Using an Auditory Model and Leaky Autocorrelators to Tune In to Speech

T. Andringa
tjeerd@bcn.rug.nl
Department of biophysics
University of Groningen
Postbus 72
9700 AB Groningen
The Netherlands

Abstract

This paper introduces a method to esitimate the spectrum of voiced speech in noise, based on an estimate of the fundamental frequency. The method uses the output of an auditory model that imitates the mechanics of the basilar membrane. The output of the segments of the model is used as an input to a set of leaky autocorrelator units (as simple neuron models) sensitive to a certain periodicity (delay). If a noisy vowel is presented to the system, the units sensitive to the fundamental period of that vowel respond most actively. The activity of the responding autocorrelator units as a function of segment number is a direct measure of the spectrum of the vowel. This technique is very robust and can, like humans, estimate the existence of a vowel in a SNR of -10 dB aperiodic speech-noise and formant frequencies in -3 to -6 dB. With this technique it is possible to split a mixture of sound sources in auditory entities (percepts) on the basis of pitch.

1 Introduction

The standard paradigm of speech preprocessing for automatic speech recognition leans heavily on the assumption that speech can be described as a stream of independent time-frames of approximately 25 ms. The success of speech recognition, at least of clean speech, proves that such frames contain enough information to give a reliable estimate of the spectral envelope. At low noise-levels the enhancement of the relevant characteristics of speech, using techniques like (RASTA-)PLP, can lead to some robustness, but at a low signal-to-noise ratio (SNR) some sort of selection of the relevant part of the incoming signal is necessary [1]. This can be done by using for example binaural sound source separation [2] and/or by using pitch information [3] to select voiced speech from a mixture of sources. This paper is aimed at the use of pitch information to select the voiced part of the speech of a target speaker.

In this paper a mechanical model of the basilar membrane (BM) movements [4] is introduced as a speech processing tool. This model is, like the real BM, continuous in both time and place. Attention is focused on processing the output of the model in the temporal domain using leaky autocorrelators that select the part of the total signal that has the desired pitch.

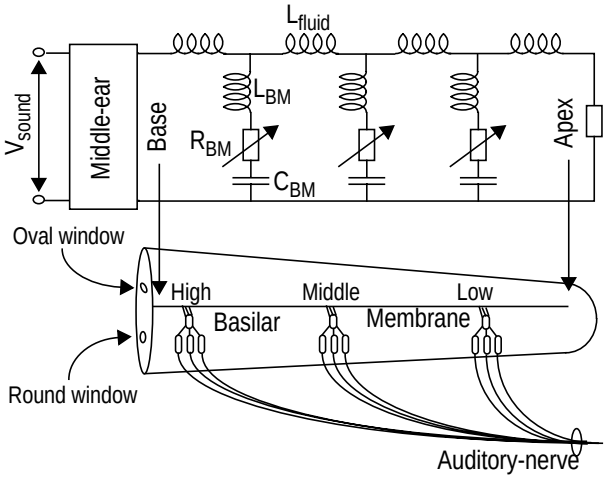

Figure 1: The basilar membrane (BM) is modelled as a cascade of 400 second order filters that model the cochlear fluid mass (M_{fluid}) and BM-segment mass (M_{BM}), the friction (damping) due to the basilar membrane movements (R_{BM}) and the stiffness of the BM (C_{BM})

2 The Auditory Model

The model of the basilar membrane consists of a series of 400 coupled second order filters, called segments. The segments are defined by their mass, a local stiffness (which decreases along the BM) and a damping that determines the energy dissipated. Figure 1 shows an electrical equivalent of the mechanical model and a schematic picture of the real cochlea. By convention the segments are numbered from base to apex and represent a decreasing resonance frequency (ω_0^2=1/LC) that is chosen according to an adapted Greenwood place-frequency relation [e], given by:

$$f_{res}[\text{kHz}] = 24[\text{kHz}] \cdot 10^{-(0.6 \cdot x[\text{mm}])} - 0.145[\text{kHz}]$$
$$x[\text{segm number}] = x[\text{mm}] \cdot \frac{400}{3.5[\text{mm}]} \quad (1)$$

This relation is roughly exponential from the base to the apex. Due to the characteristics of the coupled segments, the strongest response is slightly above the resonance frequency determined by the Greenwood relation.

