MODULATION POISSON RATE ESTIMATION FOR DOUBLY STOCHASTIC AUDITORY PROCESSES

Owen P. Kenny, David B. Grayden & Anthony N. Burkitt

The Bionic Ear Institute, Melbourne, Australia.

ABSTRACT: This paper describes a method of rate estimation for spike trains produced by auditory models. The firing of the spike train is modelled as a doubly stochastic process where the firing rate is a function of an underlying process. The underlying process is estimated from an observation counting process. The variance of the estimator is calculated for tracking a random walk stochastic FM signal to give an empirical performance measure. Formant tracking is also demonstrated as an application for the use of this type of estimation.

INTRODUCTION

Researchers have long given consideration to the relationship between speech production and the response of the auditory system. Models were derived from auditory experiments to describe and simulate the behaviour and characteristics of the auditory system (e.g., Ghitza, 1992; Wang & Shamma, 1994; Zhang *et al*, 2001). These models have demonstrated the inherent processing gains naturally found in biological auditory systems. Auditory models have found application in robust automatic speech recognition and speech coding (e.g., Ghitza, 1994; Kim *et al*, 1999).

Many of the auditory models found in the literature utilise some form of linear or non-linear filter-bank followed by a non-linear process to represent generation of action potentials in the auditory nerve. The auditory model output consists of a set of spike trains that form a spatiotemporal representation of the speech. Information about the input acoustic signal is contained in the spike amplitudes and spike rate and in the time interval between the successive spikes. In speech-processing tasks, it is the inter-spike intervals that are of particular interest, since the inverse of these give instantaneous frequencies that provide information about the dominant frequencies (formants) of the signal. When noise is present, the effect is that the time interval between spikes varies, which affects the instantaneous frequencies.

The aim of this paper is to investigate a marked Poisson process model of the spike train produced by an auditory model. The Poisson point process relates to the time interval between the successive spikes. The firing rate of the spikes is a function of an underlying modulating stochastic process. The observation time interval also includes an additive process, which accounts for the influence of the noise. The objective of the model is to obtain an estimate of the underlying modulating stochastic process from successive time interval observations, or counting process observations.

The outline of the paper is as follows. A brief overview of the auditory model is given, with some detail of the structure of the auditory filters and spike generation used to represent the auditory nerve firing. Modelling the spike train output as a doubly stochastic process and the estimation of the underlying modulation process is then discussed. The variance when tracking a random walk stochastic FM signal in noise is used to quantify the performance of the rate estimator. Finally, an application of the rate estimator to the problem of formant tracking is demonstrated.

AUDITORY MODEL

Several authors have proposed auditory models to represent biological auditory systems. The auditory model that was implemented was similar to that designed by Ghitza (1992) and Kim *et al.* (1999). The auditory model was motivated by physiological behaviour of the mammalian hearing periphery. However, it has been simplified and freed from many of the parameters typically found in auditory models.

In the mammalian periphery, sound is transmitted from the outer ear to the base of the cochlea. Here the sound is converted into travelling waves that propagate along the basilar membrane from the base to the apex of the cochlea. The maximum excursion of the basilar membrane occurs near the base for high frequencies and near the apex for low frequencies. The cilia of the inner hair cells bend with the movement of the basilar membrane causing an action potential to be generated in auditory nerve fibres.

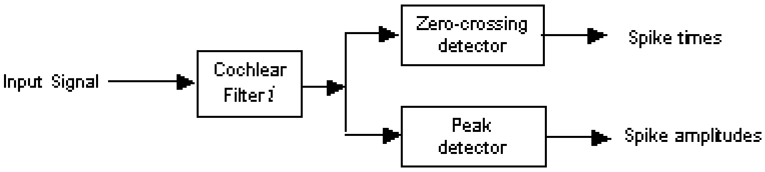

Figure 1. Auditory model.

In the auditory model (see Figure 1), a set of linear band pass filters was used to simulate the mechanical displacement of the basilar membrane. This filter bank represented frequency selectivity at various points along the basilar membrane. Thirty filters were created with centre frequencies distributed linearly from 245 Hz to 3820 Hz and bandwidths of 125 Hz. The auditory model simulated neural firing by producing a spike at the positive-going zero crossing for each filter output. The amplitude of each spike was the peak amplitude before the next negative-going zero-crossing transition (see Figure 2).

