Non-stationary self-consistent acoustic objects as atoms of voiced speech

Friedhelm R. Drepper

Forschungszentrum Jülich GmbH, 52425 Jülich, Germany, f.drepper@fz-juelich.de

Voiced segments of speech are assumed to be composed of non-stationary voiced acoustic objects which are generated as stationary (secondary) response of a non-stationary drive oscillator and which are analysed by introducing a self-consistent part-tone decomposition. The self-consistency implies that the part-tones (of voiced continuants) are suited to reconstruct a topologically equivalent image of the hidden drive (glottal master oscillator). As receiver side image the fundamental drive (FD) is suited to describe the broadband voiced excitation as entrained (synchronized) and/or modulated primary response and to serve as low frequency part of the basic time-scale separation of auditive perception, which separates phone or timbre specific processes from intonation and prosody. The self-consistent time-scale separation avoids the conventional assumption of stationary excitation and represents the basic decoding step of the phase-modulation transmission-protocol of self-consistent (voiced) acoustic objects. The present study is focussed on the adaptation of the contours of the centre frequency of the part-tone filters to the chirp of the glottal master oscillator.

1. INTRODUCTION

Many methods being conventionally used to analyze non-stationary (speech) signals like short time Fourier analysis or wavelet analysis [1, 2] are based on a complete and orthogonal decomposition of the signal into elementary components. The amplitudes of such components can be interpreted in terms of a time-frequency energy distribution. The elementary components are preferentially chosen as near optimal time-frequency atoms, which are each characterized by a reference time t_0 and angular frequency Ω_0. As a characteristic feature of wavelet analysis, the time-frequency atoms are chosen on different time scales. Time-frequency atoms are wave packets which are optimized to describe simultaneously event (particle) and wave type properties of non-stationary wave processes [3]. Their most general form can be written as second order logarithmic expansion of a complex signal around the reference time t_0 resulting in a complex Gaussian of the form

$$S_G(t) \approx \exp\left(-\frac{(t-t_0)^2}{2\sigma^2} + i\Omega_0(t-t_0)\left(1+c/2\,(t-t_0)\right)\right). \quad (1a)$$

Contrary to the conventional one [3], this parametric set of time-frequency atoms is characterized by a *quadratic* trend phase or a linear trend phase velocity (angular frequency)

$$\omega_{0,t} = \Omega_0\,(1+c\,(t-t_0)) \quad (1b)$$

with relative chirp rate c. Due to their neglect of the chirp parameter, the time-frequency atoms of short time Fourier analysis and wavelet analysis are preferentially aimed at linear time invariant (LTI) systems [1] (with a time periodic deterministic skeleton). In contrast to the latter approaches, the present one is aimed at non-stationary acoustic objects which represent a superposition of time-frequency atoms with chirped angular frequencies. The general aim, however, is not a complete and orthogonal decomposition of the speech signal, but a (potentially incomplete) decomposition into part-tones which can be interpreted as topologically equivalent images of plausible underlying acoustic modes [4-8]. The part-tones are generated by bandpass filters with impulse responses which represent optimal or near optimal time-frequency atoms. The preference of time-frequency atoms of the form (1a) results from the aim to generate part-tones with a maximal time resolution, which is compatible with a frequency resolution being necessary to isolate a sufficient number of topologically equivalent images of the underlying acoustic modes.

Like in auditory scene analysis, an *a priori* knowledge about the behaviour of the underlying acoustic modes can be used to remove a potential ambiguity of the unknown acoustic object parameters (in particular of the time course of the centre frequencies of the bandpass filters). In case of voiced speech it is "known" *a priori* that the common origin of the acoustic modes (the pulsed airflow through the glottis) and the nonlinearity of the aero-acoustic dynamics in the vocal tract lead to a characteristic phase locking of the acoustic modes [5-8].

In the situation of signal analysis the detection of a strict $(n:n')$ synchronization of the phases of *a priori* independent part-tone pairs (with non-overlapping spectral bands) represents a phenomenon, which has a low probability to happen by chance, in particular, when the higher harmonic order n has a large value. For such part-tone pairs it can therefore be assumed that there exists an uninterrupted causal link between those part-tones, including the only plausible case of two uninterrupted causal links to a common drive, which can be identified as a glottal master oscillator [4-8]. Since the $(n:n')$ phase-locking with $n \neq n'$ is generated by the *nonlinear* coupling of the acoustic modes to the glottal oscillator, a stable synchronization of *a priori* independent part-tone phases can be taken as a confirmation of topological equivalence between these part-tones and respective acoustic modes in the vocal tract of the transmitter.