Evenly spaced along the BM are approximately 3,500 *inner hair cells* (IHC) that transduce the BM motion into neuronal information which is transmitted by 30,000 neu-

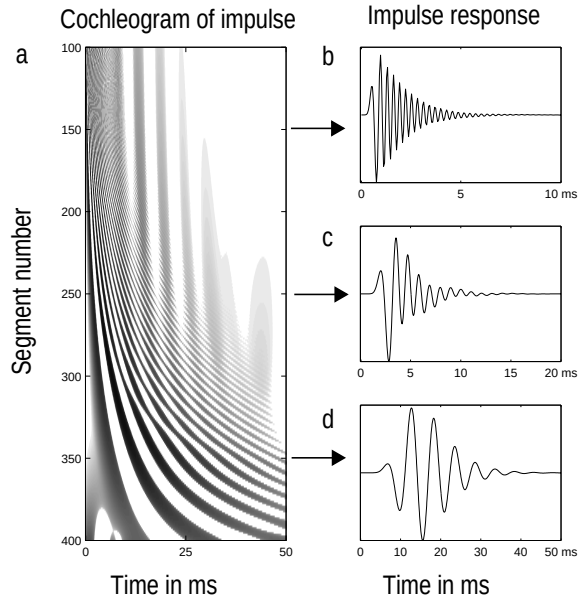

Figure 2: The impulse response of the BM model. Left the cochleogram of an impulse at t=0 ms. Right the impulse response of segments 150 (f_{res}=3770), 250 (f_{res}=1020) and 350 (f_{res}=204).

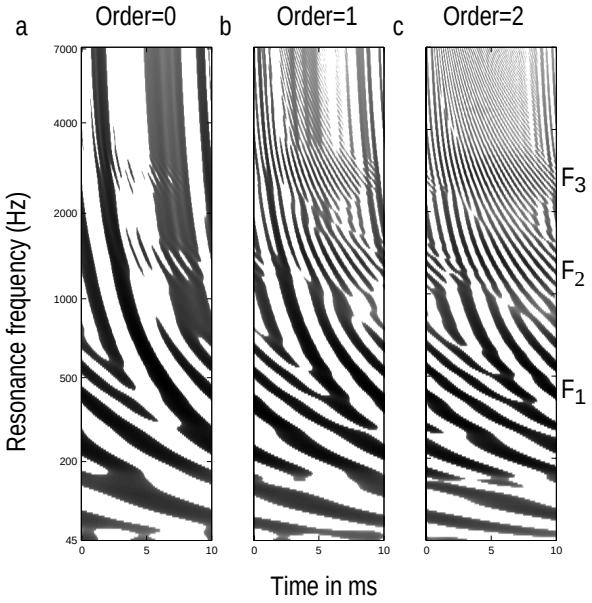

Figure 3: Effect of differentiation on the cochleogram. The first and second order derivative show much more prominent effects of formants. Note the way the harmonics are introduced.

rons of the auditory nerve to the brainstem. Each segment represents 10 IHC's and about 100 neurons of the auditory nerve. This paper assumes that all the information contained in the BM motion is available to the brain for further processing. The damping R_{BM} can be chosen nonlinear and even negative. In this paper the damping is linear and positive, which ensures that the BM-model is completely linear.

3 Cochleogram

The natural output of the model is of a series of time traces that represent the movement of each segment. An image representing this information is called a *cochleogram*. A cochleogram contains similar information as a spectrogram, but includes phase. As a special case the impulse response of the BM-model is displayed as a cochleogram for the segments 100 to 400 (corresponding frequency-range of 7000 down to 45 Hz) in figure 2a. In this picture all negative values are mapped onto zero (white) for enhanced visual contrast. The figures 2b-d give the impulse response for segments 150, 250 and 350. These segments have resonance frequencies of respectively 3770, 1020 and 204 Hz.

The impulse response differs from that of a gammatone filter mainly because there is a small delay that increases as a function of the distance from the base, and because the frequency of the oscillation decreases slightly with time. The envelope however is similar to that of a gammatone filter.

