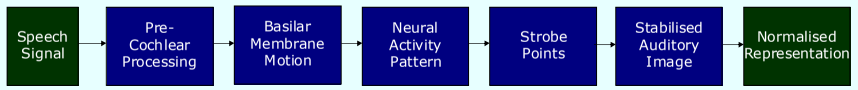
UNIVERSITY OF CAMBRIDGE
CNBH, Department of Physiology, Development and Neuroscience, University of Cambridge
www.pdn.cam.ac.uk/cnbh

Bayesian approaches to pitch-synchronous vowel normalization. P25

Thomas C. Walters and Roy D. Patterson
tcw24@cam.ac.uk rdp1@cam.ac.uk

Introduction

The Auditory Image Model (AIM) is a computational model of the early stages of human auditory processing. AIM is implemented in MATLAB.

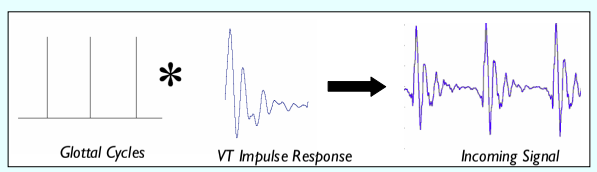


By the end of the model's processing, the representation is invariant to changes in both the glottal pulse rate (GPR) and the vocal tract length (VTL) of the input speaker. The model performs processing on the incoming sound in stages, the first stages are Pre-Cochlear Processing (**PCP**), Basilar Membrane Motion (**BMM**), Neural Activity Pattern (**NAP**) and Strobe-Points (**SP**). The **BMM** is generated by a filterbank which models the response of the basilar membrane. The **NAP** stage models the neural output from the basilar membrane up the auditory nerve. The **SP** stage attempts to find peaks of the NAP associated with the glottal pulses in the input sound. This stage can be considered as a deconvolution of a spike train from the vocal tract impulse response and from the subsequent filtering by the BMM filterbank. The **SAI** is a representation of the input sound in which the glottal pulses have been separated from the resonances of the vocal tract. A time-interval representation is built up by adding the output of the NAP to a buffer, re-starting at zero time-interval for each strobe point.

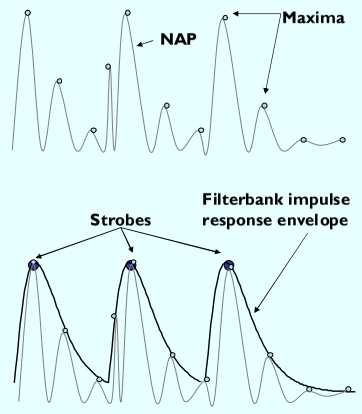
Neural activity pattern (a) and Stabilised auditory image (b) for the vowel 'aa' spoken by a man.

The Problem

In the SP stage, the algorithm should ideally find exactly one strobe point per cycle. The input signal is essentially a click train convolved with the impulse response of the vocal tract (VT) since the glottal pulses are, to a first approximation, clicks.



Due to the nature of the filterbank, the lower frequency channels have a lag in their response, giving rise to the characteristic NAP for a sound. Current algorithms use a time varying threshold on each channel independently in order to find those peaks of the NAP which were caused by glottal pulses. This works well in most cases but does not take the correlations between channels into account. These algorithms routinely find more than one strobe point per cycle, which is not optimal.



Approaches

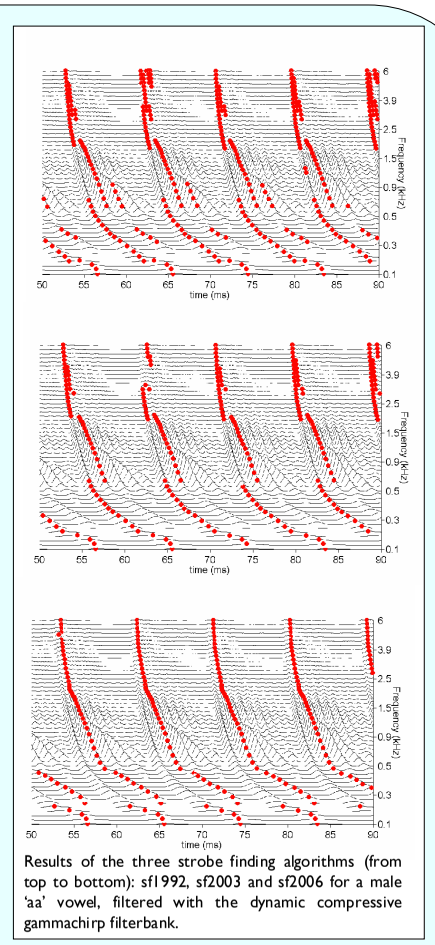
One of the keys to making AIM robust to a wide range of sounds is to ensure that the strobe-points are as accurate as possible. In order to take correlations between channels into account, and to find the optimal solution to the problem, I am developing an alternative algorithm which deals with all the channels simultaneously. In many frequency bands of the BMM, there is little or no activity other than the response of the filterbank to the click itself. Since the impulse response of the filterbank is known we can use a simple distance measure to find the probability that the output of the filterbank at a certain time was caused by a glottal cycle. Better strobes can be found by finding maxima of the likelihood function, coupled with a threshold which encodes prior information about the expected timing.

sf2006

The sf2006 algorithm mixes the approach of a distance measure from some idealized response of the system, with the thresholding used in previous versions of AIM.

