# THE TEMPORARY ENERGY DISTRIBUTION MODEL (TED) OF PITCH PERCEPTION

*Henning Reetz (henning.reetz@uni-konstanz.de)*
*Dept. of Linguistics, University of Konstanz, Germany*

## ABSTRACT

Pitch perception models assume that either the *place* along the basilar membrane or the firing *rate* of neurons encodes the perceived pitch. Place coding is widely accepted, because random sine phase components show no periodic peak pattern in the complex waveform [1], and two sine tones presented at different ears can cause a pitch percept [2], neither of which a peripheral peak-picking model can explain. Contrary to this argues this paper for a *rate* coding of the energy of the acoustic signal in the temporal domain as the source for pitch perception, resolving the two aforementioned problems. The model is implemented in a pitch extraction algorithm.

## THEORY

Sound waves reaching the outer ear are converted by the middle ear mechanics into travelling waves on the basilar membrane in the inner ear, eventually leading to neural firing of the haircells in the membrane. The *place* of maximal elongation of the membrane, and correspondingly, the maximal firing of neurons, is a frequency-dependent gradient along the membrane. This encodes different frequencies in different neuron locations along the membrane. In some way, the basilar membrane acts as a mechanical power-spectrum analyzer and the brain is supposedly able to derive the pitch of the signal from the spectral representation of it.

At the same time, individual cells fire in synchrony with the maxima of the acoustic signal, and the firing *rate*, or more precise, the distances between neural peaks encode the periodicity of the signal in the time domain. Furthermore, not only the neurons fire at the place of maximal elongation of the membrane, but all other neurons as well, although with reduced rate. And because temporal information is available in higher regions of the brain with a precision of a few microseconds, the brain could derive the pitch from the temporal representation of the signal.

In Goldstein's optimum processor theory [3], either the *place* or *rate* of the neural representation can be the input of the central pitch-processor, removing the ground for the spectral representation claimed by [2] as the only possibility to explain their experimental findings. Thus, the strongest argument for a *place* representation of the pitch in the auditory nerve is based on the experiments by [1]. He used complex signals differing only in the phase relations of their sine wave components. These signals have the same power-spectrum but differ considerably in their waveforms. The signals were perceived with the same pitch, hence he argues that only the phase insensitive spectral representation can explain the pitch perception, because a temporal representation of pitch would change with the differing waveforms. The argument is based on a waveform representation of the signal at the inner ear.

Unfortunately, the middle and inner ear are not linear systems and the basilar membrane performs a complicated three-dimensional movement with different travelling times along its length for different frequency components. The system uses active feedback components and cannot be described in linear terms. The movement of the hair-cells is a sine-wave movement for a sine-wave signal, but for complex signals, especially for time-varying signals like the speech waveform, this analogy breaks down. The ear has to be considered as an energy transformer and is not an amplitude encoder [4]. This essential point is missed by the argument that pitch perception is phase-insensitive and therefore cannot be explained in the time domain [1]. In fact, the loss of phase information in the spectral domain is not a consequence of the Fourier Transformation, but the outcome of computing the power-spectrum from it.



While the movement of the basilar membrane has not been mathematically modelled yet, we know that the membrane and its hair-cells convert signal energy into neural firing synchronized with signal maxima (for sine waves). More generally, the firing is in synchrony with maxima of the signal's energy, with firing rates of the neurons differing in intensity along the basilar membrane. Therefore, a rate coding of energy in the temporal domain is likely to exist in the auditory nerve.

## THE TED MODEL

In contrast to the 'neural firing in synchrony to signal amplitude' model, I propose a 'neural firing in synchrony to signal energy and frequency' model, where the distribution of energy in the temporal domain (Temporal Energy Distribution) encodes the pitch information. I describe now this model in the subsequent text in more detail. Numbers in the text refer to Figure 1.

The central idea of the model is the parallel representation of the acoustic signal in energy bands with different frequency responses. Taking a small window from a signal and computing its energy represents high-frequency energy, while wide windows represent low frequency energy (1). A range of windows with increasing sizes represents the energy in bands with decreasing frequencies (2). These energy bands can be understood as an instantaneous energy spectrum which is different from the classical power-spectrum. In the classical power-spectrum, the frequency distribution in a window is given under the assumption that the signal part in the window is a stationary signal. The classical power-spectrum is also usually used as data-reduction step, namely for locating harmonics in it. In opposition to this, the computation of the energy bands is made without an assumption of the type of signal, and it yields an increase in the data rate.

