Towards biologically feasible machine recognition of speech: Notes on the discussion

Phil Garner

June 22, 2017

1 Session1

Hynek began by introducing the topic and ansking if we are asking the right questions. It seems we are.

Dick continued by explaining the ASR front end, and how he likes to replace RASTA-like processing with automatic gain control. This removes the logarithm, but brings in a power law. There is some discussion of the timescales at which this works.

If you put a non-linear function after this kind of AGC arrangement, it learns to give power at low amplitudes and compression at high ones. It amounts to optimising the parameters of

$$\left[\left(\frac{x}{\bar{x}^{\alpha}}\right)^{2}+\beta\right]^{\gamma} \tag{1}$$

This prompted a fair bit of discussion, including about the features such as pitch that are typically discarded. Dick points out that ASR is now driven by companies, so any advances will actually be deployed.

Dolores asks about multi-modal ASR; using visual information. This is used by the aerospace industry. It would also dbe useful for the hearing aid industry.

Tobias asks about the features that we discard. There is a feeling that pitch may help follow the voice, especially in noise. Perhaps also lateral position. Have people tried using pitch? Some people have tried, especially for Mandarin. Generally, pitch patterns have been used, but it's not clear why.

Alain: Pitch can be useful to separate two voices.

How much spectral distortion can we tolerate? There are some DNN structures that convolve over frequency. The pattern recognition can be trained to be robust to that.

Feedback is necessary in a dynamical system to cater for limited dynamic range. An alternative may be to run several processes in parallel, then choose the right one.

What do people use to do this? People use the background initially. But to suppress background speakers we use pitch. It's difficult to do this in parallel because it's difficult to choose the right one in the end.

There is nothing going on the field that nearly approached the parallelism that we observe in the cortex.

With two sources and two microphones it's possible to separate the sources. Either sparsity based orthogonality, or point a null.

How do we deal with microphones? Both arrays, and the sensors themselves. Can we synthesise a digital stream from the outset?

Jim described the front-end from the point of view of an inner hair cell. The open-probability can move in 10 ms. It moves for low frequencies, less for high ones. It's the tip-link motor adaptation. A sinusoid will come out high amplitude to begin with, but become rapidly smaller.

Outer hair cells will generate distortion components.

Pitch can be used in several ways: as a cue that speech is present, as a feature along with spectra.

2 Session 2

Hynek begins by describing why we want to do this:

Speech has some temporal redundancy. Movements of the vocal tract are represented at all frequencies. The ear preserves the temporal spread, but separates the different bands. We need to take sufficient temporal segments to remove the redundancy that is introduced by the vocal tract. We then need to fuse the components that remain after the removal of redundancy.

There is a perfomance monitor, but people are somehow skeptical about it. Ralf points out it shouldn't be word error rate. Malcolm asks what goes in and what comes out (cat and cat).

How does it know it's right? It doesn't; it's just a model of how it performs on training data. The prediction can be seen as a confidence. Isn't turning off channels like feature selection? We think so.

Should we measure the perfomance of intermediate (low level) results? Hynek thinks no. Alain suggests maybe.

Malcolm: why nor just build an end to end neural net? Because we want to dictate intermediate results so it works better on unseen data.

Malcolm: don't we want to train and test on the same data? Hynek: No!

Ralf: it's clear that one should be doing joint training as things are too interdependent. Then again, why should it boil down to something different in the end? Malcolm feels that one big classifier that does everything is the better way. Alain: this will make it sensitive to degradation on any channel. Malcolm feels that you can avoid this by training on more data. You don't need all the data, but you need a representative subset.

Hynek: Infinity would be good, but it's practically impossible.

Isn't the fusion network what Hinton calls a manager network? The best decisions are fusions of multiple things (mixture of experts?).

There are channels that correspond to different beams. Channels that correspond to different noise levels. This is not quite mixture of experts.

Alain: averaging beams doesn't give you good results.

Ralf: question for the brain people: If you have some corrupted input, is there evidence for how it is handled in the brain? Alain thinks it fits the hypothesis that they're ignored.

You can combine different classifiers that are different technologies. In fact, a single recogniser will do better than speaker dependent.

Hynek: if some recogniser contains me and you, and you speak, let it decide which one to choose. Malcolm would use the one trained on both.

Hynek's scheme would work better if there was a strong a-priori reason for it to do so. (der...)

An inspiration is that we have so many cortical neurons compared to nerves. Also lots of redundancy.

Some methods of combining classifiers work, some don't. Cue all manner of argument about all of the above.

Don't we ignore correlations that might be important if we have separate classifiers? Alain has an answer, but I was typing.

Back up to the top, does the performance monitor feed back? Something somethong comb filter on the target source. The performance is all contingent on the ability to separate into meaningful channels. It can be argued that the cochlea evolved to do this.

Hynek strongly argues for treating the brain tonotopically. It seems the brain works this way anyway. Tonotopy exists at multiple levels in the brain; why? Parallel processing in frequency specific channels. It actually changes from fast in the periphery to slow in the cortex.