The auditory model integrates information through time, which results in a decreasing uncertainty about the periodicity of the driving signal. And because of spatial continuity it is possible to determine the spatial derivative of the BM-motion. Figure 3 shows several versions of the cochleogram of a synthetic vowel with a fundamental frequency of 100 Hz and three formants at respectively 520, 1190 and 2390 Hz (corresponding to the /AH/ in 'but'). Figure 3a shows the actual BM motion and figure 3b shows the first derivative of the BM motion. The tectorial membrane, with which the hairs of the IHC's make contact, contains fibres that are not perpendicular to the BM axis. The actual signal transduced can therefore contain some lateral difference information and resemble the first derivative of the BM motion. Compared to figure 3a the presence of the second and third formant is more prominent. Figure 3c gives the second derivative, and therefore the acceleration of every BM segment. This leads to an even clearer visualisation of the information represented by the BM.

Figure 3c shows the effect of changing resonance frequencies along the basilar membrane most clearly. At the low frequency side, corresponding to around 100 Hz, one period of the fundamental frequency fits in the corresponding fundamental period of 10 ms. At segments corresponding to 1200 Hz about twelve periods fit into 10 ms. The change of the number of harmonics represented by the BM segments entails a relatively small amplitude to ensure spatial continuity.

Figures 4a-c show the difference in information of the different orders of one period of segment 225. In figure 4a the actual form of the basilar membrane contains considerable low frequency information, while the second derivative (being an acceleration term) shows the force exercised to this segment of the BM.

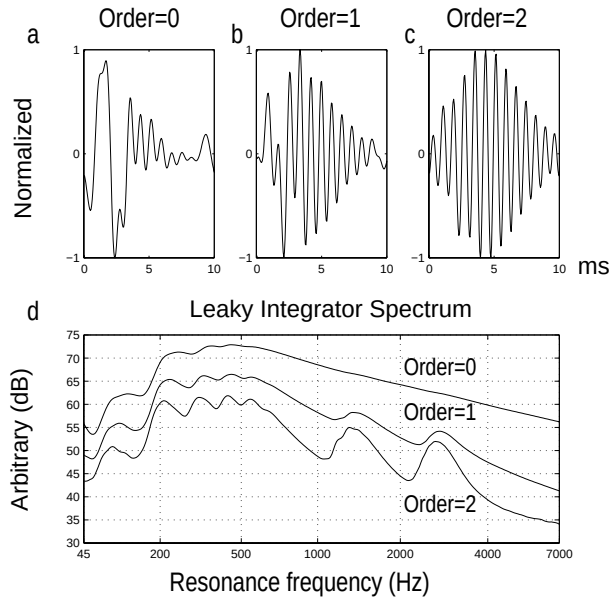

Figure 4: The effect of differentiation. Figure 4a-c all give one period of /AH/ of segment 225 for the actual form and its first and second order spatial derivative. Figure 4d gives the leaky integrator spectrum of /AH/. The values on the vertical axis are arbitrary in dB.

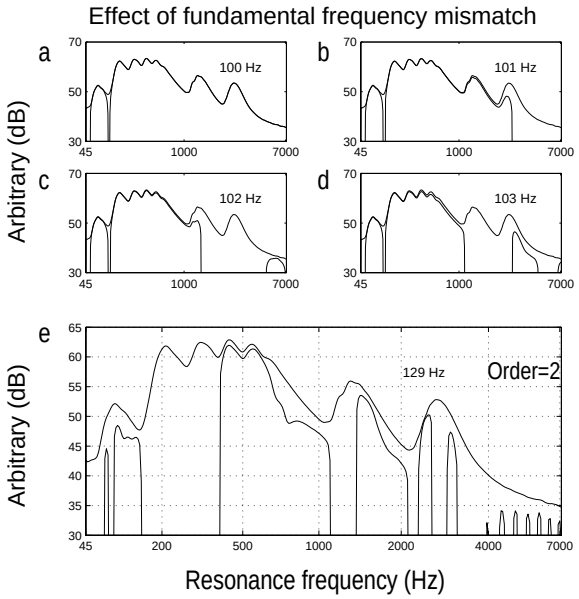

Figure 5: The effect of pitch mismatch. Figure 5a-d shows that the energy explained decreases rapidly with increasing mismatch. This enables an accurate estimate of the fundamental frequency. Figure 5e shows the effect of an arbitrarily chosen mismatch of 29%.

