

# Non-stationary acoustic objects as atoms of voiced speech

Friedhelm R. Drepper<sup>1</sup>, Ralf Schlüter<sup>2</sup>

<sup>1</sup> *Zentralinstitut für Elektronik, Forschungszentrum Jülich, 52425 Jülich, Email: f.drepper@fz-juelich.de*  
<sup>2</sup> *Lehrstuhl für Informatik 6, Computer Science Department, RWTH Aachen University*

In spite of the undisputedly high degree of non-stationarity of speech signals, the present day determination of its acoustic features is based on the assumption that *speech production* can be described as a linear time invariant (LTI) system on the time scale of about 20 ms [1]. In automatic speech recognition, the wide sense stationarity of an LTI-system is used as prerequisite for the consistent estimation of Fourier spectra or of autoregressive models [1]. As an evolutionarily plausible supplement, *speech perception* is also assumed to be focussed on acoustic features which are obtained by using the LTI assumption [2]. Present day models of pitch perception are no exception [2, 3].

The empirical mode decomposition of Huang et al. [4] represents one of very few methods, which are suited to analyze signals without assuming their stationarity. In case of voiced speech, the formant frequencies have been shown to correlate well with frequencies of the empirical modes [5]. However, it is hard to imagine how the so called sifting process of Huang et al. [4] fits into the known function of the peripheral auditory pathway. The present study proposes a different method of empirical mode decomposition, which is based on well known features of the auditory pathway. As a further contrast to decomposition [4], the pitch perception oriented mode decomposition leads to mode reconstructions, which can be confirmed to have uncorrupted phases in comparison to the phases of underlying oscillatory subsystems [6, 7]. Furthermore, it is shown that voiced phones support a fast convergent iterative reconstruction.

The empirical modes can be used advantageously to reconstruct a single fundamental oscillator, the frequency of which can be interpreted as the acoustic correlate of virtual pitch. Virtual pitch perception should be interpreted as an ingenious instrument of time scale separation, which separates the phonological time or frequency scales from the phonetically relevant ones. In contrast to numerous existing pitch tracking methods [2, 3], the virtual pitch oriented time scale separation does not rely on a frequency gap in the long term spectrum (being equivalent to temporary stationarity).

Hearing models are well known to use a subband decomposition with so called critical (audiological) bandwidths. These bandwidths limit the number of separable harmonics to the range 6-8. If we describe the input signal as  $S(t)$ , the envelope of the impulse response of the bandpass filter as  $W_j(t')$ , the distance of the envelope maximum to the start of the response as  $\tau_j$  and the centre filter frequency (CFF) as  $\omega_j/2\pi$ , we obtain a complex subband of the form

$$X(\omega_j, t) = \int_{-\infty}^{t+\tau_j} d\tau S(\tau) W_j(t-\tau) e^{i\omega_j(t-\tau)}, \quad (1)$$

which represents an analytical signal (due to the limited bandwidth) and is thus well suited to obtain subband phases

$\varphi_j(t)$ . An autoregressive approximation to a gammatone bandpass filter represents a widely used example [6, 7]. The CFFs  $\{\omega_j | j=1, \dots, J\}$  are usually chosen with the aim to achieve approximate orthogonality (and completeness) of the decomposition. In deviation to common hearing models we start the analysis with CFFs of virtual pitch perception models [2,3], i.e. with CFFs which form a harmonic (equidistant) grid (template) for the harmonics 1-8. Deviating from common pitch models we introduce a time dependent CFF into *subband* (1)

$$X_j(t) = \int_{-\infty}^{t+\tau_j} d\tau S(\tau) W_j(t-\tau) \exp(i \int_{\tau}^t \omega_j(\tau') d\tau') \cdot \quad (2)$$

To demonstrate useful properties of *part-tone* (2), we choose an input signal which is often termed as *sinusoid*

$$S(\tau) = A(\tau) \cos \Phi_j(\tau) \quad \text{with}$$

$$\Phi_j(\tau) = \int_0^{\tau} \Omega_j(\tau') d\tau', \quad (3)$$

where amplitude  $A(\tau)$  and phase velocity  $\Omega_j(\tau)$  are both limited to positive values. In analogy to phase definition (3), we define filter phase  $\psi_j(\tau)$  by a time integral of the respective CFF  $\omega_j(\tau)/2\pi$ . For times  $t \gg \tau_j$ , analytic signal (2) simplifies to

$$X_j(t) = \frac{1}{2} e^{i\psi_j(t)} \int_0^{t+\tau_j} d\tau A(\tau) W_j(t-\tau) e^{i(\Phi_j(\tau) - \psi_j(\tau))}.$$

Choosing the special case  $\omega_j(\tau') = \Omega_j(\tau')$ , i.e. a CFF contour which is precisely adapted to the sinusoid, and assuming a slowly varying amplitude  $A(\tau)$ , we get

$$X_j(t) = \frac{1}{2} A(t) e^{i\Phi_j(t)}, \quad (5)$$

i.e. a part-tone which represents the analytic signal of the input signal. Since equation (5) is valid for arbitrary times, result (5) can also be interpreted in a different way. If (in case of a slowly varying amplitude) we succeed to adapt the CFF of a part-tone to its instantaneous frequency, we are assured to have reconstructed a part-tone with an uncorrupted phase.

For a given CFF contour all input signals with a different phase velocity experience a damping due to interference. A filterbank with bandpass filters of form (2) or (4) is thus well suited to separate several empirical modes with different frequencies without corrupting their phases. However, there remains the problem of finding the appropriate CFF contours. In this situation we hypothesize that voiced speech supports a robust and efficient adaptation algorithm on the receiver side and that the adaptation of the CFFs can be achieved iteratively by using the instantaneous frequency contours of the respective part-tones.