This method produces a summary statistic for the entire BMM, linking together the responses of different channels and ensuring that strobe points only occur in sets, marking the presence of a glottal pulse which must be present in all channels, even if it is masked by other energy in some of those channels. The impulse response of the filterbank is calculated for each channel. Then the probability that the last section of the BMM was caused by a glottal pulse, is calculated.

The summary probability is summed across channels to produce a single probability signal for the whole input, which represents the probability that the incoming signal was produced by a glottal pulse at that point. The probability is then passed to the thresholding algorithm used in earlier strobe finding algorithms. The strobe points are placed, based upon the known peaks in the impulse response of the filterbank for different channels. In this way it is possible to achieve continuous 'strands' of strobe points across all channels.



Results of the three strobe finding algorithms (from top to bottom): sf1992, sf2003 and sf2006 for a male 'aa' vowel, filtered with the dynamic compressive gammachirp filterbank.

The multi-source case

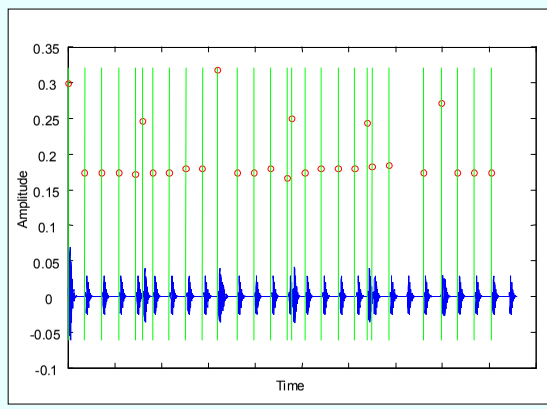
This approach uses an algorithm developed by Kaaresen (1996). The algorithm tries to fit an optimal set of spike timings and spike amplitudes to a noisy signal that was generated from a train of spikes of random amplitudes and timings which was subsequently convolved with some known impulse response function. In the form presented in the paper, it is assumed that spike amplitudes are Gaussian distributed and with timings from a Bernoulli-Gaussian process.

The version of the algorithm as presented in the paper as it provides a useful basis upon which to build in further prior knowledge about the processes which we are likely to encounter. There are some problems: it is also not the case that in all channels of the BMM, the glottal pulses will appear as filtered spikes. In those channels where there is little formant information it will be a good approximation, but in general we will need to extend the approach of this paper somewhat. Despite its limitations, this algorithm should be a good candidate for strobe finding in a more general case because:

1. The algorithm can find a set of amplitudes for the clicks it finds, as well as timings. This means that it should ultimately be able to differentiate between strobe processes generated by multiple speakers on the basis of timing information and pulse magnitude.
2. It uses a more sophisticated search algorithm than the very simple but naïve one which I implemented in sf2006, so it should be faster.

Algorithm

The algorithm was implemented in MATLAB to analyse the BMM produced by AIM. The algorithm relies upon having a known impulse response wavelet from which to make its calculations, thus I could use the same function as is used in the BMM module function to create the wavelets. From a physiological point of view this means that the brain would have to hold some sort of 'template' of the impulse response of the basilar membrane at a certain point.



The results of the iterated window maximisation algorithm on a double clicktrain filtered with the gammachirp filterbank.

The signal was two click trains, one at 20Hz and one at 88Hz at a relative amplitude of 70%.

The signal is in blue, the green lines represent the points that the algorithm has detected as being the start of impulses and the red circles represent the inferred amplitudes.

Preliminary results

I passed the algorithm a filtered double click-train (two concurrent click trains of different amplitudes and frequencies) using the gammatone module of AIM. I found that extremely accurate results could be achieved, but that they appear to be strongly dependent upon the noise variance estimates and other input parameters.

References

Stefan Bleeck, Tim Ives, and Roy D. Patterson, (2004) "Aim-mat: The auditory image model in matlab". *Acta Acustica*, **90** 781–787.
Toshio Irino and Roy D. Patterson (1997) "A time-domain, level-dependent auditory filter: The gammachirp". *Journal of the Acoustical Society of America*, **101**(1):412–419.
Toshio Irino and Roy D. Patterson, (2006) "A dynamic compressive gammachirp auditory filterbank". *IEEE Transactions on Audio, Speech, and Language Processing*, in press.
Kjetil F. Kaaresen, (1997) "Deconvolution of sparse spike trains by iterated window maximization". *IEEE Transactions on Signal Processing*, **45**(5):1173–1183.
Kjetil F. Kaaresen, (1998) "Evaluation and applications of the iterated window maximization method for sparse deconvolution". *IEEE Transactions on Signal Processing*, **46**(3):609–624.
Roy D. Patterson, Mike H. Allerhand, and Christian Giguere, (1995) "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform". *Journal of the Acoustical Society of America*, **98**(4):1890–1894.

Acknowledgments

Research supported by the UK Medical Research Council (G9901257, G9900369)

Conclusions

My initial research into strobing algorithms has opened up many new questions. A major goal of the project will be to reconcile the approaches developed here with previous methods, such as those developed by Bleeck *et al.* which were informed by data from electrophysiological experiments. Intuitively it seems highly likely that the auditory system takes an approach to analysis which is optimal for the class of signals which it encounters, and as such, maximum likelihood methods informed by well-reasoned priors are likely to be highly effective as a method of analysis.