4 Leaky autocorrelator units

In order to estimate the activity of the BM segments some form of integration is necessary. Since the leaky integrator forms a suitable model for electrically small neurons, leaky integrator units are used to integrate temporal information of the segments. A discrete version of a leaky integrator (LI) is:

$$r_n(t) = r_n(t-\Delta t)e^{-\frac{\Delta t}{\tau}} + B_n(t)\Delta t \qquad (2)$$

When the average value of the integrator input of segment *n*, $B_n(t)$, is relatively constant compared to τ, $r(t)$ will reach a steady state in which $r(t) = r(t-\Delta t)$. If $\Delta t << \tau$ we can write using a Taylor expansion:

$$\begin{aligned} r_n(t) &= r_n(t)\left(1-\frac{\Delta t}{\tau}\right) + B_n\Delta t \\ r_n(t) &= B_n\tau \end{aligned} \qquad (3)$$

where B_n denotes the average value of the integrator input per Δt ms of segment *n*. In this paper two choices of $B_n(t)$ are used:

$$B_n(t) = y_n(t)y_n(t) \qquad \text{Energy} \qquad (4)$$

$$B_n(t) = y_n(t)y_n(t+T) \qquad \text{Periodicity} \qquad (5)$$

Here $y_n(t)$ denotes the displacement (or its first or second derivative) of segment n at time t. The leaky integrator (2) in conjunction with (4) integrates the total energy of a BM segment. In conjunction with (5) the LI integrates the correlation between the BM displacement T ms apart. If the BM motion is periodic with period T this will result in a value for $B_n(t)$ as (4). When the BM motion has a different period, usually a lower or negative value results.

5 Experimental results

Figure 4d shows the activity of the leaky integrator units with τ = 10 ms, 50 ms after the onset of the vowel /AH/ for different orders of differentiation. Once again the formant positions are more prominent for the first and second order derivative. Even the fundamental frequency and the first harmonics are clearly visible.

Figure 5 shows the effect of a mismatch of the periodicity detector (5). In figure 5a the period is chosen correctly and nearly all the energy in the signal is explained. In figure 5b an error of 1% has been made. Now the energy for frequencies above 2500 Hz cannot be explained. With increasing mismatch the fraction of energy explained gets smaller and more and more random. Figure 5e shows the effect of an arbitrarily chosen mismatch of 29%.

It is clear that the portion of the total energy explained is very sensitive to the correct choice of delay T. By comparing the integrated energy and the integrated autocorrelation with delay T an estimate of the fundamental frequency with an error smaller then 1% is possible within a few periods of the fundamental frequency.

The robustness to noise is very high, as is shown in figure 6. Figure 6 shows the same information as the leaky integrator spectrum of order 0 in figure 4e, but now masked by speech noise [6], which is the most difficult aperiodic noise since every part of the spectrum is equally masked by the noise. White noise, for example, has most of the noise energy situated above the third formant and is therefore masking low frequencies less effectively.

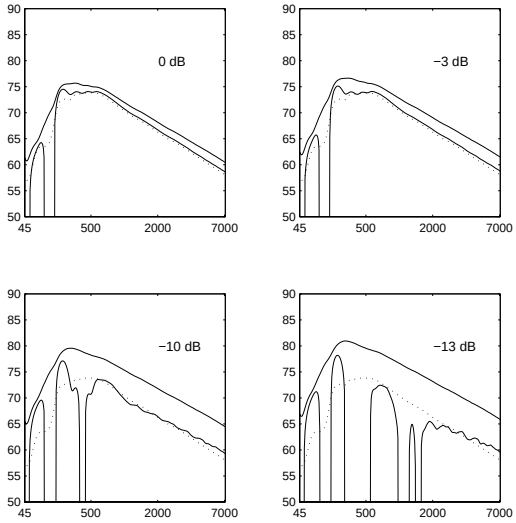

Figure 6: The effect of speech noise on the estimate of a vowel using the actual membrane movements. The uppermost line represents the total energy. The dotted line gives the energy of the vowel. The explained energy follows the reference spectrum up to -10 dB.