Based on the a priori knowledge about the phase locking of the acoustic modes, the phase velocity contours of the part-tones can be assumed to be centered around harmonic mul-

tiples ($2\pi h$ with integer h) of the frequency contour of the glottal oscillator. A cluster analysis of harmonically normalized part-tone phase velocity contours can thus be used to identify a consistent set of part-tone phases, which is suited to reconstruct a unique phase velocity of the fundamental drive [6-8]. The present study is focussed on the construction of centre frequency contours of the bandpass filters which are consistent with the corresponding part-tone phase velocity contours. Self-consistently reconstructed part-tone phases are proven to be suited for a phase-modulation transmission protocol of voiced speech.

2. VOICE ADAPTED PART-TONES

In case of the characteristic isolated pulse type events of stop consonants, single time-frequency atoms are potentially suited to describe such events. For real time analysis of voiced continuants it is unavoidable to generate part-tones which result from *causal* bandpass filters. The present study uses an all pole approximation of complex Γ-tone bandpass filters with approximately gamma-distribution like amplitudes of the impulse response [10]. For sufficiently high autoregressive order, the Γ-function like amplitude distribution guarantees a near optimal time-frequency atom property of the impulse responses. (That is why an autoregressive order (Γ-order) $\Gamma = 5$ will be used in the example instead of the more common choice $\Gamma = 4$ [11, 10].)

The choice of (roughly) audiological bandwidths for the part-tone decomposition has the effect that we can distinguish a lower range of part-tone indices characterized by guaranteed single harmonic (resolved) part-tones and a range of potentially multiple harmonic (unresolved) part-tones. In the resolved part-tone range $1 \le j \le 6$ the harmonic order h_j is identical to the part-tone index j. To avoid a substantial over-completeness (and *a priori* correlation between neighbouring part-tones) in the unresolved range $6 < j \le N$, the set of harmonic part-tones is pruned according to the respective equivalent rectangular bandwidths (ERB). A typical set of part-tones may have the harmonic orders $\{h_j\} = \{1,2,...,6,8,10,12,15,...\}$. (Diplophonic voice types may lead to rational winding numbers $h_j = n_j / m$ with a common subharmonic period number $m > 1$ [5-8].) In particular for speech segments, which correspond to nasals or vowels it is typical that some of the part-tones in the (a priori) unresolved range are also dominated by a single harmonic acoustic mode. The under-completeness of the part-tones in the (a priori) resolved range has a welcome noise suppression effect.

The all pole approximation of the gammatone filters has the advantage of a fast autoregressive algorithmic implementation [10]. For theoretical reasons we prefer its description in terms of a matrix recursion with a lower triangular matrix L of dimension Γ which plays the role of the cascade depth of the cascaded first order autoregressive filter,

$$L X_t = \lambda \exp(i\omega_t) X_{t-1} + e_1 S_t \qquad (2a)$$

with input signal S_t being sampled at discrete times t, Γ-dimensional vectors $X_t = \{v_t, w_t, \cdots, z_t\}'$, $e_1 = \{1,0,\cdots,0\}'$, $X_0 = 0$ and matrix

$$L = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -1 & 1 & \ddots & & \vdots \\ 0 & -1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & -1 & 1 & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix} . \qquad (2b)$$

The scalar λ represents the damping factor of every first order autoregressive filter and is directly related to the ERB of the Γ-tone filter, $\lambda = \exp(-a_\Gamma ERB)$, the Γ-order dependent factor a_Γ being given e.g. in [10]. The complex phase factor $\exp(i\omega_t)$ defines the instantaneous centre filter frequency $F_t = \omega_t / 2\pi$ being simply related to the instantaneous angular velocity ω_t. The unusual feature is the time dependence of the angular velocity ω_t which will be specified later. The inverse of matrix L is the lower triangular matrix with ones on and below the diagonal. It can be used to obtain X_t as a power series of matrix L^{-1},

$$X_t = \sum_{t'=0}^{t} \prod_{k=t'+1}^{t} \exp(i\omega_k)\, \lambda^{t-t'} L^{-(t+1-t')} e_1 S_{t'} . \qquad (3)$$

