoustics*, McGraw Hill: New York.
- Fant G (1960): *The acoustic theory of speech production*, The Hague: Mouton.
- Flanagan J L (1965): *Speech analysis synthesis and perception*, 1st ed, New York: Springer
- Foer F (2018): *World without mind: The existential threat of big tech*, New York: Random House.
- Friedman T L (2019): "Warning! Everything is going deep: ‘The Age of Surveillance Capitalism’", *New York Times*, Jan 29.
- Gleick J (1993): *Genius: The life and science of Richard Feynman*, Vintage Books: New York.
- Guenther F H (2016): *Neural control of speech*, Cambridge USA: MIT Press.
- Human Brain Project*, A report to the European Commission (2012): The HBP-PS Consortium, Lausanne.
- IBM debater:  
<https://www.research.ibm.com/artificial-intelligence/project-debater/live/>
- MacAfee A & Brynjolfsson E (2016): "Where computers defeat humans, and where they can't", *New York Times*, March 16
- Miller G A (1990): "Linguists, psychologists and the cognitive sciences", *Language* 66, 317-322.
- Mozur P (2017): "Google's AlphaGo defeats Chinese Go master in win for A.I.", *New York Times*, May 23.
- Stevens K N (1998): *Acoustic phonetics*, MIT Press: Cambridge Mass USA.
- SUBIC: <https://www.su.se/subic/>