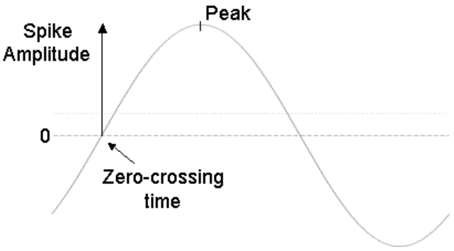

Figure 2. Spike generation from cochlear filter i. The amplitude of a spike is the following peak of the signal. The spike time is the zero-crossing time.

The auditory model exhibits two important properties of the auditory system. The first is that the average firing rates of the spikes convey spectral content of the stimulus. This is due to the fact that the inner hair cells are positioned tonotopically along the basilar membrane and, consequently, the auditory nerve fibres can be characterised by their best frequency. The second property is the presence of synchronised phase locking in the spike sequences. This synchronous firing pattern contains useful frequency information, as it usually corresponds to formant information of speech signals.

The spike train generated by the auditory model can be modelled as a marked Poisson process. This process contains the synchronous firing patterns and the formant frequency information. The means by which to extract this information is described in the next section.

MATHEMATICAL MODELS

The output of an auditory model filter can be modelled as an amplitude and frequency modulated waveform. This waveform is the result of a summation of all the components in the filter bandwidth. The first moment of this signal with respect to frequency corresponds to its instantaneous frequency and the second moment to its instantaneous bandwidth. When a speech formant passes through a filter its energy is distributed around a trajectory corresponding to its instantaneous frequency. Due to the fact that the auditory filters have large overlapping bandwidths, the outputs of the filters are locked to the nearest/strongest formant for speech signals.

An analysis of the effect of the noise on the zero crossing of the output of the auditory filter is given in Kim *et al.* (1999). A description of the influence of the noise is given here for completeness. The analytic form for the output of an auditory filter is given by

$$z(t) = a(t)e^{j\theta(t)} + \beta(t)e^{j\phi(t)}, \tag{2}$$

where $a(t)$ denotes the amplitude modulation component and $\theta(t)$ denotes the time varying phase related to the frequency modulation component of the speech signal. Similarly for the additive noise component, $\beta(t)$ and $\phi(t)$ denote the amplitude and frequency modulation processes, respectively. The zero crossing point is influenced both by the amplitude and phase of the noise. The noise component can be considered to have a complex Gaussian distribution if the amplitude has a Rayleigh distribution and the phase has a uniform distribution.

The process produced by the auditory model is described as a spike train whose Poisson intensity is determined by an underlying modulation process. The aim is to estimate this modulation process by evaluating the conditional expectation of the rate given spike arrival time observations. An equivalent estimate for the rate is obtained by evaluating the conditional expectation given counting observations. These counting observations are a sequence giving the number of spike arrivals in a set of intervals. The conditional estimate for such an observation sequence is given as

$$\hat{\lambda}_t = E\left[\lambda_t \middle| N_{t-1}, N_{t-2}, N_{t-3,} \cdots N_{t-T}\right]. \tag{3}$$

The rate estimator, similar to that of a Kalman filter, is implemented recursively utilising an innovation sequence. Innovations theory applied to point process estimation is well established and a detailed description of the theory can be found in Bremaud (1981). To summarise, the estimator consists of:

- a state representation of the process and a projection of this representation with respect to the observed history to give a prediction step,
- a representation of martingales with resect to the observation history and the innovations part of the process, and
- Martingale calculus as a means of identifying the innovations gain.

To construct an estimator the dynamics describing the doubly stochastic process is described by

$$d\lambda_t = A_r \lambda_{t|t-1} dt + dM_t \tag{4}$$

$$dN_t = \lambda_t dt + d\vartheta_t. \tag{5}$$

The Poisson rate intensity (4) produced by the underlying modulation process is modelled as an autoregressive process with coefficients, A_t, driven by a martingale, M_t. The counting observation process (5) is modelled as being related to the rate intensity of the underlying process and an observation noise process, ϑ_t. Estimating the martingale from the innovations process and integrating the differential rate equation with respect to time yields Watanbe's equation, defined as

$$\hat{\lambda}_t = \hat{\lambda}_0 + \int_0^t A_s \hat{\lambda}_s ds + \int_0^t K_s \left(dN_s - \hat{\lambda}_s ds\right). \tag{6}$$

The resulting estimation process has a recursive structure, which consists of a prediction step that estimates the rate from the previous rate estimates and a projection of the innovations process.