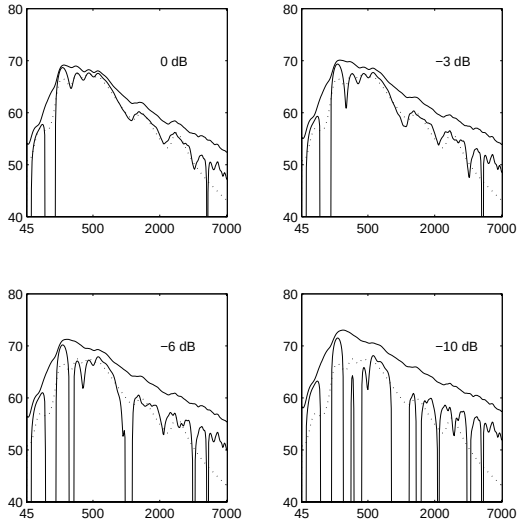

Figure 7: The effect of noise on the estimation on the formant positions. Here the first order spatial derivative is used. The formant frequencies can be reliably estimated in speech noise at a SNR of -3 to -6 dB. The dotted line gives the energy spectrum of the vowel.

The spectra of figure 6 are the average value of 80 ms of the leaky autocorrelator units sensitive to T=10 ms. With a SNR of -10 dB this technique is still able to reliably estimate the existence of the 100 Hz vowel. Only some energie between the first harmonics cannot be acounted for.

To estimate the position of the formant frequencies is somewhat more difficult and depicted in figure 7. Here the first derivative of the basilar membrane motion is used. The formant frequencies can be established well in a SNR of -3 dB, and somewhat less convincingly in -6 dB. These results can be improved considerably by choosing a higher value of the timeconstant τ and a longer period over which the output of the periodicity detector is averaged. The price to pay is a lower temporal resolution, that can become too low for a speech recognition application.

6 Conclusion

This research shows that the combination of a numerical model of the BM mechanics and a periodicity detector based on a leaky autocorrelator can reliably estimate the existence of a vowel with a certain period masked by aperiodic speech noise at SNR of -10 dB. Establishing the formant frequencies is possible at a SNR of -3 to -6 dB. By using a longer integration times then 80 ms this can be improved.

In the near future a HMM speech recognition system will be trained on clean speech preprocessed by the BM model. The proposed autocorrelation technique will be used to select relevant voiced parts in a noisy signal, resulting in a clean signal. The intervals between the voiced parts can be inspected for evidence of unvoiced sounds of the same user. The reconstruction based on both voiced and unvoiced parts will be the input of the HMM-system.

7 References

[1] M. Cooke and G. J. Brown, *Separating Simultaneous Sound Sources: Issues, Challenges and Models*, Fundamentals of Speech Synthesis, and Speech Recognition, editor E. Keller, John Wiley and Sons Ltd., Chichester, 1994.

[2] M. Bodden and T.R. Anderson, *A binaural selectivity model for speech recognition*, Proc. Eurospeech'95, pp 127-130, Madrid, 1995.

[3] W.M. Hartmann, *Pitch Perception and the Segregation and Integration of Auditory Entities*, Fundamentals of Speech Synthesis and Speech Recognition, editor E. Keller, John Wiley and Sons Ltd., Chichester, 1994.

[4] H. Duifhuis et al., *Modeling the cochlear partition with coupled Van der Pol oscilators*, Peripheral Auditory Mechanisms, J.B. Allen et al. editors, Springer, New York, 1985)

[5] D.D. Greenwood, *Critical Bandwidth and the Frequency Coordinates of the Basilar Membrane*, J. Accoust. Soc. Am, 33, pp. 1344-1356, 1961

[6] H.J.M. Steeneken and F.W.M Geurtsen, *RSG-10 noise data-base*, TNO Institute of perception, Soesterberg, 1990.