The filter output is represented by the last component z_t of vector X_t. Therefore we are interested in the matrix element in the lower left corner of any power of matrix L^{-1}. For the $(n+1)^{th}$ power this element can easily be obtained by complete induction as the ratio of three factorials $(n+\Gamma-1)!/(\Gamma-1)!n!$. Taking into account the additional dependence on the part-tone index *j,* the output of the (non-normalized) bandpass filter of part-tone *j* is thus obtained as

$$z_{j,t} = \sum_{t'=0}^{t} \exp\Big(i \sum_{k=t'+1}^{t} \omega_{j,k}\Big)\, \lambda_j^{t-t'} \frac{(t-t'+\Gamma-1)!}{(\Gamma-1)!\,(t-t')!} S_{t'} . \qquad (4)$$

For $j = 1,...,N$ the set of (normalized) part-tones can be interpreted as a highly over-sampled time-frequency decomposition of the speech signal $S_{t'}$ where the over-sampling is restricted to the time axis. The non-normalized part-tones (4) can be used to generate the part-tone phases (carrier phases)

$$\varphi_{j,t} = \arctan(im(z_{j,t}) / re(z_{j,t})) \qquad (5)$$

as well as the (harmonically) normalized part-tone phases $\varphi_{j,t} / h_j$ in the frequency range of the pitch. If the trajectory (contour) of the centre filter frequency $\omega_{j,k} / 2\pi$ is chosen as identical to the one of the instantaneous frequency $\omega_k' / 2\pi$ of a constant amplitude input signal $S_{t'} = A \exp(i \sum_{k=0}^{t'} \omega_k')$, the application of bandpass filter (4) generates the output

$$z_{j,t} = A \exp(i \sum_{k=0}^{t} \omega_k') \sum_{t'=0}^{t} \lambda_j^{t-t'} \frac{(t-t'+\Gamma-1)!}{(\Gamma-1)!\,(t-t')!} . \qquad (6)$$

This filter output has the remarkable property that its instantaneous phase velocity is identical to the one of the input signal. For a given filter frequency contour, other input signals experience a damping due to interference of the phase factors. For a given input frequency contour, other filter frequency contours generate a phase distortion of the output. In the limit $t \to \infty$, the sum in equation (6) represents an asymptotic gain

factor $g_{j,\Gamma}$. Being exclusively dependent on the bandwidth parameter λ_j and the Γ order, the gain factors can be used to obtain the normalized part-tone amplitudes $a_{j,t} = |z_{j,t}| / g_{j,\Gamma}$.

For more general voiced input signals the determination of filter frequency contours, which are identical to frequency contours of some underlying acoustic modes, represents a non-trivial problem. Conventionally [14-17] the adjustment of the filter frequency contours of part-tones (or "sinusoidal components") is achieved by introducing a short-time stationary (zero-chirp) subband decomposition which is densely sampled with respect to frequency and by determining for each point in time local maxima of the amplitudes of the subbands with respect to frequency. In a second step the maximizing frequencies of consecutive points in time are tested, whether they are suited to form continuous frequency contours. Suitable maxima are joined to form weakly non-stationary contours and part-tones. It is well known that the non-stationarity of natural voiced speech leads to frequent death and birth events of such contours, even within voiced segments [14-16].

The present approach is aimed at *self-consistent* centre filter frequency contours which are chosen as identical (or as consistent) to the frequency contours of the respective part-tones (outputs). It is based on the assumption that sustained voiced signals are composed of one or several part-tones which can iteratively be disclosed and confirmed to be self-consistent, when starting from appropriate contours of the centre filter frequency. In a first step we restrict the self-consistency to a single part-tone. In this case the self-consistency is defined as the existence of a centre filter-frequency contour of the bandpass filter being used to generate the part-tone which can be obtained as stable invariant set of the iteration of two cascaded mappings, where the first mapping uses a filter-frequency contour (out of a basin of attraction of preliminary frequency contours) to generate a part-tone phase velocity contour and the second mapping relates this part-tone phase velocity contour to an update of the mentioned filter-frequency contour. Whereas the first mapping is given by part-tone filter (4) and phase definition (5), the second mapping is chosen according to the acoustic properties of the assumed underlying physical system.