Next, the energy bands are converted into a representation about the maxima in it (3). In each energy band, the signal maximum leads to a 'firing' of a neuron according to the 'all-or-nothing' principe, i.e., information about the absolute value of the maximum is lost and only the information that the signal has reached a maximum at a certain time is encoded in the neural firing. The 'ignition' of the cell is linked to adaptation and refraction processes, preventing the neurons from firing at every local maximum, independent of their size and duration in relation to the neighboring signal.

This parallel concerto of firings is gathered into a temporal histogram of all neurons (4). This histogram is the energy distribution in all energy bands over time. The distances between the maxima in it reflect the periodicity in the signal.

## SOME CONSEQUENCES

The TED histogram can be interpreted in terms of speech production and perception. In speech production, energy is emitted either permanently (e.g., in voiceless fricatives) or impulsively, where the impulse can be a singularity (e.g., in a plosion burst) or impulses can occur repetitively (e.g., in a voiced sound). The TED histogram shows the impulsive energy emission, which can be a singularity, or a repetitive but irregular emission (e.g., in a creaky voice), or it can be a quasi-periodic sound. The difference between these three groups of sounds is reflected in the distribution of energy as being either singular, not periodic, or quasi-periodic. Especially the capability to identify any voiced sound, may it be periodic or not, gives the TED representation more power than most other pitch detection methods.

In perceptive terms, the TED model locates any energy distribution in the signal, independent of its origin. Furthermore, the TED representation has

some unusual feature for a temporal representation: random noise leads to a more or less random firing of all neurons, resulting in a nearly flat TED histogram. Periodic signals yield periodic firing in several energy bands, resulting in periodicity in the TED histogram. Spectral phase relations, number of involved harmonics, and random noise only decrease the 'peak-to-noise' ratio in the TED histogram, but does not hinder the pitch detection. As illustration, Figure 1 displays the behaviour of the model with a sine signal covered by random noise with an S/N ratio of -12 dB.

## ALGORITHM

The TED model was implemented in an algorithm whose general operation is described now with regard to Figure 1. (1) The speech signal is windowed with Hamming window sizes between 1 and 15 ms, converting the signal into parallel bands; the windows move sample-by-sample over the signal. (2) The windowed samples are squared and added up in the individual bands. (3) The local maxima within ±1 ms are selected and are represented as peaks with unitary height within each band. (4) The peaks of all bands are combined into a TED histogram. (5) Peaks with irregular distances to neighboring peaks and peaks with low amplitude are eliminated from the histogram. (6) The distances between peaks are represented as a pitch value if (i) they form a sequence of at least four peaks, and (ii) this sequence is longer than 30 ms.

The algorithm has a very simple structure but is slow on a digital general-purpose computer. Its *on-line* behaviour and its regular structure with simple computations and decisions makes it suited for realization in silicon where it could operate in real-time.

## CONCLUSION

An algorithm has been presented whose design is based on principles derived from the auditory processing in the inner ear. (These principles have been further tested with perception experiments presented elsewhere [5]). The speech signal is represented by a temporal structure in parallel energy bands which are computed in the temporal domain. This representation reflects speech production and perception issues equally well. In ongoing research I investigate the possibility to eliminate the periodicity test (step 5 of the algorithm) by incorporating more details of the intensity adaption of the inner ear into the model. Tentatively, the temporal energy distribution might also be suitable for the segmental representation of speech.

## REFERENCES

[1] Wightman, F.L. (1973) "Pitch and stimulus fine structure", *JASA, 54:* 397-406.

[2] Houtsma, A.J.M. and J.L. Goldstein (1972) "The central origin of the pitch of complex tones: evidence from musical interval recognition", *JASA, 51:* 520-529.

[3] Goldstein, J.L. (1973) "An optimum processor theory for the central formation of the pitch of complex tones", *JASA, 54:* 1496-1516.

[4] Duifhuis, H. (1992) "Cochlear modelling and physiology", In: *The auditory processing of speech - from sounds to words,* M.E.H. Schouten, (Ed.) Mouton: Berlin. p. 15-27.

[5] Reetz, H. (1995), *A temporal pitch perception model*, Doctoral diss. (unpublished)

[6] Houtsma, A.J.M., T.D. Rossing, and W.M. Wagenaars (1987) *Auditory demonstrations (CD)*, Institute for Perception Research (IPO) and Northern Illinois University (NIU), supported by the Acoustical Society of America: Eindhoven, The Netherlands.