EMPIRICAL MEASURE OF PERFORMANCE OF THE POISSON RATE ESTIMATOR

To determine an empirical measure of performance of the estimator, an instantaneous frequency was constructed using an AR 1 random walk that was centred at 200 Hz with deviation frequency of 50 Hz. A stochastic frequency modulated signal was constructed. This signal was then passed through a single 500 Hz low-pass filter and the zero crossing times were obtained. The rate was then estimated using the procedure described above. The mean square error between the known instantaneous frequency and its estimate was determined and the instantaneous phase variance determined. This procedure was repeated for different SNR values and the results are plotted in Figure 3. It can be seen from the figure that for values of SNR greater than 5 dB the mean squared error is constant. Below 5 dB the influence of the noise becomes apparent.

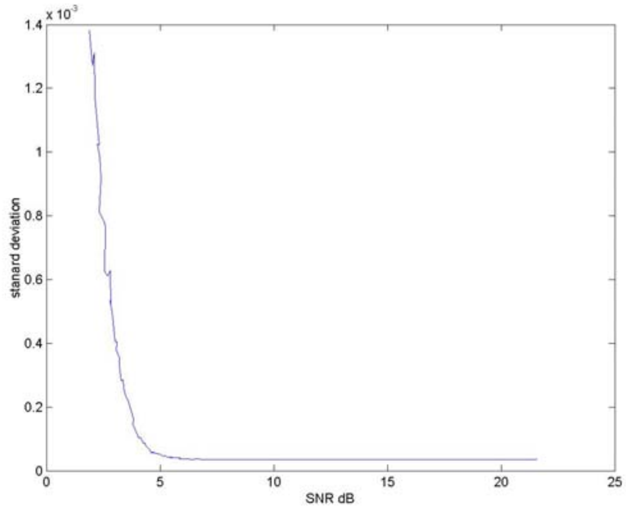

Figure 3. Instantaneous phase variance versus SNR for an AR 1 random walk.

APPLICATION OF THE ESTIMATION PROCEDURE TO FORMANT TRACKING

The benefit provided by the Poisson Rate Estimation (PRE) model is demonstrated in a formant estimation example. Conventional formant estimation routines use a windowed LPC analysis on the speech waveform and solve the predictor polynomial to locate the spectral peaks. However, when noise is present in the waveform, the spectral peaks are shifted in the LPC analysis and errors are made. The advantage of using the PRE procedure is that it makes use of the knowledge of the dynamics of the process that generate the spike rate and the statistics of the additive noise. As an example, Figure 4 shows an estimate of the Poisson rate of the spike train from the output of an auditory filter. The continuous plot shows instantaneous frequencies obtained from the inverse of the inter-spike intervals. The impulses in the figure denote estimates corresponding to 10 msec intervals. As it can be seen the effect of the noise is reduced.

In the formant-tracking task, the first formant (F1) was estimated by examination of the first five filters in the auditory model, which represented 245 – 745 Hz. The second formant (F2) was estimated using the next 12 filters, which were spaced from 870 – 2250 Hz. The first step for each formant was to determine a rough track by choosing the filter that had the most energy for each 5 msec interval. The output of each winning filter was windowed using a 10 msec Hamming window and then the PRE procedure was applied to estimate the formant. The estimator was implemented as a counting process where the number of spike arrivals within an interval was used as the observation process. The underlying process was modelled as a 100-tap AR process with a 16 kHz sampling rate. A formant estimate was determined for every sample point. In order to compare the PRE procedure with LPC, the output of the estimator was given every 10ms. A value of 0.8 was chosen for the innovations gain to minimise the mean square error for a range of SNR values from 10-30 dB.

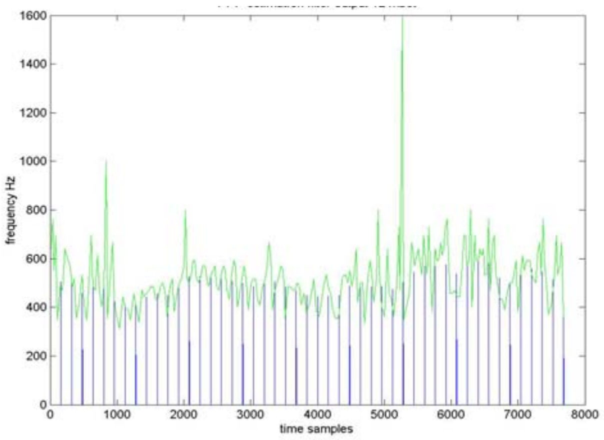

Figure 4. The instantaneous frequencies obtained from the inter-spike intervals and the Poisson rate estimates obtained from the output of an auditory filter.