The acoustic properties include the physical law *natura non facit saltus*. The resulting smoothing of the centre filter frequency is suited to improve the convergence properties of the adaptation. Being inspired by equation (1b) the smoothing step might simply be chosen as a linear approximation of the trend of the filter-frequencies within each analysis window. However, due to the time reversal asymmetry of the Γ-tone filters, a negative chirp rate leads to a singularity of the instantaneous period length of the impulse response at finite times. This singularity can be avoided, if the time dependence of centre filter frequency $\omega_{j,t}/2\pi$ of part-tone j is chosen separately depending on the sign of the (relative) chirp rate c_j. For negative chirp rate it is useful to assume alternatively a linear trend of the inverse of the respective centre filter frequency with a smooth transition at zero chirp rate

$$\omega_{j,k} = \begin{Bmatrix} \omega_{j,0}(1+c_j\,k) \\ \omega_{j,0}/(1-c_j\,k) \end{Bmatrix} \quad for \quad \begin{Bmatrix} c_j \geq 0 \\ c_j < 0 \end{Bmatrix}. \qquad (7)$$

As is well known, human pitch perception is not limited to the frequency range of the separable part-tones. In particular it is known that the modulation amplitudes (envelopes) of the higher frequency subbands play an important role in hear physiology and psychoacoustics [11, 12]. It is therefore plausible to extend the analysis of part-tone phases to the non-separable range, i.e. to phases, which can be derived from the envelopes of the part-tones. Being used preferentially for part-tones with unresolved harmonics, the envelope phases are determined e.g. as Hilbert phases of (appropriately scaled and smoothed) modulation amplitudes (envelopes) of part-tones.

To achieve a more uniform time evolution of the envelope phases and in agreement to well known results from hear-physiology and psycho-acoustics [11, 12], the normalized modulation amplitudes $a_{j,k} = |z_{j,k}| / g_{j,\Gamma}$ are submitted to a sublinear transformation (scaling) and smoothing prior to the determination of the Hilbert phases. It is common practice to choose a power law with an exponent in the range $\nu = 0.33$ [11, 12]. In contrast to the carrier phases (which do not need a correction due to their self-consistency as expressed in equation (6)) the envelope phases need a group delay correction of the respective part-tones. The part-tone index specific part of this correction has been derived from the maxima of the amplitude of the impulse responses of equation (4) [10]. The relative importance of the envelope phases is expected to increase, when the voice source changes from a modal (ideal) voice to a breathy one.

3. PART-TONES OF A SIMPLE PULSED EXCITATION

To demonstrate the generation of self-consistent part-tones of a non-stationary voiced acoustic object, a sequence of synthetic glottal pulses with a chirped frequency is chosen as input signal. For simplicity the pulses are chosen as constant amplitude saw teeth with a power spectrum, which is roughly similar to the one of the glottal excitation. The pulse shape is described by an impulse function [21] or wave shaper function [22] of the form

$$G(\psi_t) = \min\left(\mathrm{mod}(\psi_t, 2\pi), s(2\pi - \mathrm{mod}(\psi_t, 2\pi))\right), \qquad (8)$$

where ψ_t represents the phase of an artificial glottal master oscillator [5-8]. The parameter s (chosen to be 6) determines the ratio of the modulus of the downhill slope of the glottal pulses to the uphill one. The chirp of the glottal oscillator is described by a time dependent phase velocity $\omega(t) = \dot{\psi}(t)$ which is chosen in analogy to equation (7), however, with potentially different chirp rate c' and initial phase velocity ω_0'. (In the specific example, the glottal chirp parameter c' is chosen to generate a doubling of the frequency (or period length) after about 25 periods.) The fundamental phase ψ' is obtained by integrating the analogue of equation (7) with respect to time t (replacing index k)

$$\psi(t) \;=\; \begin{Bmatrix} \omega_0^{'}\,(t + c'\,t^2/2) \\ -\,\omega_0^{'}/c'\;\ln(1-c't) \end{Bmatrix} \quad for \quad \begin{Bmatrix} c' \geq 0 \\ c' < 0 \end{Bmatrix} . \qquad (9)$$