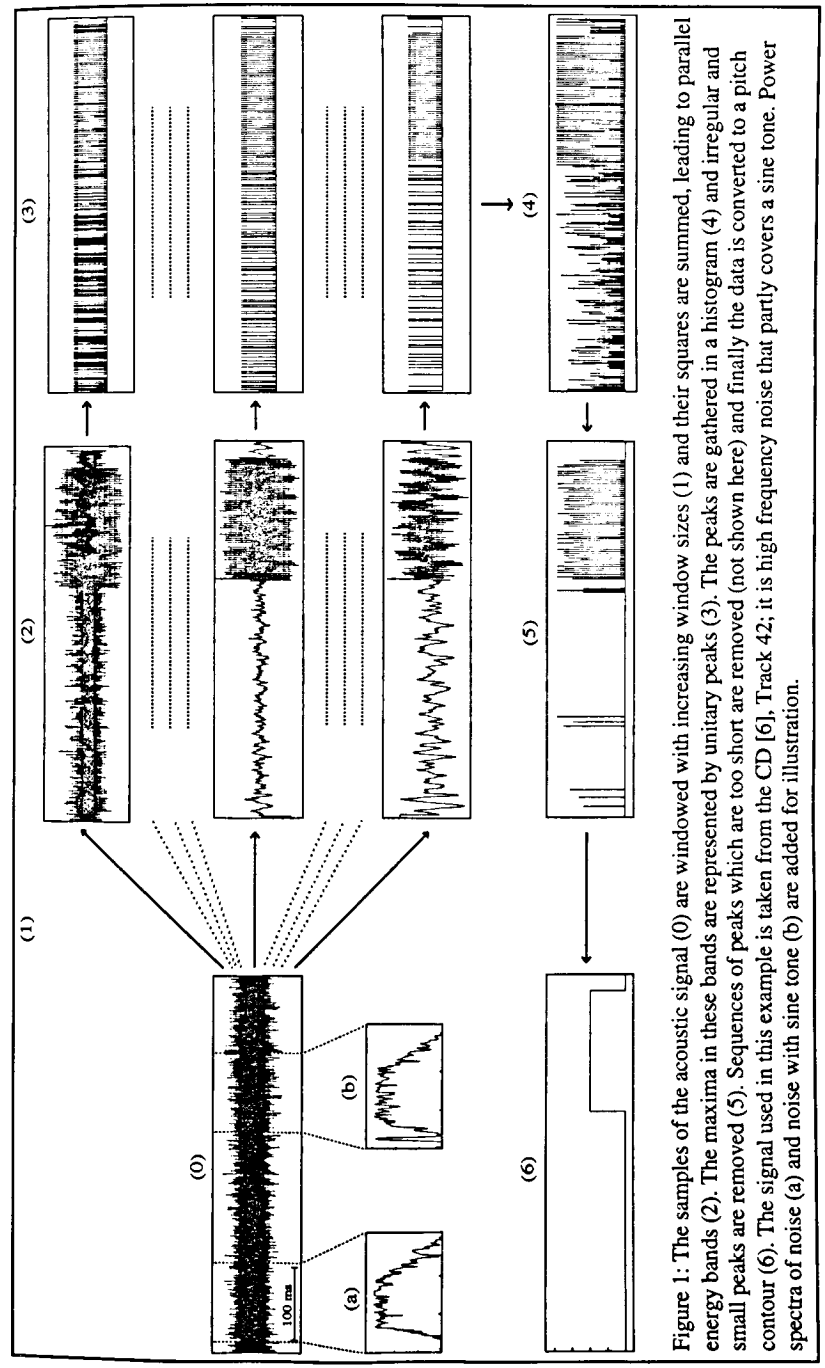


Figure 1: The samples of the acoustic signal (0) are windowed with increasing window sizes (1) and their squares are summed, leading to parallel energy bands (2). The maxima in these bands are represented by unitary peaks (3). The peaks are gathered in a histogram (4) and irregular and small peaks are removed (5). Sequences of peaks which are too short are removed (not shown here) and finally the data is converted to a pitch contour (6). The signal used in this example is taken from the CD [6], Track 42; it is high frequency noise that partly covers a sine tone. Power spectra of noise (a) and noise with sine tone (b) are added for illustration.