The TIMIT database (Garofolo *et al.*, 1993) was used to evaluate the performance of the formant estimation routine in quiet and in noise. It was compared to conventional formant estimation using the FMANAL routine provided as part of the Speech Filing System (SFS) suite of speech analysis routines (Huckvale, 2000). The default parameters were used: 20 ms analysis window, 10 msec window step size, 12 poles.

Vowel and semi-vowel phonemes and clusters that exceeded duration of 100 msec were extracted from the TIMIT database and the first and second formants (F1 & F2) were estimated using both PRE and FMANAL. Gaussian white noise was then added to give a 10 dB signal-to-noise ratio and both routines were executed again. Comparison of the formant extraction results between the clean and noisy speech segments were made using RMS error as a measure of difference. 573 phoneme clusters totalling 7517 frames were extracted from 100 sentences spoken by male speakers taken from the TIMIT training set. The first 20 msec and the last 20 msec of the phoneme clusters were ignored to avoid coarticulation and TIMIT mislabelling errors that would cause spurious results, especially for the FMANAL routine. The results are summarised in Table 1.

Table 1. Comparison of LPC and Poisson Rate Estimation (PRE) formant tracking for vowel and semi-vowel phoneme clusters between speech in quiet and in Gaussian white noise at 10 dB SNR. The measure is RMS estimation error caused by noise measured with respect to estimates for speech in quiet.

RMS error (Hz)	**LPC**	**PRE**
F1 estimation error	195 Hz	20.8 Hz
F2 estimation error	364 Hz	117 Hz

The estimation of F1 is clearly superior in noise with the Poisson Rate Estimation procedure than the LPC procedure. Estimation of F2 is also superior for PRE than LPC in noise, although the error reduction is less.

CONCLUSION

In this paper a marked Poisson process model of the spike train produced by the auditory model was described and illustrated. In a random walk stochastic FM signal in noise problem it was demonstrated that the Poisson Rate Estimation procedure shows robustness to noise down to 5 dB SNR. The combination of the auditory model and the Poisson rate estimation process is capable of providing a much greater tolerance of background noise compared to LPC in a formant estimation task. The PRE procedure will continue to be developed, especially for higher-frequency rate estimation.

ACKNOWLEDGEMENTS

The authors would like to thank Dr W.P. Malcolm from the Weapon Systems Division, Defence-Science and Technology for his fruitful discussions relating to point process estimation.

This work was supported by The Bionic Ear Institute through the Human Communication Research Centre.

REFERENCES

Bremaud, P. (1981). *Point Processes and Queues, Martingale Dynamics,* New York: Springer-Verlag.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S. & Dahlgren, N.L. (1993). *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM*, NIST Speech Disc 1-1.1.

Ghitza, O. (1992). *Auditory nerve representations as a basis for speech processing*, in S. Furui & M.M. Sondhi (eds), Advances in Speech Signal Processing, 453-485, New York: Marcel Dekker.

Ghitza, O. (1994). *Auditory models and human performance in tasks related to speech coding and speech recognition*, IEEE Transactions on Speech and Audio Processing 2, 115-132.

Huckvale, M (2000). *Speech Filing System. Tools for Research.* University College London, http://www.phon.ucl.ac.uk/resource/sfs/

Kim, D-S., Lee S-Y. & Kil, R.M. (1999). *Auditory Processing of Speech Signals for Robust Speech Recognition in Real-World Noisy Environments*, IEEE Transactions on Speech and Audio Processing 7, 55-69.

Wang, K. & Shamma, S. (1994). *Self-normalization and noise-robustness in early auditory representations*, IEEE Transactions on Speech and Audio Processing 2, 421-435.

Zhang, X. Heinz, M.G., Bruce, I.C. & Carney, L.H. (2001). *A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression,* Journal of the Acoustical Society of America 109, 648-670.