In the situation of signal analysis, appropriate contours of the centre filter frequency of the part-tone specific bandpass filters have to be obtained iteratively from the observed signal. As part of the time scale separation step of the second mapping of the last section, we assume that these contours can be described (within the current rectangular window of analysis) by a simple smooth function of time chosen as indicated in equation (7). The part-tone adaptation of the filter-frequency contour of the bandpass filter of part-tone *j* can thus be achieved by estimating the parameters of equation (7). To reduce the dependence of the estimate on the size and position of the window of analysis (and/or to avoid the adaptation of the window length to the instantaneous period length), time scale separation ansatz (7) is extended by a 2π periodic function $P_j(\varphi_{j,t}/h_j)$ of the respective normalized part-tone phase

$$\dot{\varphi}_{j,t}/h_j \;=\; \alpha_j\,t \;+\; P_j(\varphi_{j,t}/h_j) \qquad (10a)$$

$$h_j/\dot{\varphi}_{j,t} \;=\; -\alpha_j\,t \;+\; P_j(\varphi_{j,t}/h_j)\,. \qquad (10b)$$

The 2π periodic function $P_j(\varphi)$ accounts for the periodic oscillations of the phase velocity around the long term trend being generated by the characteristic auto phase-locking and is approximated by an appropriate finite order Fourier series. The Fourier coefficients as well as the trend parameter α_j are obtained by multiple linear regression.

Within a voiced segment of speech the adaptation of the parameters is performed sequentially for successive analysis windows. The initial value of the centre filter frequency of the current window is therefore typically given as result of the adaptation of the filter chirp of the preceding analysis window. Thus we treat the latter parameter as given ($\omega_{j,0} \;=\; h_j\,\omega_0^{'}$) and concentrate on the convergence properties of the chirp parameter of a filter-frequency contour. The adaptation of a single parameter can be represented graphically. To explain the approach to self-consistency we use a graph, which shows the trend parameter α_j of equations (10a or 10b) for several part-tone indices *j* as function of the common filter chirp rate *c*. To make figure 1 suited for the graphical analysis it gives the estimates of the relative trend $\alpha_j/(\omega_0'c')$ for the indices j = 2, 4, 6, 9 (corresponding to the sequence of the fixed points from bottom to top) as function of the relative filter chirp rate c/c'.

The iterative adaptation of the chirp parameter of each filter-frequency contour can be read off from figure 1 by an iteration of two geometric steps: Project horizontally from one of the described curves to the diagonal of the first quadrant (which indicates the line where the fixed points of the iteration are situated) and project vertically down (or up) to the curve again. As can be seen from figure 1, the chirp parameters of all four part-tones have a stable fixed point (equilibrium) within a well extended basin of attraction of the chirp parameter which exceeds the shown interval of the abscissa. The fixed points (corresponding to the more general invariant sets of the preceding section) indicate the final error of the filter chirp which depends not only on the part-tone index but also on the size of the analysis window (which was chosen to have a length of about five periods of the glottal process). Due to the simple least squares regression of equations (10a,b), the modulus of the trend α_j is systematically underestimated.

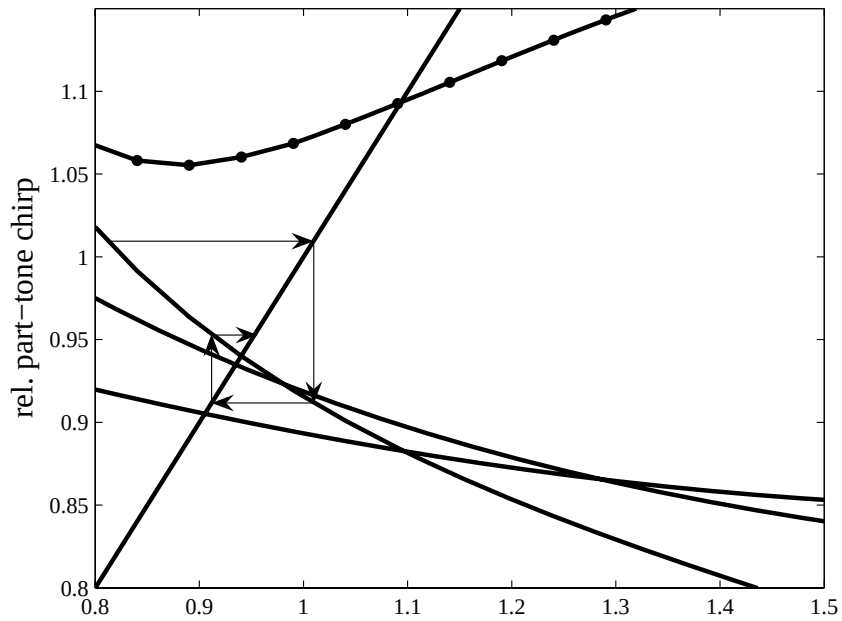

Figure 1: Estimated relative part-tone chirp rates as function of the relative chirp rate of the respective centre filter frequency, given for the envelope phase of part-tone 9 (circles, top) and the three carrier phases of part-tones 2, 4, and 6 (lines crossing the diagonal from bottom to top). All chirp rates are given relative to the chirp rate of the input sawtooth process defined in equations (8-9). The arrows and the diagonal of the first quadrant explain the algorithm, to determine the self-consistent centre filter frequencies.

4. MULTI PART-TONE STABLE ACOUSTIC OBJECTS

It is well known that human pitch perception can be trained to switch between analytic listening to a spectral pitch and synthetic listening to a virtual pitch [12, 16]. It is thus plausible to correlate the described single part-tone stable acoustic objects (with a macroscopic basin of attraction of the filter frequency contour or contour parameters) to outstanding part-tones, which are potentially perceived as spectral pitches by analytic listening [16, 27]. The number of stable invariant sets (fixed points) with a macroscopic basin of attraction depends in particular on the width of the power spectrum of the voiced signal. In the example of the last section a strong asymmetry of the sawteeth ($s >> 1$ in equation 10) favors the stability of higher order fixed points.

From psychoacoustic experiments it is also known that virtual pitch is a more universal and robust percept than spectral pitch [4, 16]. Based on the *a priori* assumption that the signal is generated by a voice production system, which generates several phase locked higher frequency acoustic modes, the observed (carrier or envelope) phase velocity of one part-tone might be used to adjust the centre filter frequency of other part-tones. This opens the possibility to use a more robust multi part-tone adaptation strategy which can be expected to converge even in cases with no single part-tone stability.

In analogy to the single part-tone stability of the last sections we relate multipart-tone stability of an acoustic object to

the existence of a fundamental phase velocity contour which can be obtained as stable invariant set of the iteration of three cascaded mappings, where the first mapping relates a preliminary fundamental phase velocity contour (out of a macroscopic basin of attraction) to a set of filter-frequency contours, the second mapping uses the set of filter-frequency contours to generate a corresponding set of part-tone phase velocity contours and the third mapping relates a subset of the part-tone phase velocity contours to update the fundamental phase velocity contour. The first mapping makes use of the characteristic auto-phase-locking of the voiced excitation ($\omega_{j,k} = h_j\, \omega_{0,k}$). The second mapping is given by filter (4) and phase (5) and the third mapping uses cluster analysis to identify invertible phase relations which are suited to reconstruct the phase velocity of the fundamental drive [4-8].

This way the contradiction between Rameau's concept of a *son fundamentale* or fundamental bass [20] and Seebeck's observation, that pitch perception does not rely on a fundamental acoustic mode as part of the heard signal [21], can be reconciled by replacing Rameau's *son fundamentale* by the described FD. Being an abstract order parameter and in need of a confirmation of its existence, the FD of a *multi-part-tone* voiced acoustic object cannot be reconstructed from a single part-tone alone. This qualifies the instantaneous fundamental phase velocity as acoustic correlate of *virtual* pitch perception. When reconstructed coherently for uninterrupted voiced speech segments, the fundamental phase becomes the central ingredient of a phase modulation decoder of voiced speech.

Contrary to the conventional psycho-acoustic theory [12, 16] (originating from Ohm and Helmholtz) which interprets the amplitudes of part-tones (with psycho-acoustically calibrated bandwidths) as primary acoustic cues, it is expected that the deviations of the phases of self-consistently determined part-tones from the synchronization manifold of the unperturbed ideally pulsed excitation have a comparable or higher relevance for acoustic perception than the corresponding amplitudes [17]. At the present state of analysis this hypothesis is mainly based on deductive arguments, which favor phase modulation features as more differentiated and robust cues for the distinction of the voiced phones of human speech as well as for the distinction of their speakers.

5. Conclusion

A transmission protocol of non-stationary self-consistent (voiced) acoustic objects is outlined, which are generated as stationary response of a non-stationary fundamental drive (FD) and which can self-consistently be decomposed into non-stationary part-tones. Self-consistent part-tones are characterized by phase velocities which are consistent with the centre filter frequencies being used to generate the part-tones. The second property of the self-consistent acoustic objects qualifies them as most elementary symbols of a voice transmission protocol which is centred on a time scale separation with a precise and robust decoding option. It is hypothesized that the self-consistent decomposition of speech segments, which are suited to transmit voiced continuants, leads to a subset of part-tones which shows generalized synchronization of their phases. The iterative identification of multi part-tone stable voiced acoustic objects relies on and enables a high precision reconstruction of a fundamental phase which can be confirmed as phase of a topologically equivalent image of a glottal master oscillator on the transmitter side. As topologically equivalent image on the receiver side, the self-consistent FD represents the long time scale part of the basic time scale separation known from human acoustic perception. The self-consistent reconstruction of the FD avoids the assumption of a frequency gap being necessary to justify the conventional assumption of a stationary or periodic voice source.

Acknowledgements: The author would like to thank M. Kob, B. Kröger, C. Neuschaefer-Rube and R. Schlüter, Aachen, J. Schoentgen, Brussels, A. Lacroix and K. Schnell, Frankfurt, and J. Rouat, Québec for helpful discussions.

6. References

[1] Rabiner L.R. and R.W. Schafer, "Digital Processing of Speech Signals",.Englewood Cliffs, NJ: Prentice-Hall (1978)
[2] Daubechies I., "Ten Lectures on Wavelets", SIAM, Philadelphia (1992)
[3] Gabor D, „Acoustic quanta and the theory of hearing", *Nature* **159**, 591-594 (1947)
[4] Drepper F.R., "Topologically equivalent reconstruction of instationary voiced speech", in C. Manfredi (editor), *MAVEBA 2003*, Firenze University Press (2004)
[5] Drepper F.R., „Selfconsistent time scale separation of non-stationary speech", *Fortschritte der Akustik-DAGA'05* (2005)
[6] Drepper F.R., "A two-level drive-response model of non-stationary speech signals", in M. Faundez-Zanuy et al. (Eds), NOLISP 2005, *LNAI* 3817, 125-138, Springer (2005)
[7] Drepper F.R., "Voiced excitation as entrained primary response of a reconstructed glottal master oscillator", *Interspeech 2005,* Lisboa (2005)
[8] Drepper F.R., „Stimmhafte Sprache als sekundäre Antwort eines selbst-konsistenten Treiberprozesses", *DAGA'06* (2006)
[10] Hohmann V., "Frequency analysis and synthesis using a Gammatone filterbank", *Acta Acustica* **10**, 433-442 (2002)
[11] Patterson R.D., "Auditory images: How complex sounds are represented", *J. Acoust. Soc. Jp*n. (E) **21**, 4 (2000)
[12] Moore B.C.J., "An introduction to the psychology of hearing", Academic Press (1989)
[14] McAulay R. and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation", *IEEE Trans. Acoust., Speech a. Signal Proc.* ASSP-**34**(4), 744-754 (1986)
[15] Heinbach W**.,** "Aurally adequate signal representation: The part-tone-time-pattern", Acustica **67**, 113-121 (1988)
[16] Terhardt E. „*Akustische Kommunikation*", Springer, Berlin (1998)
[17] Paliwal K.K. and B.S. Atal, "Frequency-related representation of speech", *Eurospeech 2003*, Genf (2003)
[20] Jean-Philippe Rameau, "Generation harmonique" (1737) reprinted in E. Jacobi (ed.), Complete Theoretical Writings Vol. 3, American Institute of Musicology (1967)
[21] August Seebeck, "Über die Definition des Tones", *Poggendorf's Annalen der Physik und Chemie* Vol. LXIII, pp 353-368 (1844)
[22] Schoentgen J., "Non-linear signal representation and its application to the modelling of the glottal waveform", *Speech Communication* **9**, pp. 189–201 (1990)